Bianca M. Colosimo
Nicola Senin *Editors*

Geometric Tolerances

Impact on Product Design, Quality Inspection and Statistical Process Monitoring

Springer

Geometric Tolerances

Bianca M. Colosimo · Nicola Senin
Editors

Geometric Tolerances

Impact on Product Design, Quality Inspection and Statistical Process Monitoring

Springer

Prof. Bianca M. Colosimo
Politecnico di Milano
Via La Masa 1
20156 Milano
Italy
biancamaria.colosimo@polimi.it

Prof. Nicola Senin
University of Perugia
Via Duranti 67
06125 Perugia
Italy
nsenin@unipg.it

ISBN 978-1-84996-310-7 e-ISBN 978-1-84996-311-4
DOI 10.1007/978-1-84996-311-4
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2010936681

Cover design: eStudioCalamar, Girona/Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

In Short

Geometric tolerances are increasingly being adopted in the design and manufacture of products, and awareness of their importance is now widespread.

This is not a book on geometric tolerances *per se*. A significant amount of literature work, including dedicated textbooks, is available that illustrates the main principles, definitions, and international standards related to geometric tolerances. Instead, this book explores the *impact* that geometric tolerances are having in specific areas of product development and manufacture, namely, product design, product quality inspection, and statistical process monitoring.

The book is structured as a collection of contributions from different authors, each highlighting one or more specific aspects related to the impact of geometric tolerances. New issues and also new opportunities are investigated.

From Dimensional to Geometric Quality Requirements

In the highly competitive scenario set by market globalization, it has become paramount for any company involved in product development to be able to reach the customer with appealing products in shorter times and with reduced costs. The capability of achieving and maintaining a consistently high level of *quality*, both of the final manufactured good and in the entire design and manufacturing process, is fundamental in achieving such a goal.

It is well known that the geometry of any manufactured product is characterized by variability with respect to its nominal counterpart: manufacturing-induced geometric variability is a long-established reality, which in turn affects functional performance and hence is tightly linked to quality. Acquiring a solid understanding of the geometric variation associated with a product has always been recognized as a key strategic element for achieving a competitive advantage. Defining acceptable boundaries for geometric variation is a mission-critical task in product development: it means identifying the best trade-off between what is achievable by the manufacturing process, what is measurable with sufficient reliability and reproducibility at quality inspection and process monitoring, and what is necessary for guaranteeing acceptable functional performance.

In recent years, the push toward the realization of ever better performing, ever more energy-efficient, and ever more cost-effective products, the constant search for innovation, and the continuous efforts directed at achieving technological breakthroughs, including the race toward miniaturization, have given rise to an increasing need for understanding, measuring, and controlling geometric variability to a larger extent than was possible with traditional means. Geometric tolerances have been gradually introduced alongside traditional dimensional tolerances in technical drawings, with the goal of providing a more comprehensive way for defining allowable variation for a given product geometry. While with dimensional tolerances geometric variation is summarized in a few synthetic indicators associated with linear dimensions, with geometric tolerances it is possible to capture a wider range of variations related to shape, position, and orientation of geometric features; those variations can be handled in a more comprehensive manner, and it is possible to define more effective constraints to control allowable geometric variability.

Geometric tolerances are now widely recognized as a key strategic element when it comes to ensuring that a certain level of quality is met by the product, whether it is a complex assembly or a single part or even a simple geometric feature. As a consequence of that, geometric tolerances are now basically commonplace in technical drawings.

The transition from dimensional tolerances to a reality where geometric tolerances are increasingly being adopted alongside dimensional tolerances implies, at a conceptual level, that we are moving from a scenario where manufacturing-induced, allowable geometric variability is defined and inspected through the variations of a few linear dimensions to a scenario where allowable geometric variability is defined and kept under control in a more comprehensive manner, and a larger amount of information pertaining to the entire shape of the manufactured part is kept under consideration.

Impact on Design, Quality Inspection, and Statistical Process Monitoring

It is clear that the introduction of more comprehensive approaches for defining and handling geometric variability is having significant repercussions on all the aspects of product development. On one hand, geometric tolerances introduce a wider array of options to designers for specifying allowable geometric variation.

While this opens up more opportunities for the accurate capture of a designer's intent, at the same time it raises new issues. With more options to define the boundaries of what could be considered as acceptable geometry, more choices must be made on what, when, and how to use such new resources. In order to make these choices wisely and effectively, deeper investigations are needed, aimed at gaining a better understanding of the potentially numerous effects, either direct or indirect, that any constraint on the shape/position/orientation of a geometric element, such as a feature, a part, or an entire assembly, may have on functional performance, manufacturability, assemblability, and ultimately cost and time to market.

Similarly, novel problems are raised at quality inspection. As quality is now related to a more comprehensive and complex representation of geometric variation – especially when compared with what could be previously achieved by looking only at linear dimensions, *i.e.*, with dimensional tolerances – it is clear that quality inspection needs to react accordingly, and traditional approaches may not be capable of successfully supporting novel requirements. It goes without saying that verifying whether a linear dimension is in tolerance or not is definitely simpler than, for example, verifying that the deviation of a free-form surface with respect to its nominal counterpart is within the specifications. The latter problem implies the existence of a solid background of solutions for solving not-so-straightforward issues, such as how to measure the actual geometry and how to assess from measurement data if the feature variation is confined within acceptable boundaries. Countless other examples could be made to testify how this conceptual transition from dimensional requirements to geometric requirements is raising significant issues, while at the same time providing the opportunity to explore new promising ground.

A similar scenario is faced by researchers operating in statistical process monitoring. Once more, how can monitoring solutions evolve so that they can handle the geometric variability associated with an entire surface, starting from traditional approaches where monitoring relies on the variations of a few synthetic indicators, such as those obtainable from a set of linear dimensions? Again, new issues appear alongside new opportunities to explore promising paths toward the development of innovative solutions.

Approach of the Book

It cannot be denied that a thorough investigation of the impact of the conceptual transition from dimensional requirements to geometric requirements is a daunting task, given the multitude of domains that characterize the complex and multifaceted process known as product development. Producing a textbook which is comprehensive enough, while at the same time achieving significant depth of detail, is close to unfeasible. The effects of introducing a potentially more effective set of constraints on geometric variability into product development are numerous, and delve deep into the process, affecting design, manufacturing process planning, quality inspection, and process monitoring activities in lots of direct and indirect ways.

This book chooses to deal with such complexity by collecting separate testimonies from different authors, each contribution concerning one or more specific issues that are deemed relevant within one of the subject areas listed earlier as product design, product quality inspection, and statistical process monitoring.

Manufacturing plays a central role in all the contributions: since geometric variability is induced by the process, manufacturing-related aspects cannot be ignored when dealing with any problem related to geometric variation. The manufacturing process, with its central importance, permeates the entirety of this book.

The contributions include original research, current literature reviews, and/or industrial practice experiences related to one of the subject areas listed above. This approach is intended to provide a significant breadth of views, while by no means attempting to achieve completeness. The hope is that a collection of significantly disparate topics covering such a wide range of subjects within product development may prove inspirational for researchers pursuing original research in one of the areas identified.

Reading Guide

The book is divided into three sections, each collecting contributions related to one of the subject areas:

- Part I: impact on product design;
- Part II: impact on product quality inspection;
- Part III: impact on statistical process monitoring.

In the following, a brief overview of the structure and contents of the book is presented. This is meant to provide an overall conceptual framework that allows the chapters to be placed in their correct context, and it also works as a reference guide, or gateway, for taking the reader toward the section he or she may be more interested in, for further reading.

Part I – Impact on Product Design

In this part of the book some key aspects related to the transition to geometric requirements are analyzed through the study of the introduction of *geometric tolerances* in product design. Geometric tolerances provide a multitude of new options for designers to specify allowable geometric variation. However, a competent and proficient use of such options implies a deeper understanding of the relationships among geometric variation, functional performance, and manufacturing and assembly processes. The principal new issues that are raised from the viewpoint of a product designer are related to *tolerance analysis* (*i.e.*, understanding how given tolerances affect functional performance) and – consequently – *tolerance synthesis* (*i.e.*, creating tolerances to be applied to a given geometry). The problem of tolerance synthesis is further divided into the problem of *tolerance specification* (*i.e.*, what tolerances to adopt, where to place them, and how to identify proper reference datums) and *tolerance allocation* (*i.e.*, what values to assign to defined tolerances).

Given its core importance in product design, *geometric tolerance specification* was chosen as the main subject for Chapter 1. In this chapter, an overview of the principal and most widespread approaches for selecting the most appropriate types of geometric tolerances, identifying where to place them on part geometry, and identifying proper reference datums is presented. Geometric tolerance specification is discussed under a dual perspective: producing protocols/guidelines to be manually applied by designers, and developing formal representations/methods to be implemented in computer-aided tools. In the chapter, tolerance specification is not addressed as a stand-alone problem; instead it is discussed in the context of its deep interconnections with tolerance allocation and tolerance analysis. The relationships between product geometry variability, manufacturing process, and fulfillment of functional requirements are highlighted, and the opportunities available for product design improvement are discussed.

Already introduced as one of the key aspects of product design, and not only because of its strong ties with tolerance specification as investigated in the previous chapter, the problem of *geometric tolerance analysis* becomes the core subject of Chapter 2. The chapter chooses to specifically focus on one of the most interesting, complex, and industrially relevant issues of tolerance analysis: the study of *tolerance chains* and their effects on the functional performance of assembly products. An overview of some of the most widely known literature approaches is presented, and they are discussed both in terms of protocols/guidelines that can be derived for designers to adopt and in terms of development of computer-aided solutions. Specifying the correct amount of allowable geometric variability of mechanical parts or simple features is of vital importance, as it has relevant repercussions both in the manufacturing domain, as it may damage the assemblability of the parts, and from a purely functional standpoint, as the overall functional performance of the assembly product may be degraded, or entirely lost, owing to the combined effects of the tolerance stack-up.

Part II – Impact on Product Quality Inspection

This part is dedicated to analyzing how the conceptual transition to geometric requirements is affecting current industrial practice and scientific research in the domain of *product quality inspection*.

The adoption of a more comprehensive set of constraints on the shape of a product raises the bar in terms of precision and complexity for all the activities related to its inspection. Current solutions (measurement instruments, measurement techniques, data analysis, and processing approaches) are in constant need of upgrading to keep pace with the evolving design scenario.

Fundamental challenges are related to how to acquire and analyze geometric information more efficiently and effectively in order to cope with the increased requirements; new opportunities surface as well.

This part of the book addresses two main subjects in the field of inspection: geometry *measurement solutions* (*i.e.*, measurement instruments and measurement processes) and *measured data analysis*. For both subjects, the selection of authored contributions was driven by the desire to highlight specific theoretical and applied subject areas that have recently attracted considerable interest, such as the inspection of microtopographic features, inspection through the adoption of sensor-fusion techniques, the development of innovative measurement process planning solutions, and the development of new formal representations for encoding a shape and its variations. *Shape* (*form*) and *shape error* (*form error*), in particular, emerge as the central subject shared by most of the contributions presented in this part of the book.

In detail, Chapter 3 is about measurement instruments such as profilometers and 3D microscopes and their application to the assessment of form error on microtopographic surface features. The problem is attracting considerable interest owing to the increased production of items such as microelectromechanical systems, semiconductors, other types of micromanufactured goods, and also more standard-sized parts characterized by microtopographic surface features. As form error assessment becomes relevant for such products, several issues must be faced to make profilometers and 3D microscopes adequate – with their peculiar modes of measurement and performance features – for such quality inspection tasks.

In Chapter 4, the problem of measurement process planning is analyzed for coordinate measuring machines (CMMs) involved in form error assessment tasks. As constraints on variability of surface shape evolve toward more comprehensive solutions, it becomes increasingly necessary to ensure that inspection captures all the relevant aspects of a surface geometry: for point-based measurement, this implies denser point clouds and/or a more thought-out placement of measurement points themselves. Significant measurement process planning problems arise, as the requirement of a more detailed inspection clashes with measurement time and cost.

Chapter 5 explores some issues related to the analysis of measured data. The focus is again shifted toward microtopography, and problems related to the assessment of form error on a microtopographic surface feature are analyzed, starting from a cloud of points as can be obtained by means of a profilometer or 3D microscopes. Peculiar aspects include the fact that the point cloud may be potentially suboptimal for a given geometry, owing to it being generally acquired by means of raster scanning, and that the exact localization of the surface feature within the acquired surface region may not be known, thus implying the need for feature identification and alignment with a nominal reference, before the actual form error can be assessed.

In Chapter 6, recent trends toward the development of multisensor measurement solutions are explored, and their role and applicability to inspection scenarios involving form error and geometric tolerances are discussed. The analysis is concerned with both measurement instruments (*e.g.*, combinations of CMMs and vision systems) and data-fusion approaches, where measurement data from different sources must be either used sequentially (*e.g.*, coarser measurement solutions used for optimal planning of more accurate, but slower, measurement processes) or merged into a single representation of the acquired geometry.

Finally, Chapter 7 is about shape coding. The need to identify efficient and effective solutions for encoding a shape into mathematical terms is central to the field of form error assessment. In this chapter, statistical shape analysis techniques, originally developed for modeling the shape of biological objects in the natural sciences, are analyzed and discussed, with particular reference to Procrustes-based methods. Possible solutions are proposed for some of the relevant issues that are related to the application of such methods to the representation of the geometry of manufactured parts.

Part III – Impact on Statistical Process Monitoring

In this part of the book, methods and tools are presented for performing *statistical process monitoring* (also known as *statistical process control* – SPC – or *statistical quality monitoring*) when the quality requirements concern the geometry of the manufactured item. The viewpoint is the one typical of the quality or process engineer who wants to quickly detect any change in the manufacturing process from its in-control or target state, given that the changes are usually associated with deteriorated process performance (*i.e.*, increase of nonconforming percentage). Traditional approaches in this area focus on dimensional requirements only, given that the basic tool in SPC, *i.e.*, the control chart, assumes that the quality characteristic of interest can be modeled as a univariate or multivariate random variable. When product quality is related to geometry rather than dimensions, the traditional control chart cannot be easily used, unless a synthetic set of indicators are used to summarize all the information contained in the cloud of measured points.

Chapter 8 presents two approaches that, given their inner simplicity, can be assumed as representative of industrial practice. The first approach consists in summarizing the information provided by the cloud of measured points in one synthetic indicator, namely, the maximum deviation of the actual shape from the nominal or ideal one, and then monitoring this indicator over time with a univariate control chart. The second approach extends a tool developed by Boeing for monitoring the upper flange angle at many different locations and consists in computing a control region where the upper and lower control limits are *k* standard deviations from the sample mean at each location. According to this method, an alarm is issued when at least one point in the whole set of data observed in a profile exceeds the control limits.

The following two chapters, *i.e.*, Chapters 9 and 10, are aimed at showing more complex but more efficient solutions for statistical process monitoring of geometric tolerances. The first type of solution relies upon a parametric model of the profile/surface geometry and hence is referred to as a “model-based” solution. In particular, two different methods are presented, depending on the specific type of model representing surface data (namely, linear regression with spatially correlated noise and principal component analysis). Chapter 10 shows a "model-free" approach where no model of the machined surface is assumed for the development of the monitoring strategy. In this case, the monitoring tool consists of a neural network operating in unsupervised mode to cluster the machined surfaces. Given a set of in-control surface data collected over a long period of time, the network is trained to detect any different patterns as representative of a different (and possibly) out-of-control state.

The performances of all three classes of approaches presented in the previous chapters of this part are compared in Chapter 11, where a real case study concerning roundness form errors of lathe-turned items is considered as a starting reference and different production scenarios are derived by slightly perturbing the case-study features. In this chapter, the best approaches in each production scenario are outlined in order to let the reader gain some insight into the advantages and disadvantages of the alternative solutions.

Milano, Italy *Bianca M. Colosimo*
September, 2010 *Nicola Senin*

Contents

Part I Impact on Product Design

Part II Impact on Product Quality Inspection

Part III Impact on Statistical Process Monitoring

Part I
Impact on Product Design

Chapter 1
Geometric Tolerance Specification

Antonio Armillotta and Quirico Semeraro

Abstract In conventional tolerancing, the efforts of designers are mainly directed at selecting suitable values for linear tolerances on part dimensions. These are either determined by trial and error through analysis calculations or optimized according to cost functions. In the transition to geometric dimensioning and tolerancing, the assignment of tolerance values must be preceded by a careful specification of the types of tolerances to be applied on part features. Along with the interrelations among features provided by datum systems, these define a tolerance model which captures design intent and is essential for the allocation and analysis of tolerance values. This chapter reviews the methods available for the specification of geometric tolerances, from common engineering practice to the development of computer-aided support tools. In the description of input data for tolerance specification, special attention is given to design requirements related to fit and function. The general strategy for the resolution of the problem is discussed, with focus on empirical specification rules and tolerance representation models which allow finite sets of tolerancing cases to be classified. The main approaches proposed in the literature for the generative specification of geometric tolerances are described and compared.

A. Armillotta
Dipartimento di Meccanica, Politecnico di Milano, Via La Masa 1, 20156 Milan, Italy,
e-mail: antonio.armillotta@polimi.it

Q. Semeraro
Dipartimento di Meccanica, Politecnico di Milano, Via La Masa 1, 20156 Milan, Italy

1.1 Introduction

For a long time, tolerancing was almost exclusively a design task. The engineering designer had the responsibility of assigning linear tolerances on some dimensions of parts according to all the precision requirements he could recognize on the product. These specifications were applied by manufacturing people (engineer, machinist, inspector), who contributed to design improvement by giving feedback on manufacturability and costs.

Things have changed considerably during the last few decades. Increasing market competition has forced companies to reduce defects and production costs. Technical standards have encouraged the adoption of geometric tolerancing, thus giving the designer more tools to ensure the assemblability and the correct function of the product. The statistical approach to production control and quality management has made it necessary to include processing constraints in product specifications. As a result, tolerancing is now a complex activity which involves both design and manufacturing personnel.

However, it is at the design stage that the diffusion of geometric tolerances has had the main impact. While geometric specifications can still be converted in the former representation language for manufacturing purposes, the designer cannot avoid dealing with the many types of geometric controls and datums now available. The complexity and the continuous evolution of standards make this adaptation process difficult and slow.

In response to such problems, much research has been done with the aim of supporting the assignment of geometric tolerances on manufactured parts (Hong and Chang 2002). The methods developed can be classified into the following three categories, related to the main design tasks of geometric tolerancing (Figure 1.1):

1. *Tolerance specification*. Starting from product data, features to be toleranced are located on all parts of the assembly. Some features among the most important in each part are selected as datums for the remaining ones. For each feature a set of tolerance types is then chosen in order to limit variation with respect to its nominal geometry and to datum features.
2. *Tolerance allocation*. Numerical values are set for all specified tolerances by either adjustment (*i.e.*, refinement of initial empirical values) or optimization (*i.e.*, minimization of a cost function subject to manufacturing constraints).
3. *Tolerance analysis*. Whenever required within an allocation procedure, design requirements are verified through the calculation of geometric entities such as gaps, angles, and dimensions involving different parts of the assembly (tolerance stackup).

These tasks are connected in a typical flow of activities, formally defined in early work on geometric tolerancing (Farmer and Gladman 1986). As said before, different routes can be taken for tolerance design (adjustment, optimization), each involving specific approaches to allocation and analysis. Whichever option he chooses, however, the designer must pay careful attention to tolerance specification, since a wrong choice of tolerance types and datum systems would affect the results of downstream design tasks.

This chapter deals with the problem of tolerance specification, which is a sort of “qualitative” tolerancing of parts (*i.e.*, without assigning values to tolerances) from product design information. Input data for this task include geometric data, possibly available in a CAD database. In addition, specification has to take into account some information not usually coded in geometric models, such as design requirements related to fit and function.

Like all tolerancing problems, specification can be studied at two different levels: empirical and generative. The former aims at constructing systematic procedures to be applied in everyday design practice, while the latter focuses on the development of formal methods to be implemented in computer-aided support tools. Both levels will be treated in the following sections, which describe the different aspects of the problem as they emerge from technical publications and the scientific literature.

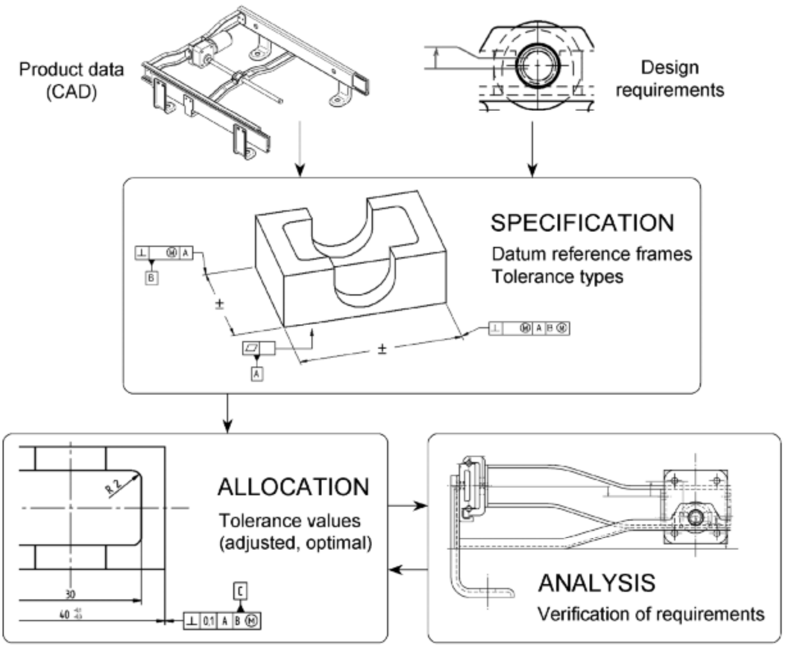


Figure 1.1 Design tasks in geometric tolerancing

1.2 From Linear to Geometric Tolerances

With linear tolerances, the specification problem is reduced to selecting functional dimensions to be controlled on the parts of a product. This is usually done by looking at the relational structure of the assembly. Each functional dimension is a distance between surfaces or other geometric entities (axes, edges, points) which establish relations with mating parts. When alternative choices are possible, manufacturing and inspection criteria help to select the proper set of dimensions to be toleranced. Figure 1.2 shows some functional dimensions selected on an example part (gearbox cover). According to common practice, dimensions are only toleranced if they contribute to design requirements (*e.g.*, gaps) for which little variation is allowed. For example, the size and the position of bolt holes are controlled by a general tolerance which still allows a correct fit owing to the large clearance available with fasteners.

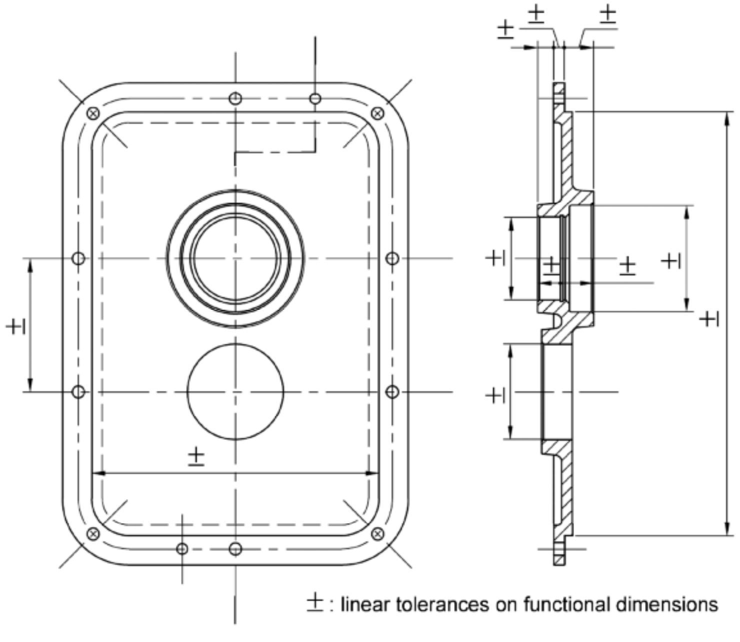


Figure 1.2 Specification of linear tolerances (past practice)

As is well known, linear tolerances have strong limitations with respect to interpretation, inspection, and effectiveness. The specification problem should definitely consider the use of geometric tolerances, in order to allow full control of geometric characteristics (form, orientation, location, profile, runout). The fundamental condition to be satisfied by specified tolerances is compliance with relevant ASME or ISO standards (ASME Y14.5.1 1994; ISO 1101 2004). Considering the difficulty of ensuring such a condition even in common design practice, it is not surprising that few of the proposed tolerancing methods generate fully consistent specifications. For example, it has been noted that some of them do not account for some key concepts of standards, such as the priority of datums in a feature control frame (Kandikjan *et al.* 2001).

It is convenient to decompose the general objective of compliance to standards into simpler acceptance conditions for tolerances resulting from a specification procedure. A set of geometric tolerances is usually considered acceptable if it guarantees an unambiguous positioning of each feature within controlled deviations (Shah *et al.* 1998). The degree of positioning depends on part function. For example, Figure 1.3 shows three different ways to tolerance an end plane of a pump housing with respect to the opposite end plane. Option a does not constrain the orientation of the feature to the datum, a condition that must always be satisfied for correct tolerancing. Option b constrains the orientation of the feature, but not its distance to the datum. Option c is preferred if additional position control is needed by some technical function (as is the case, owing to some tolerance chains involving the part).

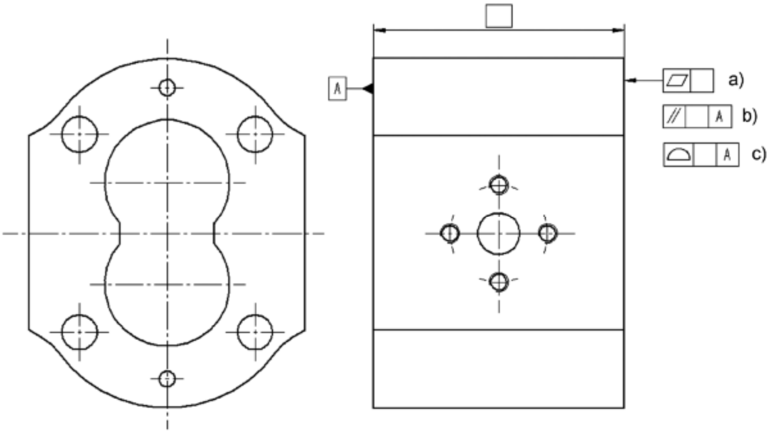


Figure 1.3 Alternative tolerance specifications for a feature

This acceptance criterion can be split into three conditions: validity, sufficiency, and consistency (Willhelm and Lu 1992). A tolerance can be invalid, *i.e.*, impossible to satisfy in practice, owing to a poor choice of its value (not accounted for in the tolerance specification). The condition of insufficiency has already been discussed for the example in Figure 1.3. An inconsistent tolerance loses its meaning owing to poor control of a referenced datum; in Figure 1.3, the lack of a flatness control on datum A affects the unambiguous positioning of the toleranced feature.

In a tolerance specification procedure, some assumptions should be made regarding the intended application domain. In most cases, the methods proposed in literature refer to the field of rotating machines and other types of assemblies with typical mechanical functions (support, sealing, gearing, *etc.*). These products, mainly composed of machined parts, do not cover the whole field of interest for tolerances. For example, special reasoning criteria apply to sheet metal and composite parts, which cannot be provided with accurate datum features and are assembled by means of positioning jigs (Wang *et al.* 2003).

Consideration should also be given to how tolerance values will be assigned. It has already been said that a specification procedure identifies tolerance types on datum and target features, leaving the allocation of tolerance values to a later phase. However, it has been pointed out that even tolerance types and datums should be optimized (Nassef and ElMaraghy 1997). According to this approach, a more correct objective should be the generation of a set of alternative tolerance specifications for a later selection according to economic criteria.

Another issue related to optimization is the choice of the right phase of product development at which the tolerance should be specified. In the literature, this task usually occurs in late stages of product design, since it needs detailed geometric models of parts as input data. It has been argued (Sudarsan *et al.* 1998; Narahari *et al.* 1999) that tolerancing decisions should be made as early as possible in product development (*design for tolerance*). This objective can be reached with a careful evaluation of the choices related to assembly configuration, design of part interfaces, and assembly planning. This concept, originally applied to linear tolerances, was reconsidered more recently in Pérez *et al.* (2006) for a possible extension to geometric tolerances.

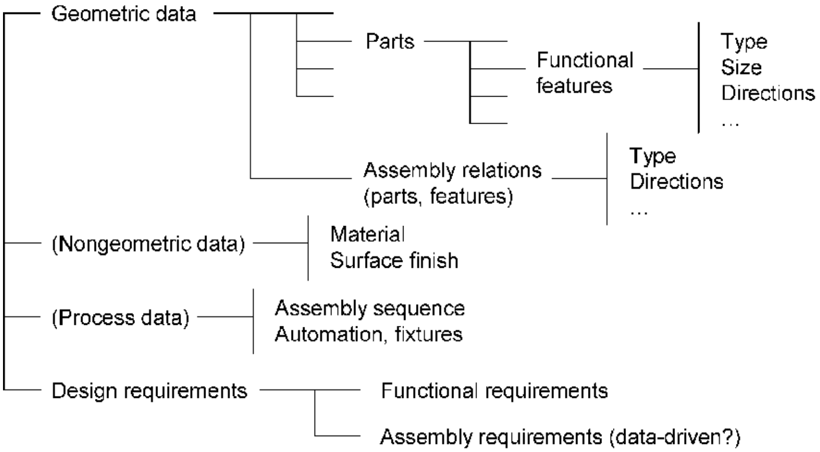


Figure 1.4 Input data for tolerance specification

1.3 Description of the Product

Figure 1.4 gives an overview of the data that a designer should collect in order to specify geometric tolerances. Except for some additional information that may be required when reasoning about geometric controls, the input data are quite similar to those needed in the context of linear tolerancing. In general, two types of data are absolutely needed for both empirical and computer-aided specification:

- a geometric description of parts and assembly relations; and
- a list of design requirements involved arranged by either product function or the need for a correct assembly.

They will be discussed in the following subsections. The remaining types of data (nongeometric and process-related) may not be strictly needed for a qualitative tolerancing of parts, although they could be of help in downstream design tasks (tolerance allocation and analysis). For instance, materials and surface finishes selected on parts lead to the choice of manufacturing processes, which may influence applicable tolerances. Similarly, any information that may be available on the assembly process (order of operations, level of automation, fixtures) could impose special precision requirements on apparently nonfunctional features (Roy and Bharadwaj 1996). Obtaining such data in advance can support early choices among alternative specifications.

The complexity of the product and the type of software support desired for tolerance specification suggest to what extent input data should be structured. To achieve a reasonable level of integration, they should be organized into formal assembly models to be extracted from product data with the interactive support of a 3D solid modeler (Kim *et al.* 2004).

1.3.1 Geometric Data

A product has a nominal geometry, upon which allowable variations will be defined through tolerances. Geometric data describe the individual parts and their mutual relations in the assembly.

Each part carries some general properties, such as the material (if deemed useful for tolerance specification) and the possible existence of identical parts in the assembly. The detailed description of the part is a list of its functional features, *i.e.*, those surfaces or features of size[1] that have geometric relations with features of mating parts. The list may be extended to nonfunctional features if they are to be controlled, as is advocated by practitioners in some industrial sectors complying to ASME standards.

[1] A feature of size is a cylindrical or spherical surface, or a set of two opposed elements or opposed parallel surfaces, associated with a size dimension (ASME Y14.5.1 1994). Pins, holes, tabs, slots, and any other feature that can be associated with a size dimension are examples of features of size.

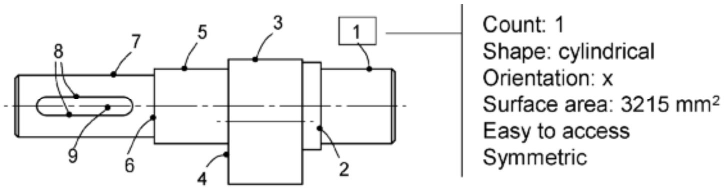


Figure 1.5 Tolerance-oriented description of part features

The functional features identified for an example part (crankshaft) are shown in Figure 1.5, along with a possible description of one of its features. Essential information for any tolerancing procedure includes the shape of the feature and the spatial orientation of its representative geometric entity (derived feature), as well as the number of equal features within a possible pattern. Further information can help to select features to be put in the *datum reference frame* (DRF) of the part. Depending on the specification method, this choice could privilege features with such favorable properties as large size, symmetry, and ease of access to reference surfaces of fixtures and gages.

Assembly-level data describe all pairwise relations between features of different parts. Figure 1.6 shows the assembly relations identified for an example product (rotary compressor) which includes the part shown in Figure 1.5. As will be clarified later, each relation is usually associated with one or more design requirements and can thus influence the choice of tolerance types on the features involved.

Assembly relations can be regarded as simple contacts between parts, without further detail. For a generative specification procedure, however, it is better to distinguish assembly relations with respect to the actual contact conditions they establish between parts. A possible classification of the most common cases is exemplified in Figure 1.6. It includes fits between features of size (a), seating relations with at least three points of actual contact (b), simple relations with at least one point of contact (c), and nominal relations with no actual contact required but a coincidence of ideal surfaces (d). Similar classifications have been proposed with an emphasis on the degrees of freedom subtracted to the relative motion of mating parts (Wang *et al.* 2003).

For empirical tolerancing, a designer should recognize geometric product data on engineering drawings and annotate them properly for later reference. Differently, the development of a generative tolerancing tool may require the extraction of product data from CAD models of assemblies and parts. Actually, only a few specification methods proposed in the literature provide such an integration as an alternative to a tedious input of data through formatted text files or graphical user interfaces.

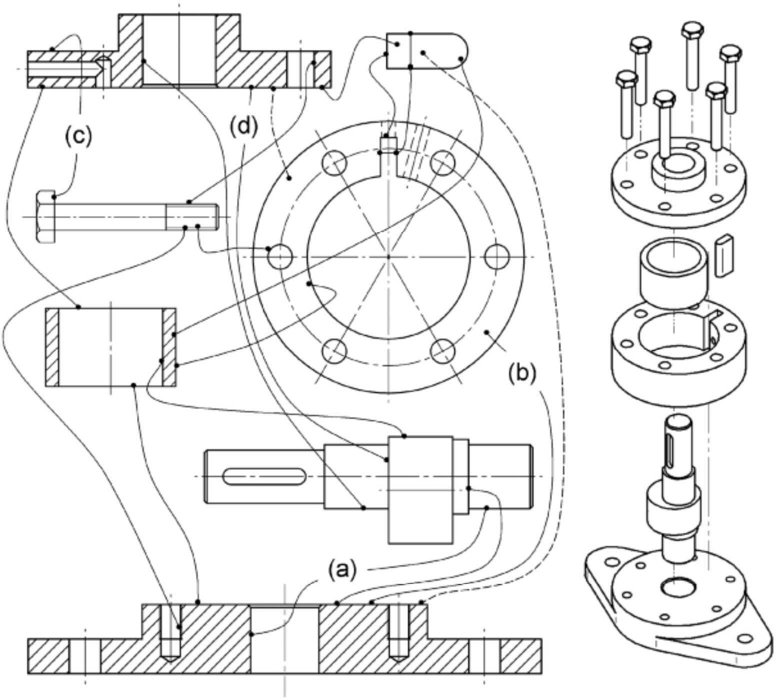


Figure 1.6 Assembly relations

1.3.2 Design Requirements

In addition to geometric data, a proper description of the product includes the design constraints to be satisfied through tolerances on part features. Such conditions may have been considered in previous design phases, but it is likely that only a few of them have been set as explicit product specifications. Now they have to be collected and classified as an input to an empirical or computer-aided procedure of tolerance specification.

1.3.2.1 Fit and Function

What is the purpose of asking for precision for a product? A first answer is related to product function, which may impose an accurate manufacture of parts in order to preserve the kinematics of mechanisms or the positioning of machine components with respect to external constraints. Some conditions of this type, which will be referred to as *functional requirements*, are shown in Figure 1.7, part a for an example product (belt drive assembly). The grooves of the pulley must be accurately positioned with respect to an external mating frame in order to be aligned with the corresponding features on the pulley of an electric motor: if such a condition is not met, the belts would slide irregularly and cause wear, noise, and loss of power. Similarly, the bottom lands of the same grooves must have a limited oscillation during rotation in order to avoid centrifugal overloads on the shaft.

A second issue to be considered is product assembly, which requires an adequate precision of parts in order to ensure that they mate correctly. Conditions of this type will be referred to as *assembly requirements* and are exemplified in Figure 1.7, part b. The pulley-side cover would not sit correctly on the housing if either a small gap with the bearing were not provided or its planar shoulder were not correctly oriented with respect to the centering boss.

In early papers on tolerancing, no distinction was made between the above two types of design requirements, which are both called "functional requirements" (Weill 1988). However, it has been noted that functional tolerancing is a more complex problem than tolerancing for assembly (Voelcker 1998). From one side, this is true because many technical functions involve variations in physical and mechanical properties of materials, which are difficult to treat in a general model. But there is another reason for distinguishing function and fit in the context of tolerance specification. Functional requirements must be explicitly set by the designer according to considerations depending on the type of product and on its relations with the outside world. Differently, assembly requirements are related to mainly geometric properties, which can, in principle, be reconstructed from product design data. For example, the gap required between the cover and the bearing in Figure 1.7, part b could be automatically recognized from the occurrence of a nominal relation associated by the designer with the planar surfaces on the two parts.

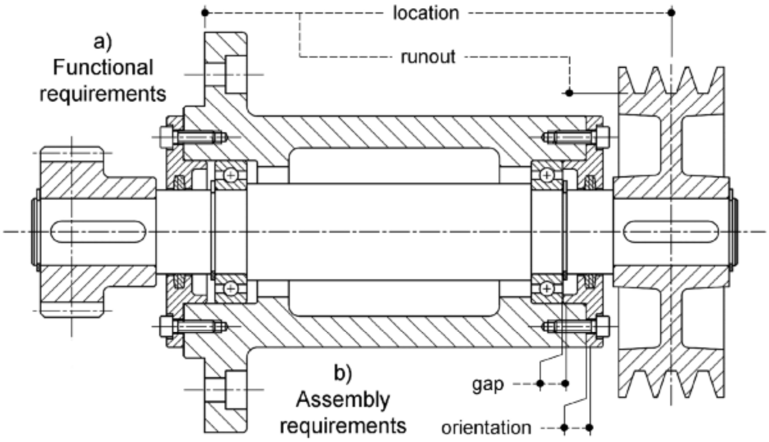


Figure 1.7 Functional and assembly requirements

Regardless of the way they are collected (interactive, data-driven), design requirements should be carefully treated as they are in a close relationship with the tolerances to be specified. Any specification method should include a *flowdown* procedure from requirements to tolerances. Although no complete flowdown procedure is available in the literature for geometric tolerances, some reasoning rules have been proposed. In one case (Wang *et al.* 2003) some types of functional requirements deriving from mechanical functions (sealing, rotation, balance, gearing, fastening, sliding, press fit, and clearance fit) are classified and related to typical choices for datums and tolerances. In another approach, not reported in detail (Roy and Bharadwaj 1996), such high-level requirements are converted into an intermediate form (equivalent functional specifications) which is claimed to be linked to tolerances through geometric rules.

1.3.2.2 Classification and Modeling of Requirements

In the tolerancing literature, assembly requirements are traditionally associated with chains of linear tolerances. Gaps, angles, and positions involving different parts are calculated as stackups of toleranced dimensions of individual parts. A classical treatise on the analysis and synthesis of statistical tolerances (Bjørke 1989) classifies tolerance chains and identifies the properties of "sum dimensions" of the different types of chains. It also defines special attributes (lumped and distributed) for the direction and the value of tolerances involved in a chain, a concept that seems to be associated with whether or not position controls are needed on part features. Moreover, it introduces the concept of compound tolerance chains, *i.e.*, building blocks corresponding to typical mechanical subassemblies for the construction of chains on complex assemblies.

The concept of *key characteristics* is an interesting attempt to provide a complete definition of design requirements and to discuss their proper treatment in assembly design (Whitney 1996; Whitney *et al.* 1999). Key characteristics are defined as a subset of applicable design requirements, which includes all conditions deemed critical by the designer. In a flowdown procedure, they are linked to datum flow chains, *i.e.*, chains of geometric relations which can be regarded as a direct extension of dimensional tolerance chains. A prioritization procedure resolves conflicts among key characteristics whenever either a product has simultaneous requirements or multiple products share common requirements. Datum flow chains are intended to be set by designers rather than reconstructed from geometric data.

The need to consider geometric tolerances in addition to linear tolerances has suggested different ways to define design requirements. In a first option, standard tolerances defined on parts are straightforwardly extended to assemblies. This leads to the definition of *linear assembly tolerances* (distance, gap, angle) and *geometric assembly tolerances* (parallelism, perpendicularity, angularity, concentricity, runout, location), which involve features of different parts (Carr 1993). These specifications are indicated on engineering drawings by a special feature control frame, which is shown in Figure 1.8 for the requirements discussed above for Figure 1.7. The same notation has been used to define functional tolerances in a generative specification method (Mejbri *et al.* 2005).

In another proposed definition, assembly requirements are treated as generalizations of fits between features of size (Voelcker 1998). This concept, referred to as *maximum material part* in analogy to the maximum material principle for fits, does not seem to have been developed further.

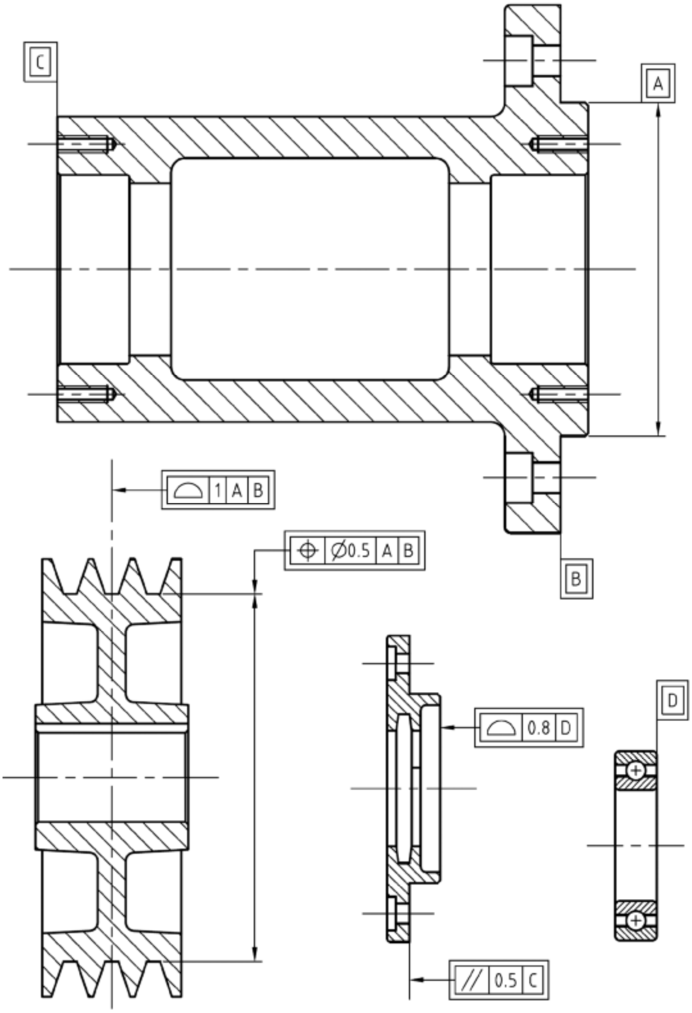


Figure 1.8 Design requirements defined as assembly tolerances

Definitions and drawing rules are not sufficient for a generative specification procedure, which also needs a representation model for design requirements. Such a model should codify requirements in a format allowing their easy association with predefined tolerance specifications.

The model of *virtual boundary requirements* has been proposed for requirements involving one or two parts, such as a fit or the protection of a minimum amount of material (Jayaraman and Srinivasan 1989; Srinivasan and Jayaraman 1989). A requirement is a set of conditions, each involving a pair of features of two different parts. For each pair of features, a virtual boundary is defined as the geometric entity which separates them at their maximum material conditions (a concept deriving from virtual conditions of clearance fits). The half-spaces constructed on the virtual boundary define the conditions of the requirement (*e.g.*, a hole surrounds a peg, and simultaneously two planes are in contact). Should the concept be generalized to a sufficiently large set of functional and assembly requirements, it could be the basis for a tolerance specification procedure (Willhelm and Lu 1992).

A more recent approach treats precision requirements in close connection with geometric relations and tolerances (Dufaure *et al.* 2005). The product is decomposed from both a structural and a functional viewpoint by defining three entities: parts, interfaces (surfaces, points, and lines), and functions. A function can represent a functional requirement, a contact condition between surfaces on different parts, or a geometric specification between surfaces on the same part. A graph provides a synthetic view of the model by showing possible relationships among entities: for example, functions are relationships among two or more interfaces, and parts are in turn technical solutions of functions. A software implementation of the model has been provided along with a demonstrative example (Teissandier and Dufaure 2007).

Another representation model connects geometric relations, precision requirements, and tolerances in a single hierarchy of hypergraphs (Giordano *et al.* 2005). The model describes all the above-mentioned items at different levels of detail, from the simple kinematic scheme of a mechanism to a full geometric description of parts. The authors foresee an automatic generation of the model, which would be the basis of an approach to tolerance specification.

1.3.2.3 Identification and Treatment of Requirements

As has already been noted, some design requirements can be identified from geometric data without having to be explicitly set by the designer. Although it is not clear how an automated recognition can be done in generic cases, some indications could derive from approaches proposed for linear tolerances.

In some cases, requirements are identified by search algorithms. In Mullins and Anderson (1998), search in a graph model of the assembly and geometric reasoning on contact surfaces among parts allowed the identification of some types of assembly requirements, such as the ending gaps of tolerance chains (mating conditions). A similar method was proposed in Zou and Morse (2004) for different types of gaps that can be created among parts in an assembly.

In Islam (2004a, b), a systematic procedure for identifying functional requirements was proposed within a method for the allocation of tolerance values on 1D chains. Again, the objective is to develop high-level requirements involving technical and safety considerations to a level of detail corresponding to part features (gaps, fits, flatness, location, *etc.*). The procedure is based on the *function analysis and system technique* (FAST), commonly used in the redesign of products by the value analysis method. The determination of limiting values for requirements is left to the experience of the designer. The development of a CAD-based software tool for the same purpose was described in Mliki and Mennier (1995), but details of identification procedures were not given.

Besides identifying design requirements, some methods can even generate their algebraic relations with dimensions of the parts involved. These are referred to as *functional equations* and are essential for tolerance analysis and allocation. The generation of functional equations has been studied in the context of linear tolerances. In Ramani *et al.* (1998), equations generated by a CAD-based tool which identifies all possible tolerance chains from functional surfaces and critical dimensions selected by the user were reported. In Söderberg and Johannesson (1999), tolerance chains were identified as loops in a network of relations among part features in an assembly model. For each part or subassembly, the model includes functions and reference frames deriving from relations with mating parts. Wang *et al.* (2006) generated 3D tolerance chains by a variational approach based on a solid modeler for geometric computations. Once a functional requirement has been interactively defined, the procedure identifies the surfaces involved in the chain. For this purpose, each surface is displaced from its original position by a small perturbation, and possible changes in the value of the functional requirement are detected after regenerating the geometric model. The chain results from the adjacencies of parts along the three reference directions.

Two additional problems can be cited to conclude the discussion of design requirements. They are less studied, but potentially important in order to integrate tolerance specification into the general process of product development. As a first question, how are design requirements (and indirectly tolerance specifications) influenced by choices related to the assembly process? Preliminary discussions and conceptual examples are included in some of the already-cited papers (Whitney 1996; Whitney *et al.* 1999; Sudarsan *et al.* 1998; Narahari *et al.* 1999). Further developments in this direction could help to plan assembly processes with reduced need for part tolerancing.

Equally interesting is the chance to reduce the number of requirements through product design choices. Empirical knowledge for this purpose exists and could be incorporated in computer-aided procedures. Some general rules are collected in McAdams (2003) and were classified according to their relevance in the different stages of product development and to the type of redesign involved (parameter calibration, detail changes, revision of product architecture).

1.4 General Approach to Tolerance Specification

The difficulties encountered by designers in the specification of geometric tolerances come from the lack of available explicit knowledge, which is limited to rules and basic examples provided by standards. Experts and researchers have tried to systematize tolerancing knowledge in the form of concepts and rules, which are also useful for generative tolerance specification. These include:

- empirical rules for the selection of datums and tolerance types; and
- classifications of tolerancing cases in order to define limited sets of solutions.

1.4.1 Empirical Specification Rules

Datum selection has been treated in tolerancing handbooks based on ASME standards (Meadows 1995; Drake 1999; Cogorno 2006) and case studies of tolerance specification (Wang *et al.* 2003). Figure 1.9 shows valid DRFs for some parts of an example product (wheel assembly, Figure 1.9a) in order to clarify the use of available rules. For all parts, datum precedences follow alphabetical order (A, B, C).

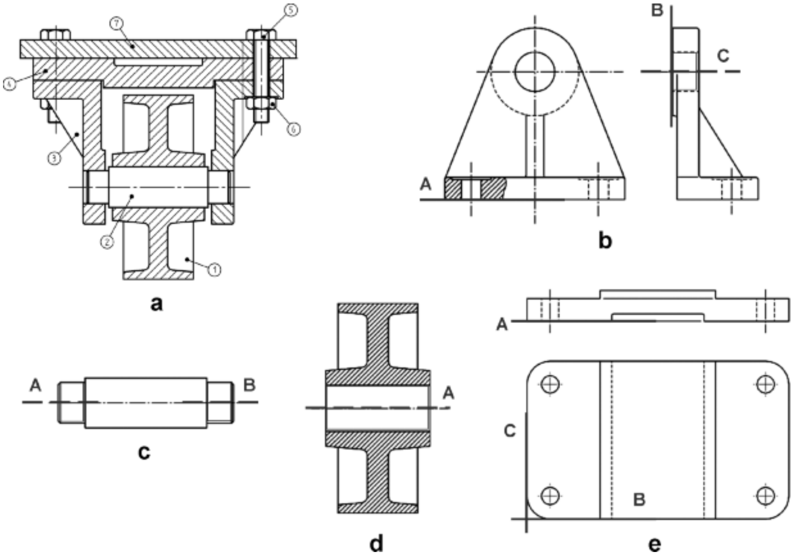


Figure 1.9 Examples of datum reference frames: **a** wheel assembly, **b** support, **c** axle, **d** wheel, and **e** plates

Since datums are often chosen among the functional features of a part, a basic rule suggests selecting the features that establish the most important assembly relations for product function. Such features may be rotation axes (support in Figure 1.9b, axle in Figure 1.9c, wheel in Figure 1.9d) or mating planes through which major forces are transmitted to the mechanism (support in Figure 1.9b, upper plate in Figure 1.9e). Actually, the importance of a functional relation cannot be easily evaluated in a data-driven procedure without decomposing it into geometric reasoning rules. A criterion proposed for this purpose is a large contact surface with mating parts, which allows repeatable fixturing and measurement (primary datums in Figure 1.9b–d). Another one is the number of degrees of freedom that mating parts subtract from part motion through the feature: rotational (seating) constraints suggest the choice of primary datums, whereas translational (locating) constraints determine secondary or tertiary datums (all parts in Figure 1.9).

Any different configuration of the DRF requires a specific design of manufacturing fixtures. Therefore, it is also desirable that datum features are easily accessed and simulated in machining and inspection, a condition that usually suggests a preference for planar and cylindrical surfaces.

The above rules are only a partial answer to the need for a guided choice of datums. To restrict the number of possible solutions, handbooks describe the most common types of DRFs. These include rotation axes defined by single or multiple features (Figure 1.9c, d), 3-2-1 schemes with three perpendicular planes (Figure 1.9e) or two planes and an axis (Figure 1.9b), and plane–cylinder sets with or without clocking. Each frame usually comes with suggestions for datum priorities and tolerance types to be selected for datums and targets. In the choice among alternative solutions, another suggested criterion is the minimum number of datum features, cited in Farmer and Gladman (1986) and Weill (1988) and demonstrated with examples of part redesign. As to current knowledge, the judgement of the designer is ultimately essential to select datum priorities for parts which do not match the most common cases (Mejbri *et al.* 2005).

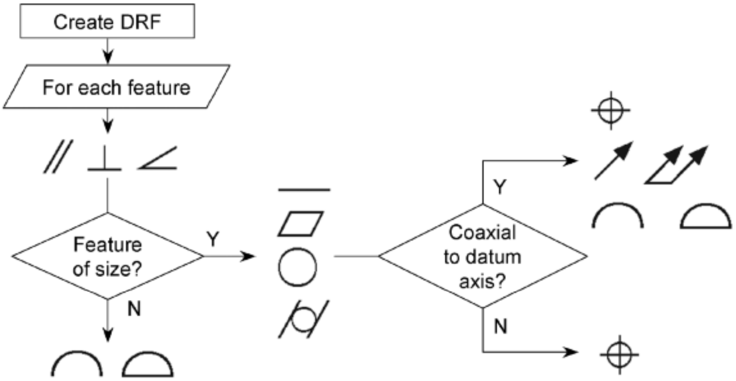


Figure 1.10 Tolerancing guidelines as per ASME Y14.5.1 (1994). *DRF* datum reference frame

Once a DRF has been selected, tolerances have to be chosen on all target features of the same part. Apart from the geometric classifications described below, some empirical guidelines for this purpose are provided by standards (ASME Y14.5.1 1994). For instance, the tolerance on a feature of size is chosen according to feature location (coaxiality to a datum axis) and to explicit design requirements (interchangeability, alignment, protection of minimum volumes, *etc.*). A partial outline of the guidelines is shown as a schematic flowchart in Figure 1.10.

1.4.2 Classification of Tolerancing Cases

Tolerance specification is an iterative process. At each step, a new target feature is toleranced with respect to a DRF including previously toleranced features. It is essential to understand how this can be done depending on the geometric properties of the features involved. Despite the huge number of practical design situations, such a task is less complex than it appears, since several studies have recognized that a finite number of tolerancing cases can be classified.

Most classifications are based on the concept of *invariance*. To control a feature, all of its possible deviations from nominal geometry must be limited within tolerance zones. Since an ideal feature is associated with the measured feature in the inspection, a deviation can be interpreted as a superposition of two transformations of nominal geometry: a variation in the geometric parameters of a feature (*e.g.*, the diameter of a cylinder) and small displacements (translations and rotations) of either the feature or its derived element (*e.g.*, the axis of a cylinder). For most features, however, there are translations and rotations that do not have to be controlled, as they do not alter feature geometry. Hence, the geometric control of a feature depends on its invariance with respect to translations and rotations in the principal directions of a given reference frame. Figure 1.11 shows the invariances for some common types of surfaces.

The concept is illustrated in the example in Figure 1.12. The features of an example part (a cover of the drive assembly in Figure 1.7) have to be toleranced in a predefined order (A, B, C, D, E, F, G) which may have been planned by a specification procedure. Given an *xyz* reference frame, feature invariances are first recognized and listed in a table. In accordance with Figure 1.11, they follow three different cases: (1) planes A, C, and G are invariant to translations along *x*, *y* and to rotations about *z*; (2) cylinders B and D are invariant to a translation along *z* and to a rotation about *z*; (3) the patterns of cylinders E and F, kinematically equivalent to prismatic surfaces, are invariant to translations along *z*.

Invariances help to find the displacements to be controlled on each feature by means of tolerances. This is done according to a general rule: when controlling a target with respect to a datum, invariances common to both features do not need to be controlled. Hence, cylinder B has to be controlled on only two rotations perpendicular to its axis, which can be done by an orientation tolerance. If A and B are selected as datums for all the other features, only their common invariance

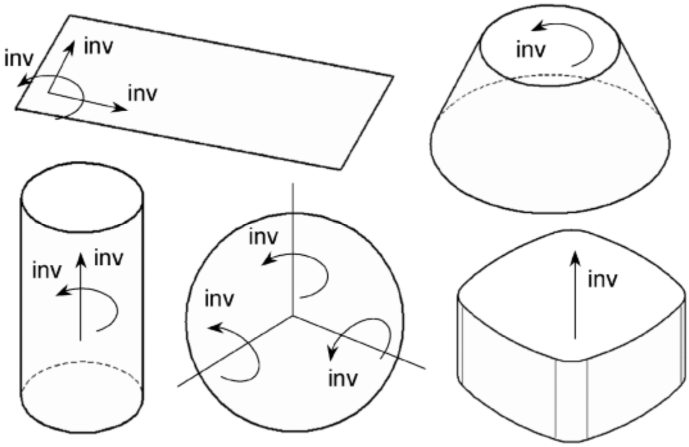
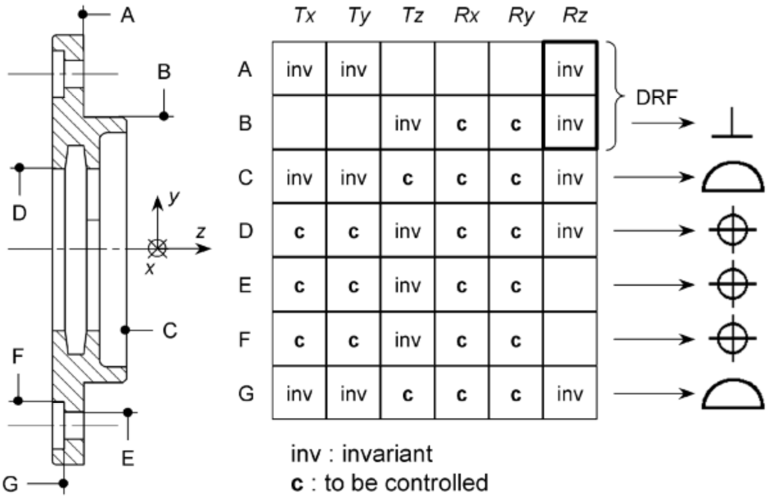


Figure 1.11 Invariances of common surfaces

(rotation about *z*) applies to the DRF. According to the above rule, plane C has to be controlled on a translation and two rotations perpendicular to it, which requires a profile tolerance. Controls of remaining features are similarly determined. It should be noted that kinematic control may not be the only criterion for the selection of the tolerance type: for example, the function of plane G (mating with a screw head) can be fulfilled by a simple orientation control rather than by a location control as suggested by its invariances.

A first list of tolerancing cases can be built by the tolerance representation model proposed in Clément *et al.* (1994) and also described in Clément and Rivière (1993), Weill (1997), and Chiabert and Orlando (2004). The model is based on a classification of part surfaces in seven classes according to their invariance with respect to translations and rotations. The defined types are spherical, planar, cylindrical, helical, rotational, prismatic, and generic (*i.e.*, noninvariant) surfaces (shown in reverse order in Figure 1.13, part a). Each invariance class is characterized by a *minimum geometric datum element* (MGDE), *i.e.*, the set of geometric entities (point, line, and plane) which is sufficient to determine the spatial position of the surface through dimensions and tolerances. Figure 1.13, part b shows the MGDE and the number of translational, rotational, and roto-translational invariances for each class.

The possible associations between two surfaces belonging to the same part are called *technologically and topologically related surfaces* (TTRSs). The composition of the MGDEs of the two associated surfaces allows one to determine the MGDE of the resulting TTRS, which is reclassified within one of the original invariance classes. In this way, the TTRS can be associated in turn with other surfaces or TTRSs, thus iterating the tolerancing process. Considering all cases deriving from possible geometries of associated surfaces, 44 composition rules are identified. These correspond to an equal number of tolerancing cases of a feature with respect to another.



	Tx	Ty	Tz	Rx	Ry	Rz
A	inv	inv				inv
B			inv	c	c	inv
C	inv	inv	c	c	c	inv
D	c	c	inv	c	c	inv
E	c	c	inv	c	c	
F	c	c	inv	c	c	
G	inv	inv	c	c	c	inv

inv : invariant
c : to be controlled

Figure 1.12 Selection of tolerance types from invariances

For example, Figure 1.13, part c shows the three cases of TTRSs that occur when associating a cylinder (target) with a plane (datum). If the cylinder is perpendicular to the plane, the resulting TTRS belongs to the invariance class of rotational surfaces; according to the same considerations made for the example in Figure 1.12, such a case corresponds to a perpendicularity tolerance on the target. If the cylinder is parallel to the plane, the TTRS belongs to the invariance class of prismatic surfaces and corresponds to a position tolerance with a basic dimension (distance of the cylinder axis to the plane). If the cylinder is oblique to the plane, the TTRS belongs to the invariance class of generic surfaces and corresponds to a position tolerance with two basic dimensions (angle of the cylinder axis to the plane, distance of an arbitrary point on the cylinder axis to the plane).

An extension to the TTRS model has been proposed in order to classify associations between surfaces of different parts (Clément *et al.* 1995). Such associations, called *pseudo-TTRSs*, are an additional way to represent precision requirements. Based on them, it has been demonstrated how functional requirements defined at a high level (*e.g.*, minimize the noise of a gear train) can be hierarchically translated down to a geometric level (Toulorge *et al.* 2003).

An interesting consequence of the above concepts is that a limited set of tolerancing cases exist and can be explored in a generative specification procedure. Regarding the number of cases, the classification based on the TTRS model is not the only one, as it assumes a definition of invariance classes which may not cover the diversity of features designed on manufactured parts.

The ASME "math standard" (ASME Y14.5.1.M 1994) classifies the cases of DRFs that can be constructed from elementary geometric entities (points, lines, and planes), considering the possible priority orders among datums. Each combination is described by its *invariants*, defined as dimensions (distances or angles)

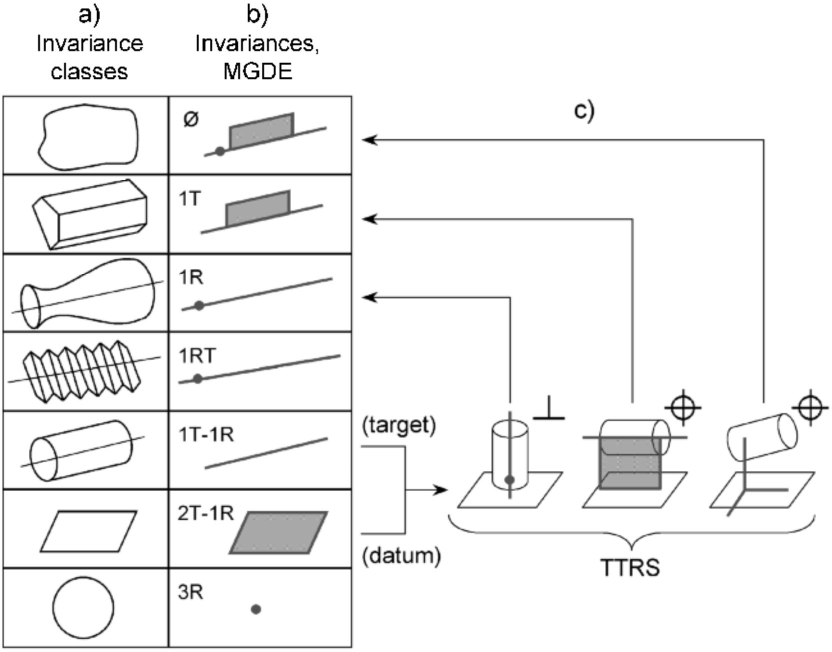


Figure 1.13 Classification of tolerancing cases by the technologically and topologically related surfaces (*TTRS*) model. *MGDE* minimum geometric datum element, *T* translational, *R* rotational *RT* roto-translational

which do not change after free transformations allowed by a datum system. Possible invariants include the three principal translations and rotations, the distance to an axis or point, and the angle to an axis. A set of 52 cases of DRFs is classified, a larger number than the number of associations between TTRSs since datum priorities are considered. Classified datum systems are then grouped according to invariants, thus giving six generic cases (one less than the number of TTRS invariance classes as the helical surface is not explicitly considered in the standard).

Another tolerance representation model (Shah *et al.* 1998; Kandikjan *et al.* 2001) is based on elementary entities coinciding with (or derived from) features on a part. Each entity has its translational and rotational *degrees of freedom*, which can be regarded as the complementary set of invariances. Any combination between two entities corresponds to the relation between a target and a datum, and maintains a number of degrees of freedom depending on the two original entities. Considering all possible combinations between two or three entities, 31 cases of DRFs and related tolerance cases are classified. Despite the analogy between the two approaches, the number of cases is smaller than in the TTRS model as the classification does not explicitly account for degenerate cases (such as coincident points, lines, or planes) and does not treat helical surfaces as separate cases.

A further model is based on similar considerations based on degrees of freedom (Hu *et al.* 2004). It consists in classifying all combinations between elementary entities defined on the same part, referred to as *cross-referenced variational geometric constraints*. Again, each combination corresponds to a way to control a target with respect to a datum. As only combinations between two entities are considered, the classification includes as few as 27 tolerancing cases.

All the above results allow one to treat the tolerancing problem as a single step of a procedure where the features to be toleranced on each part have been previously arranged in order. For this reason, they are the basis for some specification methods proposed in the literature, which will be described in the following section. These methods integrate classification rules in a reasoning procedure which analyzes features in relation to assembly geometry and design requirements.

1.5 Generative Specification Methods

Several methods have been proposed for the selection of DRFs and tolerance types on part features. In some cases, the results are directly transferred to tolerance allocation and analysis, thus using the specification method as a part of a more comprehensive tolerancing procedure. Some solutions have been implemented in software tools with possible CAD integration.

The methods developed at sufficient completeness and generality are based on six different approaches to tolerance generation:

1. analysis of a TTRS hierarchy constructed from feature relations;
2. analysis of the degrees of freedom of part features;
3. propagation of datums and geometric controls from special features (mirrors);
4. decomposition of functional requirements at the feature level by simplification of the assembly relation graph;
5. propagation of requirements on individual parts from relational information (positioning table); and
6. modeling of geometric variations on features involved in requirements (variational loop circuit).

Table 1.1 compares them with respect to the overall strategy adopted for the treatment of product data. Basic criteria include the existence of an *a priori* classification of tolerancing cases, the use of a global description model of assembly relations, and the explicit treatment of design requirements.

As said before, a classification of tolerancing cases allows one to treat single steps of the problem with reference to a set of predefined cases. To develop a specification method from a classification, features to be toleranced on each part have to be properly ordered. The approaches based on this principle differ in the way they analyze product geometry to identify priorities among features. A classification helps to improve the correctness and completeness of the tolerances generated, and usually has a mathematical foundation which can allow one to generate functional equations for downstream tolerancing tasks.

Table 1.1 Comparison of tolerance specification methods

Method	*a priori* classification	Model of relations	Explicit requirements
TTRS	X	X	–
Degrees of freedom	X	X	–
Mirrors	–	X	–
Function decomposition	–	X	X
Positioning table	–	–	X
Variational loop circuit	X	–	X

TTRS technologically and topologically related surfaces

A specification method can reason on a global, usually graph-based, model of relations among part features. Such a representation is also used in many generative methods proposed for assembly planning (Abdullah *et al.* 2003), and lends itself to search strategies based on the exploration or decomposition of the graph. However, the construction of a global assembly model is tedious and error-prone. As a consequence, this approach is especially practical when an extraction of assembly relations from CAD data is provided. The alternative approach is the construction of a relational structure including only the parts involved in a functional requirement, as a geometric extension of dimensional tolerance chains.

In some methods, functional and assembly requirements are explicitly defined by the designer or automatically recognized from geometric data. This approach guarantees that all relevant requirements are satisfied by the tolerances generated, but can be complicated by the large number of requirements occurring for complex assemblies. Moreover, if the requirements are not defined on a global relational model, the tolerances generated from different requirements have to be properly combined (the problem has not been clearly addressed).

Although sometimes oversimplified, the following description will capture some useful concepts for understanding the different solutions proposed for toler-

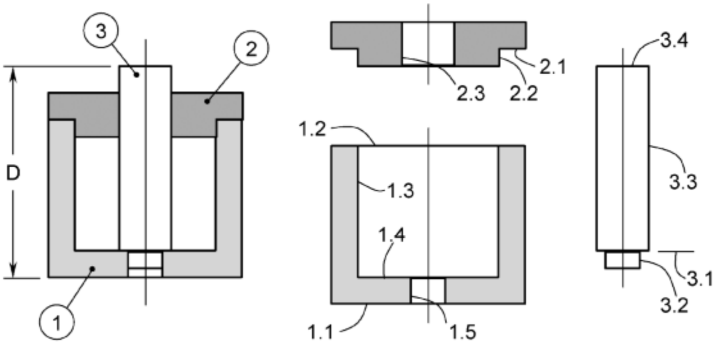


Figure 1.14 Conceptual example

ance specification. To clarify some details of the approaches, reference will be made to a simple conceptual example. This is described in Figure 1.14, which shows an assembly composed of three parts and the functional features on each of them. Assembly requirements prescribe a stable mating of planar surfaces and a tight clearance fit of cylindrical surfaces. An additional functional requirement is that a nominal value of distance *D* is satisfied within a given allowance.

1.5.1 Technologically and Topologically Related Surfaces

The first and most cited specification method is based on the TTRS model for tolerance representation and classification (Clément *et al.* 1994). The general approach, which has been described before, models each surface by an invariance class and a set of reference geometric entities (MGDEs). When two surfaces are considered in association with each other, these data allow one to determine the tolerances needed to control their relative position. The association itself (TTRS) is modeled with an invariance class and a MGDE. This concept is exploited in a recursive strategy, which builds a hierarchy of TTRSs on the functional features (here called “surfaces”) of each part.

The sequence of surface associations on a part derives from a graph which represents surface relations among parts in an assembly. Any of the loops of the graph involve two or more parts: for each part, the loop includes two surfaces, which can be associated in a TTRS. By exploration of all loops in a user-selected order, the TTRSs defined on the parts are composed to form a hierarchy. This allows one to identify applicable tolerance types according to the classified tolerancing cases from the properties of the TTRSs (invariance class and MGDE).

Figure 1.15 shows the recursive procedure that builds the TTRS hierarchy for the example. The graph is initialized with the relations between pairs of surfaces corresponding to functional features of parts (Figure 1.15a). The loops of the graph are ordered in a sequence (L1, L2, L3, L4) which is then followed when associating surfaces into TTRSs. For example, loop L1 involves two surfaces on part 1 (S12 and S13) which are associated into T1, and two surfaces on part 2 (S21 and S22) which are associated into T2. The next loops allow one to build the other TTRSs, namely, T3 and T4 (Figure 1.15b), T5 and T6 (Figure 1.15c) and eventually T7, T8, and T9 (Figure 1.15d).

For complex assemblies, the loops have to be properly ordered in order to construct a TTRS hierarchy consistent with design practice. For this purpose, predefined tolerancing templates are provided for common functional requirements on the basis of technical information published by vendors of components (bearings, slideways, gears, seals, *etc.*).

The TTRS-based specification method has been developed into an interactive software tool (Salomons 1995; Salomons *et al.* 1996). In addition to the original method, some criteria have been introduced to order the loops of the graph and to construct the MGDE of a TTRS from the MGDEs of associated features. These tasks remain mostly left to user interaction, along with the initial preparation of the assembly model in the lack of a direct CAD interface.

Loop and MGDE selection are two critical steps for an implementation of the method at a higher automation level. It has been suggested (Desrochers and Maranzana 1995) that priority should be given to the loops that (1) involve the minimum number of parts, (2) are related to tolerance chains detected from sequences of contacts along a same direction, and (3) are of large size or involve specific types of features such as planes. For the choice of the MGDE, the following criteria help to treat the cases in which a TTRS has a different invariance class from associated features: (1) specification of datums and basic dimensions with a functional interest and (2) predefinition of clearances or interferences on fits.

The construction of TTRSs on geometric models of assemblies and parts has been implemented in commercial CAD packages. In one case (Toulorge *et al.* 2003) a semiautomatic software tool is integrated in the product life cycle management system of an automotive manufacturer, with the already-described extension of the model (pseudo-TTRSs) to represent functional requirements (here referred to as *use aptitude conditions*). A more complete implementation in the

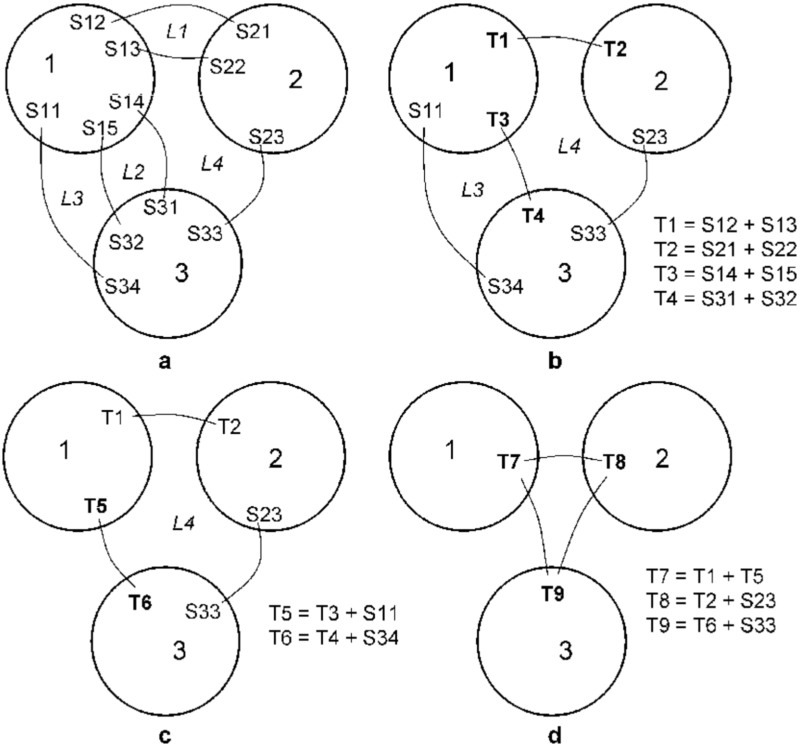


Figure 1.15 TTRS hierarchy: **a** loops,surfaces, and parts, **b** building of T3 and T4, **c** building of T5 and T6, and **d** building of T7, T8, and T9

same context was reported in Buysse *et al.* (2007): to streamline data input, TTRSs are selected by the user through a graphical procedure which displays the related MGDE as a skeleton geometric representation of parts.

1.5.2 Degrees of Freedom

The degree-of-freedom representation model (Shah *et al.* 1998; Kandikjan *et al.* 2001) and the underlying classification of DRFs have been developed into a tolerance specification method. Initially proposed as a validator of user-input tolerances in a CAD environment, the method was later extended with a rule-based specification strategy (Wu *et al.* 2003). As such, it has been implemented as a stand-alone tool based on a solid modeling engine.

A pair of elementary geometric entities (points, lines, planes) can be linked by a *metric relation* (coincidence, parallelism, distance, *etc.*), which must be controlled by tolerances. For each possible combination of entities, the classification provides the number and the types of translational and rotational degrees of freedom to be controlled along with the directions involved, as well as shape degrees of freedom to be controlled by size tolerances. Constraints on degrees of freedom are then converted to linear and geometric tolerances by means of rules, which also allow one to select datums and basic dimensions.

Target entities and metric relations are represented in a graph, which spans the whole assembly and includes size and geometric tolerances as attributes of relations. The graph is constructed from the union of characteristic subgraphs which correspond to different cases of relative tolerancing between entities.

Figure 1.16 shows the graph-based tolerance model for part 2 of the example assembly. Features of the same part (F2.1, F2.2, F2.3) are associated with self-relations of shape and size, *e.g.*, radius, and are linked to one another by metric

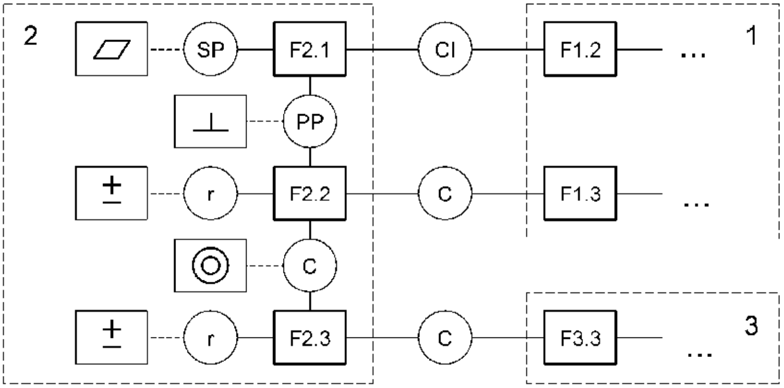


Figure 1.16 Metric relations in the degree-of-freedom method. *SP* shape, *r* radius, *PP* perpendicularity, *CI* coincidence, *C* coaxiality

relations of perpendicularity and coaxiality. A proper type of tolerance is specified to control each metric relation. Additional metric relations of coincidence and coaxiality establish links with features of other parts.

The tolerance specification method consists in the semiautomatic construction of the graph of metric relations completed with tolerance attributes. In the first step, each metric relation among entities is converted into a directional relation between a datum and a target. The selection of datums gives priority to planar faces, or entities with many metric relations, or entities with a large size. In a second phase, linear and geometric tolerances are determined for each metric relation according to the classification, in an ordered procedure that allows one to set datum priority. The tolerances generated are eventually validated by verifying that each target is correctly restrained by one or more datums. A further option is the automatic recognition of tolerance chains for allocation and analysis.

1.5.3 Mirrors

Another approach to tolerance specification (Wang *et al.* 2003) is based on *mirrors*, *i.e.*, planar surfaces involved in assembly relations. They are classified into strong mirrors (corresponding to seating relations) and weak mirrors (corresponding to relations which do not require a minimum of three actual contact points). Among all assembly relations between features, some involve mirrors and have a special influence on tolerance generation. The part with the maximum number of mirrors is considered more important and is the first to be toleranced. The existence of mirror relations with already-toleranced parts allows one to select the remaining parts in order of priority.

Tolerances are generated on each part in two steps. In the first one (temporary tolerancing), datums are selected according to rules involving mirrors and design requirements: *e.g.*, strong mirrors are used as primary datums, locating features as secondary datums, *etc.* To construct a DRF with priority, features are clustered according to their adjacency to mirrors. If the part has a mirror relation with a

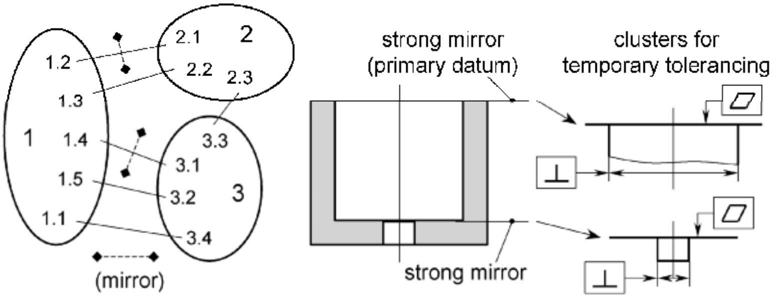


Figure 1.17 Temporary tolerancing from mirrors

previously toleranced part, the selection of datums is influenced by the relation (in practice, the datum system of the previous part is "reversed" on the new part). Form, orientation and location, tolerances are then assigned to datums and targets by a rule-based procedure. In the second step (final tolerancing), other rules are applied to harmonize tolerances on parts based on the whole set of assembly relations with other parts.

Figure 1.17 shows the use of mirrors on the example. Part 1 is considered as the most important since it has two mirrors, both strong owing to the occurrence of a minimum of three points of actual contact. One of the two mirrors (feature 1.2) is preferred as the primary datum of the part since it is more readily accessible in manufacturing and inspection. Each mirror is taken as a starting point to cluster adjacent features, which are temporarily toleranced by the already-cited ASME classification of tolerancing cases. For both clusters of the example, the temporary tolerancing scheme consists of a flatness control on the plane and a perpendicularity control on the cylindrical feature.

The mirror method has been demonstrated on a realistic example with results in good accordance with design practice. No details are available on software implementation and further validation of the approach.

1.5.4 Function Decomposition

This approach consists of a procedure, originally intended as a training tool, which was later integrated in a generative specification method (Ballu and Mathieu 1999). It combines a global strategy based on assembly relations with an explicit treatment of functional requirements. These are set by the designers as links between features or parts in the graph of assembly relations.

For each functional requirement, the graph is iteratively simplified by recognizing special patterns of contacts (*e.g.*, series or parallel). At each iteration, the features involved in patterns are aggregated in compound features for the next iteration. The features are thus arranged in a hierarchy, which is then analyzed in order to recognize the features involved by the requirement. This is done by rules which take account of the geometry of features at intermediate levels (feature types and associated direction).

After this procedure has been repeated for all functional requirements, all influenced translations and rotations are known for each feature of each part. Therefore, it is possible to select the type of geometric control on the feature (orientation or location) and the related datums in order of priority (*e.g.*, planes and cylinders of permanent contact precede cylinders of floating contact).

To demonstrate the approach for the example, Figure 1.18 shows the selection of the features involved in the functional requirement considered for the assembly. In Figure 1.18a, the graph of relations is built and completed with the functional requirement, related to the distance between features 1.1 and 3.4. The graph is then simplified three times through the recognition of contact patterns in parallel (first and third steps) and in series (second step). By keeping note of graph simplifications, one obtains a decomposition of the functional requirement as a hierarchy of features (Figure 1.18b). Within the hierarchy, features which do not influence the requirement are discarded from further consideration: specifically, all cylindrical contacts do not contribute to the translational constraint of functional requirement and thus do not need to be toleranced (at least with respect to the functional requirement).

The procedure described has been incorporated in interactive specification tools, aimed at supporting designers during the whole product development cycle, starting from conceptual design (Dantan *et al.* 2003a). To better link input data to product design information, functions defined at a high level are converted into a set of key characteristics related to part features. In addition, an attempt is made to consider the effect of the assembly plan on tolerances. In some cases, the fulfillment of a functional requirement (*e.g.*, a clearance fit) depends on the configuration established in previous assembly operations. The dependence is modeled by means of rules for the selection of maximum-material and least-material modifiers (Dantan *et al.* 2003b, 2005), defined from a tolerance representation based on the virtual boundary model (Jayaraman and Srinivasan 1989).

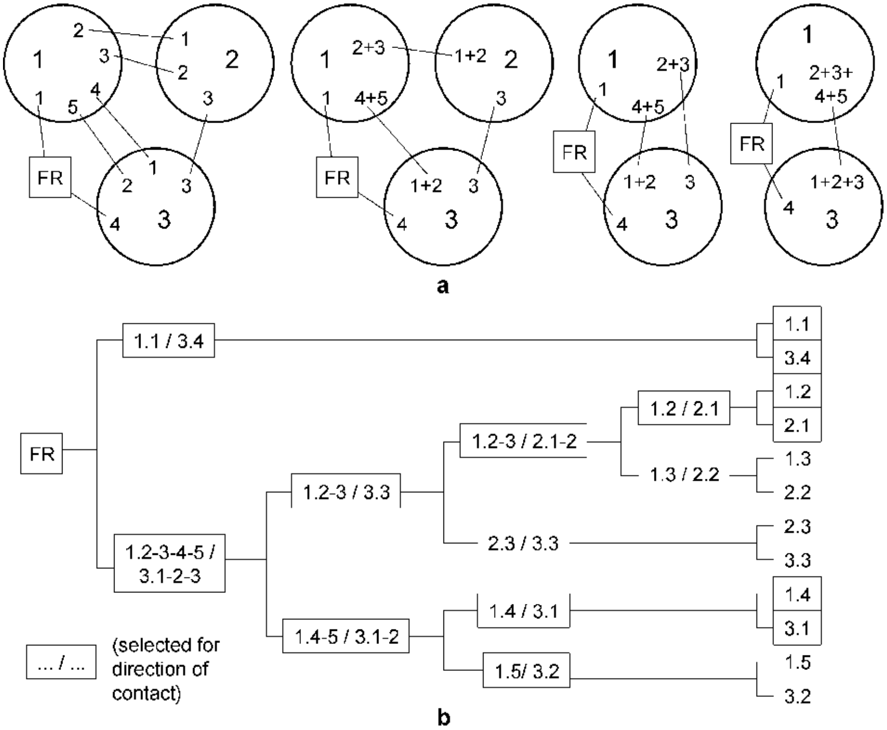


Figure 1.18 Decomposition of a functional requirement (*FR*): **a** graph of relations completed with the functional requirement, and **b** decomposition of the functional requirement as a hierarchy of features

1.5.5 Positioning Table

Each design requirement usually involves only a limited number of parts. For this reason, the generation of the tolerances needed to satisfy a requirement can be done without considering the whole graph of relations among part features in an assembly. This basic consideration introduces the specification method proposed in Anselmetti and Mawussi (2003) and Mejbri *et al.* (2003, 2005).

A functional requirement is defined as a geometric tolerance assigned to an ending entity on a part with respect to external datums on a base component. Such a "generalized" tolerance must be converted into a set of regular tolerances (*i.e.*, with only internal datums) on the individual parts involved in the requirement. For this purpose, a *positioning table* is defined for each part to represent contact relations with mating parts. The table lists the contacts in an ordered set, corresponding to a priority order of datums for all requirements involving the part. Datum order is defined by the designer according to the types of actual contacts required between features (seating contacts on three actual points, simple contacts, *etc.*) and to the need to define complex cases of datum frames (composite datums, symmetries, feature patterns, *etc.*).

With use of the positioning tables of parts, a chain of relations (mechanical joints) up to the base component is constructed for each functional requirement. The chain is then explored recursively from the first part by following the chain of relations. At each step, the external datums of the geometric tolerance related to the requirement are converted to internal datums.

The DRF used for each tolerance generated is validated, *i.e.*, reduced to the simplest subsystem that is able to unambiguously control the spatial position of the tolerance zone. The validation criterion is based on the invariant degrees of freedom, *i.e.*, the translations and rotations with no effect on the tolerance zone. If a datum deriving from the contact with another part is not needed to constrain the degrees of freedom of the tolerance zone, that part is considered inactive. The validation is based on a classification of tolerance zones, considered in combination with possible feature types. A set of validation rules allow one to verify if a certain system of datums contributes to constraining the degrees of freedom, *i.e.*, eliminates degrees of freedom which are noninvariant for the tolerance zone.

Figure 1.19, part a shows the positioning table for part 3 of the example. The first column of the table means that a contact exists between planar features 3.1 and 1.4. The remaining columns are similarly interpreted with other types of features and relations. In Figure 1.19, part b, each relation generates a tolerance on the feature involved according to a rule drawn from a classification. The column order is set by the designer and determines the priority of datums (A and B for this part).

The approach was later completed with an automated generation of functional equations for tolerance allocation (Anselmetti 2006). In addition to functional requirements defined by the user, some assembly requirements are automatically identified from geometric data. These include interferences and alignments be-

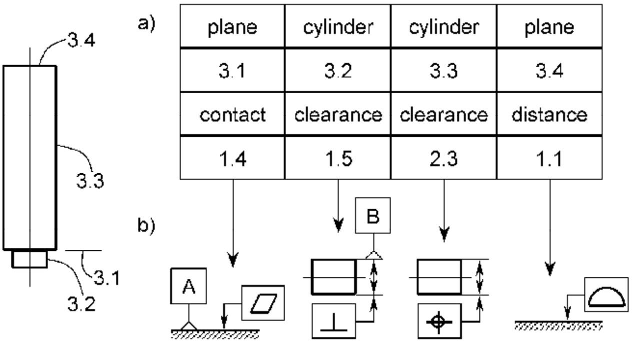


Figure 1.19 Positioning table and tolerancing rules

tween surfaces and constraints on the mounting surfaces of standard components (threaded fasteners, dowel pins, *etc.*). The method is implemented in an interactive software tool based on a commercial CAD package.

1.5.6 Variational Loop Circuit

Another method derived from a classification of tolerancing cases was proposed in Hu *et al.* (2004) and Hu and Xiong (2005a, b) as part of a comprehensive method for the design of geometric tolerances. Unlike approaches based on TTRS and degree-of-freedom models, it needs assembly requirements as explicit input data. The method is based on an assembly model which includes both nominal features (*e.g.*, planes, cylinders) and nominal derived features (*e.g.*, cylinder axes). The relation between two entities is based on nominal parameters which correspond to nominal dimensions of the parts along reference directions.

For each design requirement, a directed graph called a *nominal loop circuit* (NLC) represents all relations among the entities involved and the related nominal parameters. The NLC is constructed from subgraphs corresponding to fits and other predefined patterns of relations. Equations linking the nominal parameters are generated from the structure of the graph.

Geometric variations on features are taken into account by defining associated features and associated derived features, *i.e.*, features obtained by measurement of real surfaces. The NLC for each requirement is thus enriched with associated entities and their relations with nominal entities. These relations are expressed through additional parameters (variations of parameters), which allow one to construct a second graph called a *variational loop circuit* (VLC). Equations linking nominal parameters and variations of parameters are also generated from the VLC and used in tolerance allocation and analysis. Each VLC is eventually completed with geometric tolerances, expressed as inequations involving variations of parameters and identified by geometric rules.

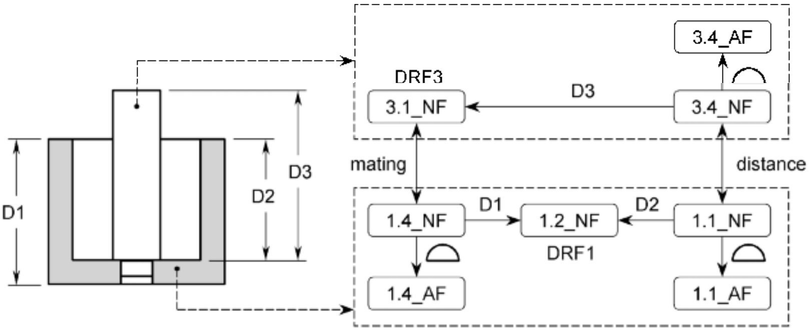


Figure 1.20 Variational loop circuit for a functional requirement. *DRF* datum reference frame, *NF* nominal feature, *AF* associated feature

Figure 1.20 shows the corresponding graphs for the functional requirement considered for the example. Since the requirement involves only planar surfaces of just two parts (1 and 3), the circuit is very simple and does not have derived features (*e.g.*, axes). The NLC is first built with only nominal features. Each nominal feature is linked to the features (a single plane in this case) of the DRF of the same part (DRF1 or DRF3). Each link represents the geometric relation to the datum, which consists of one of distances D1, D2, and D3. To build the VLC, the associated features are added. Each of them is linked to the corresponding nominal feature through geometric parameters (not shown in the figure) that allow one to select the proper type of tolerance and the possible need for basic dimensions from a classification of tolerancing cases.

1.6 Conclusions

What can be learned from this review on tolerance specification? It seems clear that the introduction of geometric tolerances in design practice has created a problem that did not exist with linear tolerancing. The designer is bound to declare the degree of geometric control of every part feature, in order to guarantee the fit and function of the product. This is not easy, since many types of geometric controls exist and their standard representation language is complex. While this difficulty can be overcome through training and experience, it remains hard for a designer to reason on the combinatorial explosion of assembly relations and design requirements of a complex product.

Research is attempting to remove these obstacles. Classifications of tolerancing cases and guided procedures are increasingly available for training and design practice. Generative methods are being developed and integrated in CAD software as a further help when dealing with complex assemblies. They are not intended as push-button tools and usually involve a moderate amount of user interaction.

Some indications can be drawn for future work on tolerance specification. The main developments are expected in three areas: the treatment of product data, the general strategy for problem solution, and the compatibility with downstream tolerancing tasks.

To allow a complete description of the product, some improvements need to be made to the treatment of design requirements. Owing to their large number, it is impractical for the identification of requirements to be left completely to the designer. Procedures will have to be developed for the automated identification of assembly requirements, and functional requirements will have to be better modeled and classified. The effect of the assembly process on tolerances will have to be better clarified. Improved CAD interfaces will help to streamline the construction of product description models without tedious and error-prone interactive procedures.

To improve tolerance generation, a balance will have to be found between an explicit processing of precision requirements, ideal to treat complex chains of geometric tolerances, and a global strategy guided by feature relations at the assembly level. The formalization of technical knowledge will have to be extended by rules for the selection of tolerance types and datums in typical design situations. The need to cover the whole set of tolerances provided by technical standards will probably require an extension of existing reasoning rules and classifications of tolerancing cases.

To extend the scope of specification methods, an integration with computer-aided tools for tolerance analysis will have to be ensured. The interface will have to involve an automated generation of geometric tolerance chains and related functional equations. This will allow a seamless integration of tolerancing activities and a significant efficiency gain in the design of mechanical assemblies.

References

Abdullah TA, Popplewell K, Page CJ (2003) A review of the support tools for the process of assembly method selection and assembly planning. Int J Prod Res 41(11):2391–2410

Anselmetti B (2006) Generation of functional tolerancing based on positioning features. Comput Aided Des 38:902–919

Anselmetti B, Mawussi K (2003) Computer aided tolerancing using positioning features. ASME J Comput Inf Sci Eng 3:15–21

ASME Y14.5.1 (1994) Dimensioning and tolerancing. American Society of Mechanical Engineers, New York

ASME Y14.5.1.M (1994) Mathematical definition of dimensioning and tolerancing principles. American Society of Mechanical Engineers, New York

Ballu A, Mathieu L (1999) Choice of functional specifications using graphs within the framework of education. In: Proceedings of the CIRP seminar on computer-aided tolerancing, Twente, pp 197–206

Bjørke O (1989) Computer-aided tolerancing. ASME Press, New York

Buysse D, Socoliuc M, Rivière A (2007) A new specification model for realistic assemblies simulation. In: Proceedings of the CIRP international conference on computer-aided tolerancing, Erlangen

Carr CD (1993) A comprehensive method for specifying tolerance requirements for assembly. ADCATS report 93-1

Chiabert P, Orlando M (2004) About a CAT model consistent with the ISO/TC 213 last issues. J Mater Proc Technol 157/158:61–66

Clément A, Rivière A (1993) Tolerancing versus nominal modeling in next generation CAD/CAM system. In: Proceedings of the CIRP seminar on computer-aided tolerancing, Cachan, pp 97–113

Clément A, Rivière A, Temmerman M (1994) Cotation tridimensionelle des systèmes mécaniques. PYC, Yvry-sur-Siene

Clément A, Rivière A, Serré P (1995) A declarative information model for functional requirements. In: Proceedings of the CIRP seminar on computer-aided tolerancing, Tokyo, pp 3–16

Cogorno GR (2006) Geometric dimensioning and tolerancing for mechanical design. McGraw-Hill, New York

Dantan JY, Anwer N, Mathieu L (2003a) Integrated tolerancing process for conceptual design. CIRP Ann 52(1):135–138

Dantan JY, Ballu A, Martin P (2003b) Mathematical formulation of tolerance synthesis taking into account the assembly process. In: Proceedings of the IEEE international symposium on assembly and task planning, Besançon, pp 241–246

Dantan JY, Mathieu L, Ballu A, Martin P (2005) Tolerance synthesis: quantifier notion and virtual boundary. Comput Aided Des 37:231–240

Desrochers A, Maranzana R (1995) Constrained dimensioning and tolerancing assistance for mechanisms. In: Proceedings of the CIRP seminar on computer-aided tolerancing, Tokyo, pp 17–30

Drake PJ (1999) Dimensioning and tolerancing handbook. McGraw-Hill, New York

Dufaure J, Teissandier D, Debarbouille G (2005) Influence of the standard components integration on the tolerancing activity. In: Proceedings of the CIRP international seminar on computer-aided tolerancing, Tempe, pp 235–244

Farmer LE, Gladman CA (1986) Tolerance technology: computer-based analysis. CIRP Ann 35(1):7–10

Giordano M, Pairel E, Hernandez P (2005) Complex mechanical structure tolerancing by means of hyper-graphs. In: Proceedings of the CIRP international seminar on computer-aided tolerancing Tempe, pp 105–114

Hong YS, Chang TC (2002) A comprehensive review of tolerancing research. Int J Prod Res 40(11):2425–2459

Hu J, Xiong G (2005a) Dimensional and geometric tolerance design based on constraints. Int J Adv Manuf Technol 26:1099–1108

Hu J, Xiong G (2005b) Concurrent design of a geometric parameter and tolerance for assembly and cost. Int J Prod Res 43(2):267–293

Hu J, Xiong G, Wu Z (2004) A variational geometric constraints network for a tolerance types specification. Int J Adv Manuf Technol 24:214–222

Islam MN (2004a) Functional dimensioning and tolerancing software for concurrent engineering applications. Comput Ind 54:169–190

Islam MN (2004b) A methodology for extracting dimensional requirements for a product from customer needs. Int J Adv Manuf Technol 23:489–494

ISO 1101 (2004) Geometrical product specifications (GPS) – geometrical tolerancing – tolerances of form, orientation, location and run-out. International Organization for Standardization, Geneva

Jayaraman R, Srinivasan B (1989) Geometric tolerancing: I. Virtual boundary requirements. IBM J Res Dev 33(2):90–104

Kandikjan T, Shah JJ, Davidson JK (2001) A mechanism for validating dimensioning and tolerancing schemes in CAD systems. Comput Aided Des 33:721–737

Kim KY, Wang Y, Mougboh OS, Nnaji BO (2004) Design formalism for collaborative assembly design. Comput Aided Des 36:849–871

McAdams DA (2003) Identification and codification of principles for functional tolerance design. J Eng Des 14(3):355–375

Meadows JD (1995) Geometric dimensioning and tolerancing: applications and techniques for use in design, manufacturing and inspection. Dekker, New York

Mejbri H, Anselmetti B, Mawussi B (2003) A recursive tolerancing method with sub-assembly generation. In: Proceedings of the IEEE international symposium on assembly and task planning, Besançon, pp 235–240

Mejbri H, Anselmetti B, Mawussi K (2005) Functional tolerancing of complex mechanisms: identification and specification of key parts. Comput Ind Eng 49:241–265

Mliki MN, Mennier D (1995) Dimensioning and functional tolerancing aided by computer in CAD/CAM systems: application to Autocad system. In: Proceedings of the INRIA/IEEE symposium on emerging technologies and factory automation, Paris, pp 421–428

Mullins SH, Anderson DC (1998) Automatic identification of geometric constraints in mechanical assemblies. Comput Aided Des 30(9):715–726

Narahari Y, Sudarsan R, Lyons KW et al (1999) Design for tolerance of electro-mechanical assemblies: an integrated approach. IEEE Trans Robot Autom 15(6):1062–1079

Nassef AO, ElMaraghy HA (1997) Allocation of geometric tolerances: new criterion and methodology. CIRP Ann 46(1):101–106

Pérez R, De Ciurana J, Riba C (2006) The characterization and specification of functional requirements and geometric tolerances in design. J Eng Des 17(4):311–324

Ramani B, Cheraghi SH, Twomey JM (1998) CAD-based integrated tolerancing system. Int J Prod Res 36(10):2891–2910

Roy U, Bharadwaj B (1996) Tolerance synthesis in a product design system. In: Proceedings of NAMRC, Ann Arbor

Salomons OW (1995) Computer support in the design of mechanical products – constraints specification and satisfaction in feature based design and manufacturing. PhD thesis, University of Twente

Salomons OW, Jonge Poerink HJ, Haalboom FJ et al (1996) A computer aided tolerancing tool I: tolerance specification. Comput Ind 31:161–174

Shah JJ, Yan Y, Zhang BC (1998) Dimension and tolerance modeling and transformations in feature based design and manufacturing. J Intell Manuf 9:475–488

Söderberg R, Johannesson H (1999) Tolerance chain detection by geometrical constraint based coupling analysis. J Eng Des 10(1):5–24

Srinivasan B, Jayaraman R (1989) Geometric tolerancing: II. Conditional tolerances. IBM J Res Dev 33(2):105–125

Sudarsan R, Narahari Y, Lyons KW et al (1998) Design for tolerance of electro-mechanical assemblies. In: Proceedings of the IEEE international conference on robotics and automation, Leuven, pp 1490–1497

Teissandier D, Dufaure J (2007) Specifications of a pre and post-processing tool for a tolerancing analysis solver. In: Proceedings of the CIRP international conference on computer-aided tolerancing, Erlangen

Toulorge H, Rivière A, Bellacicco A et al (2003) Towards a digital functional assistance process for tolerancing. ASME J Comput Inf Sci Eng 3:39–44

Voelcker HB (1998) The current state of affairs in dimensional tolerancing: 1997. Integr Manuf Syst 9(4):205–217

Wang H, Roy U, Sudarsan R et al (2003) Functional tolerancing of a gearbox. In: Proceedings of NAMRC, Hamilton

Wang H, Ning R, Yan Y (2006) Simulated tolerances CAD geometrical model and automatic generation of 3D tolerance chains. Int J Adv Manuf Technol 29:1019–1025

Weill R (1988) Tolerancing for function. CIRP Ann 37(2):603–610

Weill RD (1997) Dimensioning and tolerancing for function. In: Zhang HC (ed) Advanced tolerancing techniques. Wiley, New York

Whitney DE (1996) The potential for assembly modeling in product development and manufacturing. Technical report. The Center for Technology, Policy and Industrial Development, MIT, Cambridge

Whitney DE, Mantripragada R, Adams JD et al (1999) Designing assemblies. Res Eng Des 11:229–253

Willhelm RG, Lu SCY (1992) Tolerance synthesis to support concurrent engineering. CIRP Ann 41(1):197–200

Wu Y, Shah JJ, Davidson JK (2003) Computer modeling of geometric variations in mechanical parts and assemblies. ASME J Comput Inf Sci Eng 3:54–63

Zou Z, Morse EP (2004) A gap-based approach to capture fitting conditions for mechanical assembly. Comput Aided Des 36:691–700

Chapter 2
Geometric Tolerance Analysis

Wilma Polini

Abstract This chapter focuses on five main literature models of geometric tolerance analysis – *vector loop*, *variational*, *matrix*, *Jacobian*, and *torsor* – and makes a comparison between them in order to highlight the advantages and the weaknesses of each, with the goal of providing a criterion for selecting the most suitable one, depending on the application. The comparison is done at two levels: the first is qualitative and is based on the analysis of the models according to a set of descriptors derived from what is available in the literature; the second is quantitative and is based on a case study which is solved by means of the five models. Finally, in addition to providing comparative insight into the five tolerance analysis models, some guidelines are provided as well, related to the development of a novel approach which is aimed at overcoming some of the limitations of those models.

2.1 Introduction

Increasing competition in industry leads to the adoption of cost-cutting programs in the manufacturing, design, and assembly of products. Current products are complex systems, often made of several assemblies and subassemblies, and including complex part geometries; a wide variety of different requirements must be satisfied at the design and manufacturing stages in order for such products to fulfill the desired functional requirements. The paradigm of concurrent engineering enforces an

W. Polini
Dipartimento di Ingegneria Industriale, Università degli Studi di Cassino,
Via Gaetano di Biasio 43, 03043 Cassino, Italy.
e-mail: polini@unicas.it

approach where product design and manufacturing process planning activities are carried out in parallel in highly communicative and collaborative environments, with the aim of reducing reworking times and discard rates. In recent years, the importance of assessing the effects of part tolerances on assembly products has increasingly been acknowledged as one of the most strategic activities to be pursued with the goal of ensuring higher production qualities at lower costs. In fact, while the need for assigning some type of dimensional and geometric tolerances to assembly components is widely recognized as a necessary step for ensuring a standardized production process and for guaranteeing the correct working of the assembly to the required levels of satisfaction, the relationships between the values assigned to such tolerances and final product functionality are more subtle and need to be investigated in greater detail. As defined by designer intent, assembled products are built to satisfy one or more functional prerequisites. The degree to which each of such functional prerequisites is satisfied by the product is usually strongly related to a few key dimensions of the final assembly itself. In related literature, such key dimensions are often called *key functional parameters* or *design dimensions*, *or functional requirements*, which is the term that will be adopted in this chapter. Functional requirements are typically the result of the stack-up of geometries and dimensions of several parts; and their final variability, also fundamental in determining overall functional performance, arises from the combined effects of the variabilities associated with the parts involved in the stack-up. Tolerance values assigned to parts and subassemblies become critical in determining the overall variability of functional requirements and, consequently, the functional performance of the final product. Moreover, when analyzing the chain of connected parts, it becomes relevant to identify those tolerances that – more than others – have an influence on the final outcome, since it has been shown that often the *70–30 rule* applies, meaning that 30% of the tolerances assigned to the components are responsible for 70% of the assembly geometric variation. The variabilities associated with dimensions and geometries of the assembly components combine, according to the assembly cycle, and generate the variability associated with the functional requirements. *Tolerance stack-up functions* are mathematical models aimed at capturing such combinations. They have that name because they are designed as functions whose output is the variation range of a functional requirement, and whose inputs are the tolerances assigned to assembly components. In other words, a tolerance analysis problem implies modeling and solving a tolerance stack-up function, in order to determine the nominal value and the tolerance range of a functional requirement starting from the nominal values and the tolerance ranges assigned to the relevant dimensions of assembly components.

In a typical tolerance analysis scenario, linear or nonlinear tolerance stack-up functions may need to be solved, depending on the problem being studied, and on the way it was modeled in the analysis. Alternative assembly cycles may be considered within the analysis in order to identify the one allowing assembly functionality with the maximum value of the tolerance range assigned to the components. The importance of an effectual tolerance analysis is widely recognized: significant problems may arise during the actual assembly process if the tolerance analysis on a part or subassembly was not carried out or was done to an unsatisfactory level (Whitney 2004). It may even happen that the product is subjected to significant redesign because of unforeseen tolerance problems, which were not detected prior to actual assembly taking place. In this case business costs may be significantly high, especially when considering that 40–60% of the production costs are estimated as due to the assembly process (Delchambre 1996).

Many well-known approaches, or *models*, exist in the literature for tolerance analysis (Hong and Chang 2002; Shen *et al.* 2004). The *vector loop model* adopts a graph-like schematization where any relevant linear dimension in the assembly is represented by a vector, and an associated tolerance is represented as a small variation of such a vector (Chase *et al.* 1997; Chase 1999). Vectors are connected to form chains and/or loops, reflecting how assembly parts stack up together in determining the final functional requirements of the assembly. Stack-up functions are built by combining the variations associated with vectors involved in each chain into mathematical expressions, which can then be solved with different approaches.

The *variational model* has its roots in parametric geometric modeling, where geometry can be modeled by mathematical equations that allow shape and size attributes to be changed and controlled through a reduced set of parameters (Gupta and Turner 1993; Whitney *et al.* 1999; Li and Roy 2001). Parametric modeling can be used as the starting point for reproducing small variations in an assembly part, within the ranges defined by a given tolerance. In the *matrix model*, the approach aims at deriving the explicit mathematical representation of the geometry of each tolerance region (the portion of space where a feature is allowed to be, given a set of tolerances); this is done through displacement matrices, which describe the small displacements a feature is allowed to have without violating the tolerances (Desrochers and Rivière 1997; Clément *et al.* 1998).

In the *Jacobian model*, tolerance chains are modeled as sequences of connected pairs of relevant surfaces; displacement between such surfaces (whether nominal or due to variations allowed by tolerances) is modeled through homogeneous coordinate transformation matrices (Laperrière and Lafond 1999; Laperrière and Kabore 2001). The way the matrices are formulated draws inspiration from a common approach adopted in robotics which involves the use of Jacobian matrices.

Finally, in the *torsor model*, screw parameters are introduced to model 3D tolerance zones (Ballot and Bourdet 1995, 1997). The name derives from the data structure adopted to collect the screw parameters (*i.e.*, the torsor).

The five modeling approaches introduced above propose different solutions to specific aspects of the tolerance analysis problem; all have strong points and weaknesses that may make them inadequate for specific applications. Most aspects differentiating the approaches are related to how geometric variability of parts and assemblies is modeled, how joints and clearance between parts are represented, how stack-up functions are solved, and so forth. Moreover, it is difficult to find literature work where the different approaches are compared systematically with the help of one or more case studies aimed at highlighting the advantages and disadvantages of each.

Section 2.2 describes a practical case study, which will be used as a reference to illustrate the tolerance analysis models presented in Sections 2.3–2.5. Section 2.6 provides some guidelines for the development of a new model aimed at overcoming the weaknesses of the literature models.

2.2 The Reference Case Study

To compare the tolerance analysis models, the case study shown in Figure 2.1 is introduced. The 2D geometry of the example assembly is made of a rectangular box containing two disk-shaped parts. The width *g* of the gap between the top disk and the upper surface of the box is assumed as the functional requirement to be investigated by the analysis. The goal of the tolerance analysis problem is to identify the tolerance stack-up function that defines the variability of *g*, and describes it as a function of the geometries and tolerances of the components involved in the assembly.

Tolerance analysis is based on the dimensional and geometric tolerances illustrated in Figure 2.1. The example is adapted from a real-life industrial application and properly simplified to make it easier to present and discuss in this context. The tolerancing scheme applied, which may not appear as entirely rigorous under the

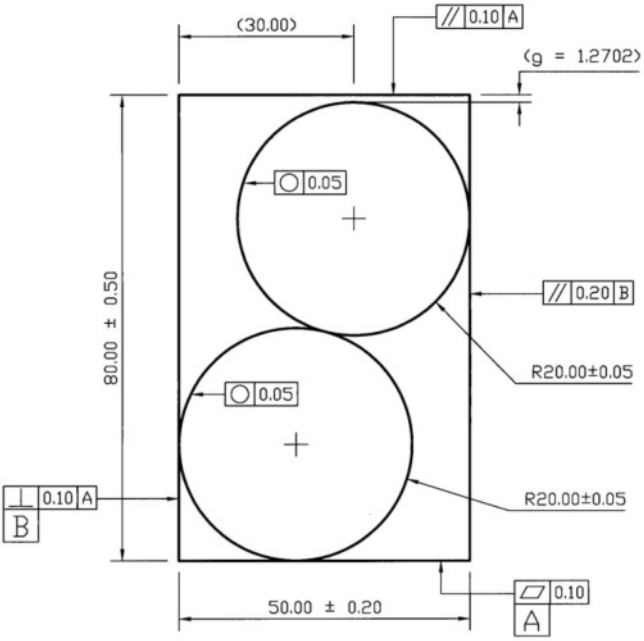


Figure 2.1 Dimensional and geometric tolerances applied to the case study

viewpoint of a strict application of standardized tolerancing rules, is directly derived from the current practice adopted for the actual industrial product.

The case study is representative of all the main aspects and critical issues involved in a typical tolerance analysis problem, and it is simple enough to allow for the application of a simplified manual computation procedure to obtain the extreme values of the gap *g* for the special case where *only dimensional tolerances are considered*. The manual computation is based on searching for the worst-case conditions, *i.e.*, the combinations of part dimensions that give rise to the maximum and minimum gap values; since no geometric tolerances are considered, part geometries are assumed at nominal states.

The maximum value of the gap is calculated by considering the maximum height and width of the box, together with the minimum value of the radius of the disks:

$$\begin{aligned} g_{\max_{\dim}} &= 80.5 - 19.95 - \sqrt{(19.95 \cdot 2)^2 - (50.40 - 19.95 - 19.95)^2} \\ &-19.95 = 2.1064\,\text{mm}. \end{aligned} \tag{2.1}$$

In the same way, the minimum value of the gap is

$$\begin{aligned} g_{\min_{\dim}} &= 79.5 - 20.05 - \sqrt{(20.05 \cdot 2)^2 - (49.8 - 20.05 - 20.05)^2} \\ &-20.05 = 0.4909\,\text{mm}. \end{aligned} \tag{2.2}$$

The variability of the gap is the difference between the maximum or the minimum values and the nominal one:

$$\begin{aligned} \Delta g_{\dim_1} &= +\left(g_{\max_{\dim}} - g_N\right) = (2.1064 - 1.2702) \cong +0.84\,\text{mm}, \\ \Delta g_{\dim_2} &= -\left(g_N - g_{\min_{\dim}}\right) = -(1.2702 - 0.4909) \cong -0.78\,\text{mm}. \end{aligned} \tag{2.3}$$

Albeit operating on a simplified problem (geometric tolerances are neglected) the manual computation of the gap boundary values provides a useful support for the quantitative comparison of the five methods, at least when they are applied by considering dimensional tolerances only. The manually obtained, extreme gap values will be used as reference values later on, then the results of the five methods will be discussed.

Furthermore, the manual computation procedure highlights one of the fundamental issues analyzed in this chapter, *i.e.*, how hard it actually is to include geometric tolerances in any model attempting to represent geometric variability. The investigation of how this challenge is handled in the analysis of tolerance chains is of fundamental importance when analyzing the performance and limitations of the five approaches: under this assumption, it will be shown how each approach provides a different degree of support to the inclusion of geometric tolerances, and how each requires different modeling efforts, simplifications, and workarounds, in order to let geometric tolerances be included in the tolerance chain analysis problem.

2.3 The Vector Loop Model

The vector loop model uses vectors to represent relevant dimensions in an assembly (Chase *et al.* 1995, 1996). Each vector represents either a component dimension or an assembly dimension. Vectors are arranged in chains or loops to reproduce the effects of those dimensions that stack together to determine the resultant assembly dimensions. Three types of variations are modeled in the vector loop model: *dimensional variations*, *kinematic variations*, and *geometric variations*.

In a vector loop model, the magnitude of a geometric dimension is mapped to the length (L_i) of the corresponding vector. Dimensional variations defined by dimensional tolerances are incorporated as ± variations in the length of the vector. Kinematic variations describe the relative motions among mating parts, *i.e.*, small adjustments that occur at assembly time in response to the dimensional and geometric variations of the components. In the vector loop model, kinematic variations are modeled by means of *kinematic joints*, *i.e.*, schematizations such as the *slider*. In vector loop models, there are six common joint types available for 2D assemblies and 12 common joints for 3D assemblies. At each kinematic joint, assembly adjustments are turned into ranges for the motions allowed by the joint (*i.e.*, degrees of freedom). A local datum reference frame (DRF) must be defined for each kinematic joint.

Geometric variations capture those variations that are imputable to geometric tolerances. These are modeled by adding additional degrees of freedom to the kinematic joints illustrated above. This introduces a simplification: although geometric tolerances may affect an entire surface, in vector loop models they are considered only in terms of the variations they induce at mating points, and only in the directions allowed by the type of kinematic joint. Depending on what type of geometric variation is represented by the tolerance and what motions are allowed at the kinematic joint, a geometric tolerance is typically modeled as an additional set of translational and rotational transformations (*e.g.*, displacement vectors, rotation matrices) to be added at the joint.

To better understand the vector loop model, the basic steps for applying it to a tolerance analysis problem are provided below (Gao *et al.* 1998; Faerber 1999; Nigam and Turner 1995):

1. *Create the assembly graph*. The first step is to create an assembly graph. The assembly graph is a simplified diagram of the assembly representing the parts, their dimensions, the mating conditions, and functional requirements, *i.e.*, the final assembly dimensions that must be measured in order to verify that the product is capable of providing the required functionality. An assembly graph assists in identifying the number of vector chains and loops involved in the assembly.
2. *Define the DRF for each part*. The next step is to define the DRF for each part. DRFs are used to locate relevant features on each part. If there is a circular contact surface, its center is considered as a DRF too.

3. *Define kinematic joints and create datum paths.* Each mating relation among parts is translated into a kinematic joint. Kinematic joints are typically located at contact points between parts. Datum paths are geometric layouts specifying the direction and orientation of vectors forming the vector loops; they are created by chaining together the dimensions that locate the point of contact of a part with another, with respect to the DRF of the part itself.
4. *Create vector loops.* With use of the assembly graph and the datum paths, vector loops are created. Each vector loop is created by connecting datums. Vector loops may be open or closed; an open loop terminates with a functional requirement, which can be measured in the final assembly (it could be either the size of a relevant gap in the final assembly, or any other functionally relevant assembly dimension); a closed loop indicates the presence of one or more adjustable elements in the assembly.
5. *Derive the stack-up equations.* The assembly constraints defined within vector-loop-based models may be mathematically represented as a concatenation of homogeneous rigid body transformation matrices:

$$\mathbf{R}_1 \cdot \mathbf{T}_1 \cdot \ldots \cdot \mathbf{R}_i \cdot \mathbf{T}_i \cdot \ldots \cdot \mathbf{R}_n \cdot \mathbf{T}_n \cdot \mathbf{R}_f = \mathbf{H}, \tag{2.4}$$

where $\mathbf{R}_1$ is the rotational transformation matrix between the x-axis and the first vector; $\mathbf{T}_1$ is the translational matrix for the first vector, $\mathbf{R}_{i,n}$ and $\mathbf{T}_{i,n}$ are the corresponding matrices for the vector at node i or node n, and $\mathbf{R}_f$ is the final closure rotation, again with respect to the x-axis. $\mathbf{H}$ is the resultant matrix. For example, in the 2D case the rotational and the translational matrices are as follows:

$$\mathbf{R}_i = \begin{bmatrix} \cos\phi_i & \sin\phi_i & 0 \\ \sin\phi_i & \cos\phi_i & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \mathbf{T}_i = \begin{bmatrix} 1 & 0 & L_i \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where ϕ_i is the angle between the vectors at node i, and L_i is the length of vector i. If the assembly is described by a closed loop of constraints, $\mathbf{H}$ is equal to the identity matrix, otherwise $\mathbf{H}$ is equal to the $\mathbf{g}$ vector representing the resultant transformation that will lead to the identification of a functional requirement.

6. *Tolerance analysis* – assuming an assembly as made of p parts. Each part is represented by an $\mathbf{x}$ vector of its relevant dimensions and by an $\boldsymbol{\alpha}$ vector containing additional dimensions, added to take into account geometric tolerances. When parts are assembled together, the resulting product is characterized by a $\mathbf{u}$ vector of the assembly variables and by a $\mathbf{g}$ vector of measurable functional requirements. It is possible to write $L = J - P + 1$ closed loops, where J is the number of the mates among the parts and P is the number of parts. For each closed loop

$$\mathbf{H}(x, u, \alpha) = 0, \tag{2.5}$$

while there is an open loop for each functional requirement that looks like

$$\mathbf{g} = \mathbf{K}(x,u,\alpha)\,. \tag{2.6}$$

Equation 2.5 allows one to calculate **g** after having solved the system of equations in Equation 2.4. The equations in Equation 2.4 are usually not linear; they can be solved in different ways, for example, by means of the direct linearization method:

$$\mathbf{dH} \cong \mathbf{A}\cdot\mathbf{dx} + \mathbf{B}\cdot\mathbf{du} + \mathbf{F}\cdot\mathbf{d\alpha} = 0\,, \tag{2.7}$$

$$\mathbf{dH} \cong -\mathbf{B}^{-1}\cdot\mathbf{A}\cdot\mathbf{dx} - \mathbf{B}^{-1}\cdot\mathbf{F}\cdot\mathbf{d\alpha}\,, \tag{2.8}$$

$$\mathbf{dg} \cong \mathbf{C}\cdot\mathbf{dx} + \mathbf{D}\cdot\mathbf{du} + \mathbf{G}\cdot\mathbf{d\alpha} = 0\,, \tag{2.9}$$

with $\mathbf{A}_{ij} = \partial\mathbf{H}_i / \partial x_j$, $\mathbf{B}_{ij} = \partial\mathbf{H}_i / \partial u_j$, $\mathbf{F}_{ij} = \partial\mathbf{H}_i / \partial\alpha_j$, $\mathbf{C}_{ij} = \partial\mathbf{K}_i / \partial x_j$, $\mathbf{D}_{ij} = \partial\mathbf{K}_i / \partial u_j$, and $\mathbf{G}_{ij} = \partial\mathbf{K}_i / \partial\alpha_j$.

From Equations 2.7–2.9,

$$\begin{aligned}\mathbf{dg} &\cong [\mathbf{C} - \mathbf{D}\cdot\mathbf{B}^{-1}\cdot\mathbf{A}]\cdot\mathbf{dx} + [\mathbf{G} - \mathbf{D}\cdot\mathbf{B}^{-1}\cdot\mathbf{F}]\cdot\mathbf{d\alpha}\\ &= \mathbf{S_x}\cdot\mathbf{dx} + \mathbf{S_\alpha}\cdot\mathbf{d\alpha},\end{aligned} \tag{2.10}$$

where $\mathbf{S_x} = [\mathbf{C} - \mathbf{D}\cdot\mathbf{B}^{-1}\cdot\mathbf{A}]$ and $\mathbf{S_\alpha} = [\mathbf{G} - \mathbf{D}\cdot\mathbf{B}^{-1}\cdot\mathbf{F}]$ are named the "*sensitivity*" matrices. When the sensitivity matrices are known, it is possible to calculate the solution in the worst-case scenario as

$$\Delta g_i = \sum\nolimits_k |\,\mathbf{S}_{x_{ik}}\cdot\mathbf{t}_{x_k}\,| + \sum\nolimits_l |\,\mathbf{S}_{\alpha_{il}}\cdot\mathbf{t}_{\alpha_l}\,|, \tag{2.11}$$

while in the statistical scenario the solution can be obtained as a root sum of squares, as follows:

$$\Delta g_i = \left[\sum\nolimits_k (\mathbf{S}_{x_{ik}}\cdot\mathbf{t}_{x_k})^2 + \sum\nolimits_l (\mathbf{S}_{\alpha_{il}}\cdot\mathbf{t}_{\alpha_l})^2\right]^{1/2}, \tag{2.12}$$

where k and l are the number of x dimensions and α geometric tolerances that influence the variable g_i, $\mathbf{S}_{xik}$ is the matrix of the coefficients of the k x variables inside the i-stack-up function of Equation 2.10, $\mathbf{S}_{\alpha il}$ is the matrix of the coefficients of the l α variables inside the i stack-up function of Equation 2.10, and $\mathbf{t}_{xk}$ and $\mathbf{t}_{\alpha l}$ are the vectors of the dimensional or the geometric tolerances of the xk and αl variables, respectively.

The direct linearization method is a very simple and rapid method, but it is approximated too. When an approximated solution is not acceptable, it is possible to use alternative approaches, such as numerical simulation by means of a Monte Carlo technique (Gao *et al.* 1998; Boyer and Stewart 1991).

2.3.1 Results of the Case Study with Dimensional Tolerances

With reference to Figure 2.2, let x_1 and x_2 be the dimensions of the box, and x_3 and x_4 the diameters of the two disks; u_1, u_2, u_3, and u_4 are the assembly (dependent) dimensions and g is the width of the gap between the top side of the box and the second disk. The dimension g is the functional requirement. Therefore, the assembly graph in Figure 2.3 has been built. It shows two joints of "*cylinder slider*" kind between the box and disk 1 at point A and point B, respectively; one joint of "*parallel cylinder*" kind between disk 1 and disk 2 at point C; one joint of "*cylinder slider*" kind between disk 2 and the box at point D; and the measurement to be performed (g).

DRFs have been assigned to each part; they are centered at point Ω for the box and at the centers O_1 and O_2 of the two disks. All the DRFs have a horizontal x-axis. Datum A has been assumed to be nominal. The DRF of the box is also assumed as the global DRF of the assembly. Figure 2.4 shows the created datum paths that chain together the points of contact of a part with another with respect to the DRF of the part itself. Vector loops are created using the datum paths as a guide. There are $L=J-P+1=4-3+1=2$ closed loops and one open loop. The first closed loop joins the box and disk 1 passing through contact points A and B.

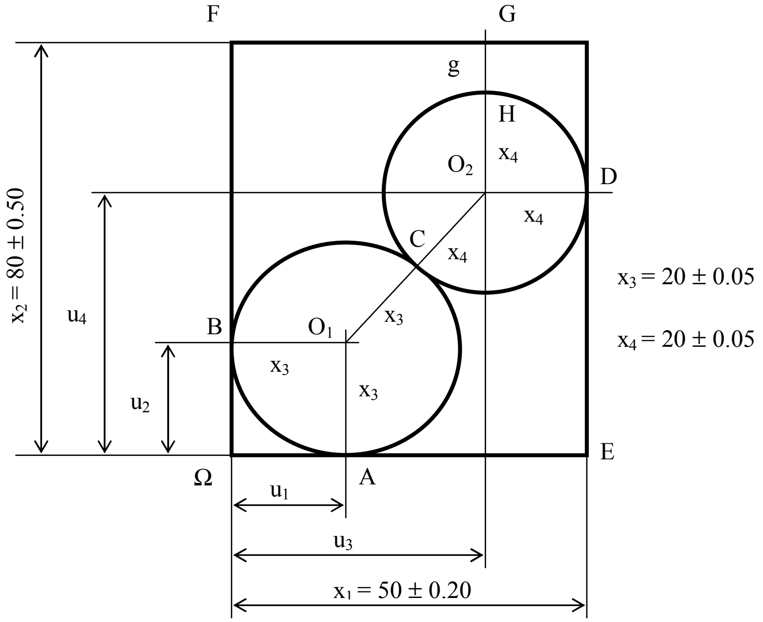


Figure 2.2 Assembly variables and tolerances of the vector loop model with dimensional tolerances

The second closed loop joins the subassembly box disk 1 and disk 2 through contact points D and C. The open loop defines the gap width *g*. All the loops are defined counterclockwise. The **R** and **T** matrices are 2D; their elements are shown in Table 2.1.

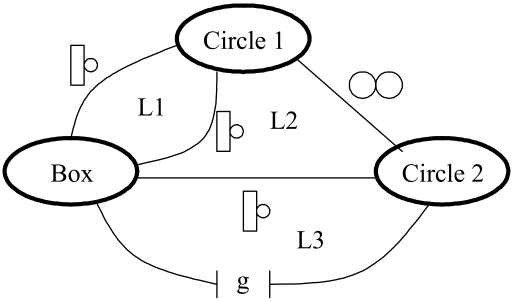


Figure 2.3 Assembly graph

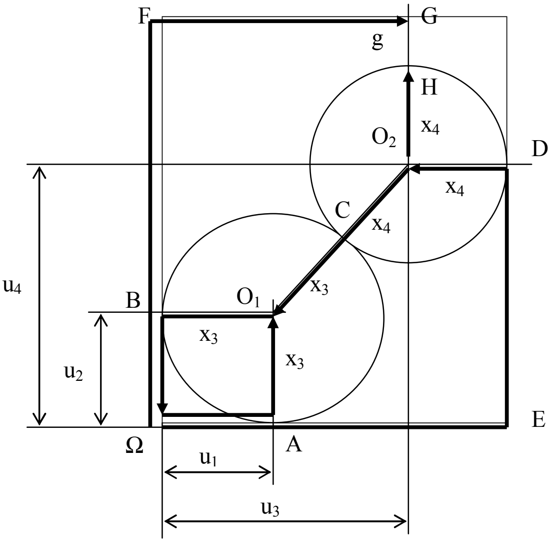


Figure 2.4 Datum paths of the vector loop model

Table 2.1 Elements of **R** and **T** matrices when the case study considers dimensional tolerances

	Loop 1		Loop 2		Loop 3	
i	ϕ_i of R_i	L_i of T_i	ϕ_i of R_i	L_i of T_i	ϕ_i of R_i	L_i of T_i
1	0	u_1	0	x_1	0	x_1
2	90°	x_3	90°	u_4	90°	u_4
3	ϕ_{13}	x_3	90°	x_4	90°	x_4
4	90°	u_2	ϕ_{24}	x_4	ϕ_{34}	x_4
5	90°		0°	x_3	0°	g
6			ϕ_{26}	x_3	90°	u_3
7			90°	u_2	90°	x_2
8			90°		90°	

For the first loop, Equation 2.4 becomes

$$\mathbf{R}_1 \cdot \mathbf{T}_1 \cdot \mathbf{R}_2 \cdot \mathbf{T}_2 \cdot \mathbf{R}_3 \cdot \mathbf{T}_3 \cdot \mathbf{R}_4 \cdot \mathbf{T}_4 \cdot \mathbf{R}_f = \mathbf{I}, \tag{2.13}$$

which gives the system

$$\begin{aligned} & u_1 + x_3\cos(90+\varphi_{13}) + u_2\cos(180+\varphi_{13}) = 0, \\ & x_3 + x_3\sin(90+\varphi_{13}) + u_2\sin(180+\varphi_{13}) = 0, \\ & \varphi_{13} - 90 = 0. \end{aligned} \tag{2.14}$$

For the second loop,

$$\mathbf{R}_1 \cdot \mathbf{T}_1 \cdot \mathbf{R}_2 \cdot \mathbf{T}_2 \cdot \mathbf{R}_3 \cdot \mathbf{T}_3 \cdot \mathbf{R}_4 \cdot \mathbf{T}_4 \cdot \mathbf{R}_5 \cdot \mathbf{T}_5 \cdot \mathbf{R}_6 \cdot \mathbf{T}_6 \cdot \mathbf{R}_7 \cdot \mathbf{T}_7 \cdot \mathbf{R}_f = \mathbf{I}, \tag{2.15}$$

which gives the system

$$\begin{aligned} & x_1 - x_4 + x_4\cos(180+\varphi_{24}) + x_3\cos(180+\varphi_{24}) + \\ & + x_3\cos(180+\varphi_{24}+\varphi_{26}) = 0, \\ & x_1 - x_4 + x_4\cos(180+\varphi_{24}) + x_3\cos(180+\varphi_{24}) + \\ & + x_3\cos(180+\varphi_{24}+\varphi_{26}) = 0, \\ & \varphi_{24} + \varphi_{26} = 0. \end{aligned} \tag{2.16}$$

For the third loop,

$$\mathbf{R}_1 \cdot \mathbf{T}_1 \cdot \mathbf{R}_2 \cdot \mathbf{T}_2 \cdot \mathbf{R}_3 \cdot \mathbf{T}_3 \cdot \mathbf{R}_4 \cdot \mathbf{T}_4 \cdot \mathbf{R}_5 \cdot \mathbf{T}_5 \cdot \mathbf{R}_6 \cdot \mathbf{T}_6 \cdot \mathbf{R}_7 \cdot \mathbf{T}_7 \cdot \mathbf{R}_f = \mathbf{G}, \tag{2.17}$$

which gives

$$g = x_2 - u_4 - x_4. \tag{2.18}$$

From the "*sensitivity*" analysis,

$$\mathbf{A} \cdot \mathbf{dx} + \mathbf{B} \cdot \mathbf{du} = 0\,, \tag{2.19}$$

which gives

$$\mathbf{du} = -\mathbf{B}^{-1} \cdot \mathbf{A} \cdot \mathbf{dx} = \mathbf{S}^u \cdot \mathbf{dx}\,, \tag{2.20}$$

where

$$\mathbf{du} = \{du_1, du_2, du_4, d\varphi_{13}, d\varphi_{24}, d\varphi_{26}\}^{\mathrm{T}}, \tag{2.21}$$

$$\mathbf{dx} = \{dx_1, dx_2, dx_3, dx_4\}^{\mathrm{T}} = \{0.20, 0.50, 0.05, 0.05\}^{\mathrm{T}}, \tag{2.22}$$

$$\mathbf{S}^{\mathrm{u}} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ -0.2582 & 0 & 2.2910 & 1.2910 \\ 0 & 0 & 0 & 0 \\ -0.0258 & 0 & 0.0323 & 0.0323 \\ 0.0258 & 0 & -0.0323 & -0.0323 \end{bmatrix}. \tag{2.23}$$

The gap g depends on the following x variables through the sensitivity coefficients:

$$\begin{aligned} dg &= dx_2 - dx_4 - du_4 = 0.2582 \cdot dx_1 + dx_2 - 2.2910 \cdot dx_3 \\ &-2.2910 \cdot dx_4. \end{aligned} \tag{2.24}$$

It is possible to calculate the solution in the worst case as

$$\Delta g_{WC} = \pm \sum | S_i | \cdot \Delta x_i = \pm 0.7807 \cong \pm 0.78 \,\mathrm{mm}. \tag{2.25}$$

The solution obtained is lower than the value obtained by means of the manual resolution method of about 4% [= (1.56–1.62)/1.62].

It is possible to calculate the solution in the statistical scenario (root sum of square) as

$$\Delta g_{Stat} = \pm \left[\sum \left(S_{x_{ik}} \cdot t_{x_k} \right)^2 \right]^{1/2} = \pm 0.5158 \cong \pm 0.52 \,\mathrm{mm}. \tag{2.26}$$

2.3.2 Results of the Case Study with Geometric Tolerances

With reference to Figure 2.5, let x_1 and x_2 be the dimensions of the box, and x_3 and x_4 the diameters of the two disks; u_1, u_2, u_3, and u_4 are the assembly (dependent) dimensions and g is the width of the gap between the top side of the box and the second disk. The DRFs and the datum paths are the same as in the previous case (see Section 2.3.1).

The vector loops are the same as in the previous case, but they have to take into consideration the geometric tolerances. To include geometric tolerances, the following variables must be added to the **x** vector (note that mating points A and B are named after datums A and B the points lie upon):

- The flatness tolerance applied to the bottom surface of the box (datum A in the drawing) can be represented as a translation of the A point in the direction perpendicular to datum A, *i.e.*, perpendicular to the *x*-axis; this translation is described by the variable $\alpha_1 = T_{A1} = 0 \pm 0.10/2 = 0 \pm 0.05$ mm.
- The perpendicularity applied to the vertical left surface of the box (datum B) can be represented as a translation of point B in the direction perpendicular to datum B (the *y*-axis); this translation is described by the variable $\alpha_2 = T_{B2} = 0 \pm 0.10/2 = 0 \pm 0.05$ mm.
- The parallelism applied to the right side of the box (with respect to datum B) can be represented as a translation of point D, again in the direction perpendicular to datum B. It can be described by the variable $\alpha_3 = T_{D3} = 0 \pm 0.20/2 = 0 \pm 0.10$ mm.
- The circularity applied to disk 1 can be seen as points A, B, and C translating along the radius, and can be described by the variables $\alpha_4 = T_{A4} = 0 \pm 0.05/2 = 0 \pm 0.025$ mm, $\alpha_5 = T_{B4} = 0 \pm 0.05/2 = 0 \pm 0.025$ mm, and $\alpha_6 = T_{C4} = 0 \pm 0.05/2 = 0 \pm 0.025$ mm.
- The circularity applied to disk 2 can be represented as points C, D, and H translating along the radius, and can be described by the variables $\alpha_7 = T_{C5} = 0 \pm 0.05/2 = 0 \pm 0.025$ mm, $\alpha_8 = T_{D5} = 0 \pm 0.05/2 = 0 \pm 0.025$ mm, and $\alpha_9 = T_{H5} = 0 \pm 0.05/2 = 0 \pm 0.025$ mm.
- The parallelism applied to the top side of the box can be represented as a vertical translation of point G, and is described by the variable $\alpha_{10} = T_{G6} = 0 \pm 0.10/2 = 0 \pm 0.05$ mm.

The **R** and **T** matrices are 2D; their elements are shown in Table 2.2.

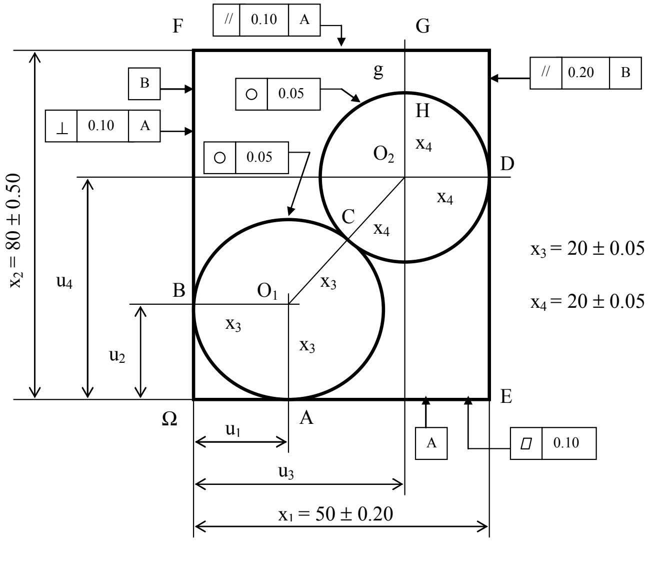


Figure 2.5 Assembly variables and tolerances of the vector loop model with geometric tolerances

Table 2.2 Elements of **R** and **T** matrices when the case study considers geometric tolerances

	Loop 1		Loop 2		Loop 3	
i	ϕ_i of R_i	L_i of T_i	ϕ_i of R_i	L_i of T_i	ϕ_i of R_i	L_i of T_i
1	0	u_1	0	x_1	0°	x_1
2	90°	$\alpha_1 = 0 \pm 0.05$	90°	u_4	90°	u_4
3	0°	$\alpha_4 = 0 \pm 0.025$	90°	$\alpha_3 = 0 \pm 0.1$	90°	$\alpha_3 = 0 \pm 0.1$
4	0°	x_3	0°	$\alpha_8 = 0 \pm 0.025$	0°	$\alpha_8 = 0 \pm 0.025$
5	ϕ_{13}	x_3	0°	x_4	0°	x_4
6	0°	$\alpha_5 = 0 \pm 0.025$	ϕ_{24}	x_4	ϕ_{34}	x_4
7	0°	$\alpha_2 = 0 \pm 0.05$	0°	$\alpha_7 = 0 \pm 0.025$	0°	$\alpha_9 = 0 \pm 0.025$
8	90°	u_2	0°	$\alpha_6 = 0 \pm 0.025$	0°	g
9	90°		0°	x_3	0°	$\alpha_{10} = 0 \pm 0.05$
10			ϕ_{26}	x_3		
11			0°	$\alpha_5 = 0 \pm 0.025$		
12			0°	$\alpha_2 = 0 \pm 0.05$		
13			90°	u_2		
14			90°			

Once the vector loops have been generated, the relative equations can be defined and solved. For the first loop, Equation 2.4 becomes

$$\begin{aligned}&\mathbf{R}_1 \cdot \mathbf{T}_1 \cdot \mathbf{R}_2 \cdot \mathbf{T}_2 \cdot \mathbf{R}_3 \cdot \mathbf{T}_3 \cdot \mathbf{R}_4 \cdot \mathbf{T}_4 \cdot \mathbf{R}_5 \cdot \mathbf{T}_5 \cdot \mathbf{R}_6 \cdot \mathbf{T}_6 \cdot \mathbf{R}_7 \cdot \mathbf{T}_7 \cdot \mathbf{R}_8 \cdot \mathbf{T}_8 \cdot\\ &\mathbf{R}_f = \mathbf{I},\end{aligned} \tag{2.27}$$

which gives the system

$$\begin{aligned}&u_1 + (x_3 + \alpha_2 + \alpha_5) \cdot \cos(90 + \varphi_{13}) + u_2 \cos(180 + \varphi_{13}) = 0,\\ &x_3 + \alpha_1 + \alpha_4 + (x_3 + \alpha_2 + \alpha_5) \cdot \sin(90 + \varphi_{13}) + u_2 \sin(180 + \varphi_{13}) = 0,\\ &\varphi_{13} - 90 = 0.\end{aligned} \tag{2.28}$$

For the second loop,

$$\begin{aligned}&\mathbf{R}_1 \cdot \mathbf{T}_1 \cdot \mathbf{R}_2 \cdot \mathbf{T}_2 \cdot \mathbf{R}_3 \cdot \mathbf{T}_3 \cdot \mathbf{R}_4 \cdot \mathbf{T}_4 \cdot \mathbf{R}_5 \cdot \mathbf{T}_5 \cdot \mathbf{R}_6 \cdot \mathbf{T}_6 \cdot \ldots \cdot \mathbf{R}_{13} \cdot \mathbf{T}_{13} \cdot\\ &\mathbf{R}_f = \mathbf{I},\end{aligned} \tag{2.29}$$

which gives the system

$$\begin{aligned}&x_1 - (x_4 + \alpha_3 + \alpha_8) + (x_3 + x_4 + \alpha_6 + \alpha_7) \cdot \cos(180 + \varphi_{24}) +\\ &(x_3 + \alpha_2 + \alpha_5) \cdot \cos(180 + \varphi_{24} + \varphi_{26}) = 0,\\ &u_4 + (x_3 + x_4 + \alpha_6 + \alpha_7) \cdot \sin(180 + \varphi_{24}) + (x_3 + \alpha_2 + \alpha_5) \cdot\\ &\sin(180 + \varphi_{24} + \varphi_{26}) - u_2 = 0,\\ &\varphi_{24} + \varphi_{26} = 0.\end{aligned} \tag{2.30}$$

For the third loop,

$$\begin{aligned}&\mathbf{R}_1 \cdot \mathbf{T}_1 \cdot \mathbf{R}_2 \cdot \mathbf{T}_2 \cdot \mathbf{R}_3 \cdot \mathbf{T}_3 \cdot \mathbf{R}_4 \cdot \mathbf{T}_4 \cdot \mathbf{R}_5 \cdot \mathbf{T}_5 \cdot \mathbf{R}_6 \cdot \mathbf{T}_6 \cdot \ldots \cdot \mathbf{R}_{11} \cdot \mathbf{T}_{11} \cdot \\ &\mathbf{R}_f = \mathbf{G},\end{aligned} \tag{2.31}$$

which gives

$$g = x_2 + \alpha_{10} - u_4 - x_4 - \alpha_9. \tag{2.32}$$

Concerning the “*sensitivity*” analysis,

$$\mathbf{du} = -\mathbf{B}^{-1} \cdot \mathbf{A} \cdot \mathbf{dx} - \mathbf{B}^{-1} \cdot \mathbf{C} \cdot \mathbf{d\alpha} = \mathbf{S}^{ud} \cdot \mathbf{dx} + \mathbf{S}^{u\alpha} \cdot \mathbf{d\alpha}, \tag{2.33}$$

where

$$\mathbf{du} = \left\{du_1, du_2, du_4, d\varphi_{13}, d\varphi_{24}, d\varphi_{26}\right\}^{\mathrm{T}}, \tag{2.34}$$

$$\mathbf{dx} = \left\{dx_1, dx_2, dx_3, dx_4\right\}^{\mathrm{T}} = \left\{0.20, 0.50, 0.05, 0.05\right\}^{\mathrm{T}}, \tag{2.35}$$

$$\mathbf{d\alpha} = \left\{d\alpha_1, \ldots, d\alpha_{10}\right\}^{\mathrm{T}} = \left\{\begin{matrix} 0.05, 0.05, 0.10, 0.025, 0.025, 0.025, \\ 0.025, 0.025, 0.025, 0.05 \end{matrix}\right\}^{\mathrm{T}}, \tag{2.36}$$

$$\mathbf{S}^{ud} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ -0.2582 & 0 & 2.2910 & 1.2910 \\ 0 & 0 & 0 & 0 \\ -0.0258 & 0 & 0.0323 & 0.0323 \\ 0.0258 & 0 & -0.0323 & -0.0323 \end{bmatrix}, \tag{2.37}$$

$$\mathbf{S}^{u\alpha} = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0.2582 & 0 & 0.0258 & -0.0258 \\ 0 & 0 & 0.2582 & 0 & 0.0258 & -0.0258 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0.2582 & 0 & 0.0258 & -0.0258 \\ 0 & 0 & 1.0328 & 0 & 0.0064 & -0.0064 \\ 0 & 0 & 1.0328 & 0 & 0.0064 & -0.0064 \\ 0 & 0 & 0.2582 & 0 & 0.0258 & -0.0258 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^{\mathrm{T}}. \tag{2.38}$$

The solution shows that the variability of the gap width g can be represented by the following function of the **x** vector:

$$\begin{aligned}&dg = dx_2 + d\alpha_{10} - dx_4 - du_4 - d\alpha_9 = 0.2582 \cdot dx_1 + dx_2\\&-2.2910 \cdot dx_3 - 2.2910 \cdot dx_4 - d\alpha_1 - 0.2582 \cdot d\alpha_2 - 0.2582 \cdot d\alpha_3\\&-d\alpha_4 - 0.2582 \cdot d\alpha_5 - 1.0328 \cdot d\alpha_6 - 1.0328 \cdot d\alpha_7 - 0.2582 \cdot d\alpha_8\\&-d\alpha_9 + d\alpha_{10}\end{aligned} \tag{2.39}$$

It is possible to compute the solution with the worst-case approach:

$$\Delta g_{WC} = \pm\left(\sum |S_i| \cdot \Delta x_i + \sum |S_j| \cdot \Delta\alpha_j\right) = \pm 1.0340 \cong \pm 1.03\,\text{mm}. \tag{2.40}$$

It is also possible to compute the solution with the statistical approach (root sum of squares):

$$\Delta g_{Stat} = \pm\left[\sum\left(S_{x_{ik}} \cdot t_{x_k}\right)^2 + \sum\left(S_{\alpha_{ij}} \cdot \Delta\alpha_j\right)^2\right]^{1/2} = \pm 0.5361 \cong \pm 0.54\,\text{mm}. \tag{2.41}$$

2.4 Further Geometric Tolerance Analysis Models

2.4.1 The Variational Model

A mathematical foundation of this model was proposed first by Boyer and Stewart (1991), and then by Gupta and Turner (1993). Later, several additional variants were proposed as well, and nowadays commercial computer aided tolerancing (CAT) software packages are based on this approach, such as eM-TolMate from UGS®, 3-DCS from Dimensional Control Systems®, and VisVSA from UGS®.

The basic idea of the variational model is to represent the variability of an assembly, due to tolerances and assembly conditions, through a parametric mathematical model.

To create an assembly, the designer must define the nominal shape and the dimensions of each assembly component (this information is usually retrieved from CAD files). Then, the designer identifies the relevant features of each component and assigns dimensional and geometric tolerances to them. Each feature has its local DRF, while each component and the whole assembly have their own global DRF. In nominal conditions, a homogeneous transformation matrix (called **TN**) is defined that identifies the position of the feature DRF with respect to the part DRF. In real conditions (*i.e.*, manufactured part), the feature will be characterized by a roto-translational displacement with respect to its nominal position. This displacement is modeled to summarize the complete effects of the dimensional and geometric variations affecting the part by means of another matrix: the differential homogeneous transformation matrix (called **DT**). The variational model may take into account the precedence among the datums by setting the parameters of the **DT** matrix.

The variational model is not able to deal with the form tolerances such as the vector loop model does; this means that the actual feature shape is assumed unchanged, *i.e.*, feature shape variations are neglected. The position of the displaced feature in the part DRF can be simply obtained by matrix multiplication as a change of DRF.

The model is parametric because different types and amounts of variations can be modeled by simply altering the contents (parameters) of the **DT** matrix. In some cases, the localization of a feature affected by a variation may be defined by a transformation with respect to another feature in the same part which is affected by variations as well. Therefore, the material modifier condition is modeled by setting the parameters of the **DT** matrix.

Once the variabilities of the parts have been modeled, they must be assembled together. Another set of differential homogeneous transformation matrices is introduced to handle the roto-translational deviations introduced by each assembly mating relation. Such matrices are named **DA**, with the letter *A* (for "assembly") to distinguish them from the matrices that have been used for parts. Those matrices are hard to evaluate, since they depend on both the tolerances imposed on the parts in contact and the assembly conditions. This model is not able to represent mating conditions with clearance. The problem of evaluating the differential matrix is analyzed in several literature works. A possible strategy consists in modeling the joint between the coupled parts by reconstructing the coupling sequence between the features (Berman 2005). Another possibility is to impose some analytical constraints on the assembly parameters (Whitney 2004).

When all the transformation matrices have been obtained, it is possible to express all the features in the same global DRF of the assembly. Finally, the functional requirements can be modeled in the form of functions, as follows:

$$\mathrm{FR} = f\left(p_1, p_2, \ldots, p_n\right), \tag{2.42}$$

where FR is the assembly functional requirement, $p_1, \ldots, p_n$ are the model parameters, and $f(p)$ is the stack-up function (usually not linear) obtained from the matrix multiplications described above. This model may be applied to assemblies involving joints which make a linear structure among the parts (linear stack-up function, see Figure 2.6a) and joints which make a complex structure among the parts (networks of stack-up functions, see Figure 2.6b), such as a vector loop does.

Once the stack-up functions have been modeled, there are two approaches to solve them: the worst-case approach and the statistical approach. The worst-case analysis consists in identifying the extreme configurations of the assembly under a given set of tolerances. In the variational approach, the problem is generally handled as an optimization (maximization and/or minimization) problem, under constraints defined by the tolerances themselves. The statistical approach is generally handled by assigning predefined probability density functions, *e.g.*, Gaussian, to the parameters identifying the main elements that contribute to the variation of each feature (often assumed independent, by simplification), and then solving the stack-up functions accordingly (Salomons *et al.* 1996).

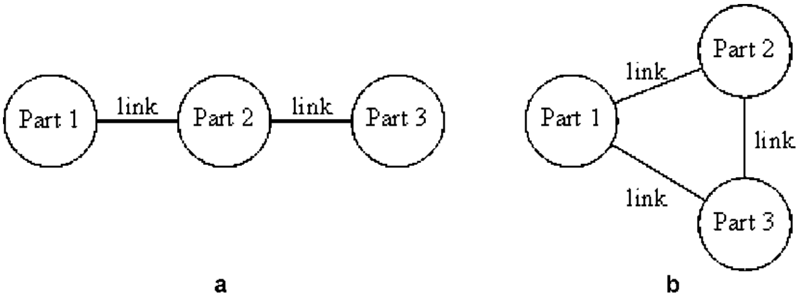


Figure 2.6 **a** Linear stack-up function, and **b** network stack-up function

To better illustrate the variational method, its basic steps are illustrated in the following:

1. *Create the assembly graph*. The first step is to create an assembly graph. The assembly graph is a simplified diagram of the assembly representing the parts, the features, the mating conditions, and the functional requirements.
2. *Define the DRF of each feature, of each part, and of the assembly*. The next step is to identify the local DRF of each feature and the global DRF of each part and of the assembly (usually the DRF of the assembly coincides with the DRF of the first part). DRFs are positioned depending on the surface type; from the DRFs, local parameters and the differential homogeneous transformation matrices **DT** are defined.
3. *Transform the features*. Once the transformation matrices are known, each feature of a part is transformed into the global DRF of the part.
4. *Create the assembly*. With use of the assembly graph and the transformed features, the assembly conditions are extracted, *i.e.*, the assembly parameters included in the matrix **DA** are calculated.
5. *Derive the equations of the functional requirements*. Once the assembly parameters are known, all the features can be expressed in the same global DRF of the assembly. At this point, the functional requirements are defined in terms of functions that can be solved by means of the previously described worst-case and/or statistical approaches.

2.4.2 The Matrix Model

Instead of deriving equations that model a specific displacement of a part or assembly as a function of given set of geometric dimensions (parameters) assuming specific values within the boundaries defined by tolerances (like in the variational approach), the matrix model aims at deriving an explicit mathematical representation of the boundary of the entire spatial region that encloses all possible displacements due to one or more variability sources. In order to do that, homogeneous transformation matrices are again considered as the foundation of the mathematical representation. A displacement matrix **DT** is used to describe any roto-translational variation a feature may be subjected to; the matrix is defined with respect to a local DRF. Since the goal is to represent the boundaries of the region of possible variations (*i.e.*, extreme values), the approach is intrinsically a worst-case approach. No statistical approach may be implemented, such as vector loop and variational models do. To represent boundaries, constraints must be added to the displacements modeled within the **DT** matrices. Displacement boundaries resulting from complex series of tolerances are solved by modeling the effects of each tolerance separately and by combining the resulting regions. Analogously, gaps/clearances are represented as if they were tolerance regions. Finally, by classifying the surfaces into several classes, each characterized by some type of invariance with respect to specific displacement types (*e.g.*, a cylinder is invariant to any rotation about its axis), one can simplify displacements and the resulting displacement matrix (Clément *et al.* 1994).

A similar approach is followed to model the dimensions acting as functional requirements of the assembly; since in this case the resulting region (of possible values) is essentially contained in a segment, segment boundaries must be computed by means of a worst-case approach (minimum–maximum distances between the two points). The two points defining the boundaries of the segment must be defined as the result of stack-up functions (Desrochers and Rivière 1997).

The matrix model is based on the positional tolerancing and the technologically and topologically related surfaces (TTRS) criteria (Clément *et al.* 1998). Geometric features are assumed as ideal, *i.e.*, the form tolerances are neglected, such as in the variational model. To better understand the matrix method for tolerance analysis, its basic steps are provided below:

1. *Transform the tolerances applied to the drawing.* The first step is to transform the tolerances applied to the drawing to make them compliant with the positional tolerancing and the TTRS criteria.
2. *Create the assembly graph.* The second step is to create an assembly graph. The assembly graph allows for identification of the global DRF and the linkages among the features to which the tolerances are assigned. The assembly parts should be in contact; the joints with clearance may not be considered.
3. *Define the local DRF of each part feature.* A DRF must be assigned to each part feature.
4. *Identify the measurable points for each functional requirement.* Points that locate the boundaries of each functional requirement must be identified and the path that connects them to the global DRF must be defined, taking into account all the tolerances stacking up along the way.
5. *Define the contributions of each single displacement and the related constraints.* It is necessary to define the contribution of each displacement to the total displacement region, and the constraints necessary to identify its boundaries. Each surface can be classified into one of the seven classes of invariant surfaces; this allows one to discard some displacements and to obtain a simpli-

fied displacement matrix. Additional information is necessary to specify the constraints ensuring that the feature remains inside the boundaries of the tolerance zone.

6. *Apply the superimposition principle and run the optimization*. If more than one tolerance is applied to the same part, the total effect is computed through the superimposition principle. For example, if *n* tolerances are applied to the same feature, in the local DRF, the displacement of a generic point belonging to the feature is simply defined as a sum of the single contributions. The aggregation of expressions obtained for each toleranced feature results in a constrained optimization problem, which can be solved with known, standard approaches. This model has been developed for assemblies involving joints which make a linear structure among the parts (linear stack-up function), while it is not able to deal with joints which make a complex structure among the parts (network stack-up function). The worst-case approach may be applied to the matrix model, since the statistical one has not been developed yet.

All the details of the model are described in depth in Marziale and Polini (2009b).

2.4.3 The Jacobian Model

In the terminology adopted by the Jacobian model approach, any relevant surface involved in the tolerance stack-up is referred to as a *functional element*. In the tolerance chain, functional elements are considered in pairs: the two paired surfaces may belong to the same part (internal pair), or to two different parts, and are paired since they interact as mating elements (kinematic pair, also referred to as an external pair). The parts should be in contact to be modeled by this model.

Transformation matrices can be used to locate a functional element of a pair with respect to the other: such matrices can be used to model the nominal displacement between the two functional elements, but also additional small displacements due to the variabilities modeled by the tolerances. The form tolerances are neglected. The main peculiar aspect of the *Jacobian* approach is how such matrices are formulated, *i.e.*, by means of an approach derived from the description of kinematic chains in robotics. The transformation that links two functional elements belonging to a pair, and that includes both nominal displacement and small deviations due to tolerances, can be modeled by a set of six virtual joints, each associated with a DRF. Each virtual joint is oriented so that a functional element may have either a translation or a rotation along its *z*-axis. The aggregation of the six virtual joints gives rise to the transformation matrix linking one functional element to the other functional element of the pair (Laperrière and Lafond 1999; Laperrière and Kabore 2001). The position of a point lying on the second functional element of a pair, which may be assumed as depicting the functional requirement under scrutiny, with respect to the DRF of the first functional element (assumed as the global DRF) may be expressed by considering the three small translations and the three small rotations of the point in the global DRF through the product of a Jacobian matrix associated with the functional element with tolerances of all the functional element pairs involved (internal or kinematic) and a vector of small deviations associated with the functional element with tolerances of all the functional element pairs involved, expressed in the local DRF. The main element of the expression is the Jacobian matrix, which is relatively easy to compute, starting from the nominal position of the geometric elements involved. The tricky part, however, is to turn the assembly tolerances into displacements to assign them to the virtual joints defined for each functional element pair in the chain.

The main steps of the approach are described below:

1. *Identify the functional element pairs*. The first step is the identification of the functional element pairs (*i.e.*, pairs of relevant surfaces). The functional elements are arranged in consecutive pairs to form a stack-up function aimed at computing each functional requirement.
2. *Define the DRF for each functional element and the virtual joints*. The next step is to define a DRF for each functional element, and to create the chain of virtual joints representing the transformation that links the pair of functional elements. Once such information is available, the transformation matrix for each functional element can be obtained.
3. *Create the chain and obtain the overall Jacobian matrix*. The transformation matrices can be chained to obtain the stack-up function needed to evaluate each functional requirement. This model has been developed for assemblies involving joints which make a linear structure among the parts (linear stack-up function), while it is not able to deal with joints which make a complex structure among the parts (network of stack-up functions), such as the matrix model does.
4. Once the required stack-up function has been obtained, it may be solved by the usual methods in the literature (Salomons *et al.* 1996) for the worst-case or statistical approaches.
5. Finally, it is necessary to observe that this model is based on the TTRS criterion (Clément *et al.* 1998) and on the positional tolerancing criterion (Legoff *et al.* 1999). Therefore, the tolerances of a generic drawing need to be converted in accordance with the previously defined criteria, before carrying out the tolerance analysis.

2.4.4 The Torsor Model

The torsor model uses screw parameters to model 3D tolerance zones (Chase *et al.* 1996). Screw parameters are a common approach adopted in kinematics to describe motion, and since a tolerance zone can be seen as the region where a surface is allowed to move, screw parameters can be used to describe it. Each real surface of a part is modeled by a substitution surface. A substitution surface is a nominal surface characterized by a set of screw parameters that model the deviations from the nominal geometry due to the applied tolerances. Seven types of tolerance zones are defined. Each one is identified by a subset of nonzero screw parameters, while the remaining ones are set to zero as they leave the surface invariant. The screw parameters are arranged in a particular mathematical operator called a *torsor*, hence the name of the approach. Considering a generic surface, if u_A, v_A, and w_A are the translation components of its point A, and α, β, and $\gamma$are the rotation angles (considered small) with respect to the nominal geometry, the corresponding torsor is

$$\mathbf{T}_A = \left\{ \begin{matrix} \alpha & u_A \\ \beta & v_A \\ \gamma & w_A \end{matrix} \right\}_{\mathbf{R}} , \tag{2.43}$$

where **R** is the DRF that is used to evaluate the screw components.

To model the interactions between the parts of an assembly, three types of torsors (or small displacement torsor, SDT) are defined (Ballot and Bourdet 1997): a *part SDT* for each part of the assembly to model the displacement of the part; a *deviation SDT* for each surface of each part to model the geometric deviations from the nominal geometry; a *gap SDT* between two surfaces linking two parts to model the mating relation. The form tolerances are neglected and they are not included in the deviation SDT.

A union of SDTs is used to obtain the global behavior of the assembly. The aggregation can be done by considering that the worst-case approach computes the cumulative effect of a linear stack-up function of *n* elements by adding the single components of the torsors. This is not true for a network of stack-up functions, which has not been developed by the torsor model yet. The torsor method does not allow one to apply a statistical approach, since the torsor's components are intervals of the small displacements; they are not parameters to which it is possible to assign easily a probability density function.

The torsor model operates under the assumption that the TTRS and the positional tolerancing criteria are adopted, which means that the tolerances in the drawing may need to be updated before carrying out the tolerance analysis. The solution of stack-up functions arranged in a network has not been completely developed. Finally, it is worth pointing out that, in the relevant literature, the use of SDTs for modeling tolerance analysis problems tends to follow two main approaches: on one hand, SDTs are used to develop functions for computing the position of geometric elements (belonging to the assembly) as they are subjected to displacement allowed by tolerances (*e.g.*, see Chase *et al.* 1996); on the other hand, SDTs are used to model entire spatial volumes that encapsulate all the possible points in space that may be occupied by geometric elements during their variations (*e.g.*, see Laperrière *et al.* 2002). In the analysis of the case study, only the second approach has been considered, since it looks more promising.

The basic steps of the torsor model are described in the following (Villeneuve *et al.* 2001; Teissandier *et al.* 1999):

1. *Identify the relevant surfaces of each part and the relations among them.* The first step is to identify the relevant surfaces belonging to each part and the relationships among them; this information is usually collected in a *surfaces graph*. In this step the chains to relate the functional requirements to the relevant surfaces are identified.
2. *Derive the SDTs.* A deviation SDT needs to be associated with each relevant surface of each part. This leads to the evaluation of a global SDT for each part. Finally, the shape of the gap SDT is associated with each joint according to the functional conditions of the assembly.
3. *Obtain the functional requirement stack-up functions.* Compute the cumulative effects of the displacements and obtain the final linear stack-up function of each functional requirement.

2.5 Comparison of the Models

A first comparison of the previously described five models can be done by devising a set of indicators describing features, capabilities, and issues related to the application of such models to given tolerance stack-up problems. The indicators and their results for the five models are summarized in Table 2.3. The indicators were designed by drawing inspiration from what is available from the literature (Salomons *et al.* 1996), with the necessary adaptations. Each descriptor may assume one of the following three states: “X” if the model has a property, or it is capable of handling a specific aspect or issue of the problem, “–” if the same property is missing, or the model is not able to handle the aspect of the problem; “?” if the answer is uncertain, because it may depend on the specific tolerance analysis problem, or because there is not enough information to verify the capability. The first descriptor is the “analysis type”, which refers to the type of approach that can be adopted to solve the stack-up functions, *i.e.*, worst-case or statistical. The descriptor “tolerance type” indicates the kind of tolerance that the model may take into account: dimensional, form, or other geometric (no form) tolerances. The “envelope and independence” descriptors refers to the possibility of the model representing a dimensional tolerance when the envelope principle or the independence principle is specified. The “parameters from tolerances” descriptor indicates whether the model allows for translation of the applied tolerance ranges into the model parameter ranges. The “tolerance stack-up type” descriptor refers to the possibility of a model building and solving linear stack-up functions or networks of stack-up functions. The “joint type” descriptor refers to the joint types that the model may take into account, either with contact between the surfaces or with clearance. The “functional requirement schematization” descriptor refers to how a functional requirement can be represented by a feature or by a set of points belonging to a feature. The “tolerance zones interaction” descriptor indicates the capability of representing the interaction among more than one tolerance applied to the same surface. The "datum precedence" descriptor indicates whether a model can represent a sequence of datums. Finally, the "material modifiers condition" descriptor indicates the capability of a model to take into account material modifiers.

According to the results of the first comparison, the *vector loop* model and the *variational* model appear more developed than the others; they are the only ones that provide support for solving tolerance stack-up functions involving networks. Moreover, they provide a method for assigning probability density functions to model parameters, given the applied tolerances. However, the *vector loop* model and the *variational* model are not completely consistent with the actual ISO and ASME standards and they do not provide support handling interactions among tolerance zones.

The *vector loop* model is the only model providing actual support for modeling form tolerances; all the other models adopt the simplification consisting in considering the real features as coincident with their substitute ones.

The *variational* model supports the inclusion of precedence constraints among datums, and also the presence of material modifiers conditions.

The *matrix* model and the *torsor* model support only the worst-case approach for solving the tolerance analysis problem. This is a limitation, but their formalization allows them to handle joints with clearance, and interaction among tolerance zones.

Table 2.3 Results of the comparison by descriptors

		Vector loop	Variational	Matrix	Jacobian	Torsor
Analysis type	Worst case	X	X	X	X	X
	Statistical	X	X	–	X	–
Tolerance type	Dimensional	X	X	X	X	X
	Form	X	–	–	–	–
	Other geometric	X	X	X	X	X
Envelope and independence		–	–	–	–	–
Parameters from tolerances		–	–	–	–	–
Stack-up type	Linear	X	X	X	X	X
	Network	X	X	–	–	–
Link type	With contact	X	X	X	X	X
	With clearance	–	?	X	?	X
Functional requirement schematization	With feature	X	X	–	X	X
	With points	X	X	X	X	X
Tolerance zones interaction		–	–	X	X	X
Datum precedence		?	X	?	?	X
Material modifiers condition		?	X	?	?	?

X possible, – not possible,? unclear

The *Jacobian* model has the advantage that the Jacobian matrix can be easily calculated from nominal conditions, while displacements of the functional requirements can be directly related to displacements of the virtual joints; however, it is difficult to derive such virtual joint displacements from the tolerances applied to the assembly components. On the other hand, the *torsor* model may allow for an easy evaluation of the ranges of the small displacements directly from the tolerances applied to the assembly components, but then it is very difficult to relate these ranges to the ranges of the functional requirements of the assembly.

These two considerations have suggested the idea of a *unified Jacobian–torsor model* to evaluate the displacements of the virtual joints from the tolerances applied to the assembly components through the torsors and, then, to relate the displacements of the functional requirements to the virtual joint displacements through the Jacobian matrix (Laperriére *et al.* 2002; Desrochers *et al.* 2003). Although this is theoretically possible, since the deviations are usually small and, therefore, the equations can be linearized, the actual feasibility of this approach is still the subject of research.

Finally, the models considered have some common limitations. The first deals with the envelope rule: the models do not allow one to apply the envelope rule and the independence rule to different tolerances of the same part. The second is that there do not exist any criteria to assign a probability density function to the model parameters joined to the applied tolerances and that considers the interaction among the tolerance zones. The last point deals with the assembly cycle: the models are not able to represent all the types of coupling with clearance between two parts.

The solution of the case study by the five models considered is described in detail in Marziale and Polini (2009b). Table 2.4 summarizes the results for the functional requirement Δg as obtained by the application of the five models and compared with the solution obtained with manual computation when only dimensional

Table 2.4 Results of the comparison among the models applied to the case study with dimensional tolerances

Model	Type of analysis	Results (mm)
Exact solution	Worst case	+0.84 –0.78
Vector loop	Worst case	±0.78
	Statistical case	±0.52
Variational	Worst case	±0.78
	Statistical case	±0.51
Matrix	Worst case	±0.70
	Statistical case	–
Jacobian	Worst case	±0.78
	Statistical case	±0.53
Torsor	Worst case	±0.78
	Statistical case	–

tolerances are considered. Table 2.5 shows the results when both dimensional and geometric tolerances are applied.

The results obtained by considering only the dimensional tolerances show that all the models give slightly underestimated results with the worst-case approach, when compared with the results obtained manually with the approach described earlier. The *matrix* model has the highest error (–14%), while all the other models provide the same result (–4%). This is probably due to the way the dimensional tolerances are schematized (*i.e.*, the first datum is nominal, the variability due to the dimensional tolerance is considered to be applied only on one of the two features delimiting the dimension). Moreover, the statistical approach gives similar results for all the models considered.

The results obtained by considering both dimensional and geometric tolerances show that all the models, except the *vector loop* model, give similar results with the worst-case approach. This is probably due to the fact that the *vector loop* model considers the effect of a set of tolerances applied to a surface as the sum of the effects due to each single tolerance applied to the same surface. The effects of the different tolerances are considered to be independent. Therefore, increasing the number of tolerances applied to the same surface increases the variability of the functional requirement. This means that the interaction among tolerances defined on the same surface are not properly handled.

All five models produce very similar results when the statistical approach is applied.

Moreover, the results in Tables 2.4 and 2.5 obtained from the *Jacobian* model and from the *torsor* model are basically identical. This is due to the fact that a simplification was adopted when modeling the problem, *i.e.*, the angles of the box were considered fixed at 90°. This assumption is due to the need to avoid the networks of stack-up functions that the two models are not able to deal with. It means that all the tolerances applied may involve only translations of the sides of the box.

Table 2.5 Results of the comparison among the models applied to the case study with dimensional and geometric tolerances

Model	Type of analysis	Results (mm)
Vector loop	Worst case	±1.03
	Statistical case	±0.54
Variational	Worst case	±0.78
	Statistical case	±0.50
Matrix	Worst case	±0.69
	Statistical case	–
Jacobian	Worst case	±0.78
	Statistical case	±0.53
Torsor	Worst case	±0.78
	Statistical case	–

2.6 Guidelines for the Development of a New Tolerance Analysis Model

None of the models proposed in the literature provide a complete and clear mechanism for handling all the requirements included in the tolerancing standards (Shen *et al.* 2004). This limitation is reflected also in the available commercial CAT software applications, which are based on the same models (Prisco and Giorleo 2002). As already discussed in detail in previous work (Marziale and Polini 2009b), the main limitations of the actual models are the following: they do not properly support the application of the envelope rule and of the independence rule to different dimensional tolerances on the same part as prescribed by the ISO and ASME standards; they do not handle form tolerances (except for the vector loop model); they do not provide mechanisms for assigning probability density functions to model parameters starting from tolerances and considering tolerance zone interactions; finally, they are not capable of representing all the possible types of part couplings that may include clearance.

Some guidelines are now presented, aimed at the development of a new model that addresses at least some of the limitations highlighted above.

A dimensional tolerance assigned to the distance between two features of a part or of an assembly (Figure 2.7) may be required with the application of the envelope rule or of the independence principle. In the second case, to correctly define the relationship between the features, it is necessary to add a geometric tolerance in order to limit the form and the orientation deviations. The envelope principle states that when a feature is produced at its maximum material condition (MMC), the feature must have a perfect form. The MMC of a feature is the size at which the most material is in the part. The MMC size establishes a perfect form boundary and no part of the feature must extend outside this boundary. As the size of a feature departs from the MMC, its form is permitted to vary. Any amount of variation is permitted, as long as the perfect form boundary is not violated. However, the size limits must not be violated either. Therefore, if the envelope principle is applied to a dimensional tolerance, an additional constraint has to be considered to build and to solve the stack-up functions of the assembly. Both the ASME Y14.5M (1994) and the ISO 8015 (1985) standards foresee the possibility that, also on the same part, dimensional tolerances may be assigned with or without the application of the envelope principle. Consequently the mathematical model that is used to schematize a dimensional tolerance in order to build and to solve the stack-up functions should necessarily take into consideration these two possibilities. To overcome this limitation of the models in the literature that are not able to consider these two cases (Desrochers *et al.* 2003), a possible solution is to consider a greater set of parameters to model the degrees of freedom of the planes delimiting the dimension considered due to the applied tolerances (dimensional, orientation, form). The envelope rule and the independence rule constrain those parameters in different ways.

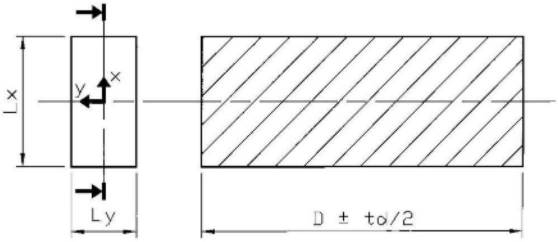


Figure 2.7 Two-dimensional dimensional tolerance

To model the form tolerance, it is possible to introduce a virtual transformation that is assigned to points of the surface to which a form tolerance is assigned. This approach was introduced by Chase *et al.* (1996) in their vector loop method for tolerance analysis.

In a statistical approach, a probability density function is assigned to each model parameter. Therefore, the tolerance analysis model has to determine the probability density function of each parameter according to the interaction of the tolerance zones. To do this, a possible solution is to decompose the possible deviation of a feature into different contributions – the dimensional, the form, the position, and the orientation ones – whose thickness is described by a model parameter. Each parameter may be considered independent of the others and it may be simulated by a probability density function which may be modeled by a Gaussian probability density function with a standard deviation equal to one sixth of the corresponding tolerance range. If more than one tolerance is applied to the same feature, the sum of the squares of the ranges of the applied tolerances is equal to the square of the range of the overall tolerance. For example, if the envelope principle is applied, the overall tolerance is the dimensional tolerance applied to the feature.

When two parts are assembled together, the mating surfaces form a joint. If the joint involves clearance, the clearance affects only some of the six small kinematic adjustments that define the position of a part with respect to the other. To evaluate the model parameters of the joint, it must be observed that they depend on the cumulative effects of the assembly constraints that must be satisfied by the coupled surfaces of the joint. The admissible values of the model parameters must be considered for each assembly constraint; therefore, all the constraints have to be considered and the resulting admissible values of the model parameters may be calculated as the intersection of the values previously defined. If the calculated domain of the admissible values of the model parameters is empty, the assembly is not possible. If the admissible domain contains a set of points, the convex hull, representing the boundary that encloses the points, may be determined. Once the boundary has been evaluated, it can be used in the tolerance analysis model as an additional constraint acting on the model parameters in the worst-case approach. It can be used to define the range of the probability density function that is assigned to the model parameters in the statistical approach (Marziale and Polini 2009a).

2.7 Conclusions

In this work, five different models available from the literature on tolerance analysis were compared through their application to a case study. None of the models proposed in the literature provide a complete and clear mechanism for handling all the requirements included in the tolerancing standards, and this limitation is reflected also in the available commercial CAT software. The main limitations include the following: no proper support for the application of the envelope rule and of the independence rule; cannot handle form tolerances (except for the vector loop model); no mechanisms for assigning probability density functions to model parameters starting from tolerances and considering tolerance zone interactions; no proper representation of all the possible types of part couplings that include clearance.

Guidelines were presented for the development of a new model aimed at addressing such highlighted limitations. Some suggestions were given to consider dimensional tolerance with the application of both the envelope principle and the independence principle, to take into account the real features and the interaction of the tolerance zones, to consider joints with clearance among the assembly components, and to adopt both the worst-case and the statistical approaches to solve the stack-up functions. The implementation of those suggestions in a new model and its application to case studies is the subject of ongoing research.

Acknowledgements This research was partially funded by the Italian Ministry of University and Research grant PRIN 2005 – “Characterization of the multiscale geometry of manufactured surfaces for process control and tolerance analysis”.

References

ASME Y14.5M (1994) Dimensioning and tolerancing. American Society of Mechanical Engineering, New York

Ballot E, Bourdet P (1995) Geometrical behaviour laws for computer aided tolerancing. In: Proceedings of the 4th CIRP seminar on computer aided tolerancing, University of Tokyo, April 1995

Ballot E, Bourdet P (1997) A computational method for the consequences of geometric errors in mechanisms. In: Proceedings of the 5th CIRP seminar on computer aided tolerancing, Toronto, 27–29 April 1997

Berman YO (2005) Shape and position uncertainty in mechanical assembly. PhD thesis, The Hebrew University, Jerusalem

Boyer M, Stewart NF (1991) Modeling spaces for toleranced objects. Int J Robot Res 10:470–582

Chase KW (1999) Multi-dimensional tolerance analysis (automated method). In: Drake PJR (ed) Dimensioning and tolerancing handbook. McGraw-Hill, New York

Chase KW, Gao J, Magleby SP (1995) General 2-D tolerance analysis of mechanical assemblies with small kinematic adjustments. J Des Manuf 5:263–274

Chase KW, Gao J, Magleby SP et al (1996) Including geometric feature variations in tolerance analysis of mechanical assemblies. IIE Trans 28:795–807

Chase KW, Gao J, Magleby SP (1997) Tolerance analysis of 2- and 3D mechanical assemblies with small kinematic adjustments. In: Zhang HC (ed) Advanced tolerancing techniques. Wiley, New York

Clément A, Rivière A, Temmerman M (1994) Cotation tridimensionelle des systèmes mécaniques, théorie & pratique. Cachan, France

Clément A, Riviére A, Serré P et al (1998) The TTRSs: 13 constraints for dimensioning and tolerancing. In: ElMaraghy HA (ed) Geometric design tolerancing: theories, standards and applications. Chapman & Hall, London

Delchambre A (1996) CAD method for industrial assembly. Concurrent design of product, equipment and control systems. Wiley, New York

Desrochers A, Rivière A (1997) A matrix approach to the representation of tolerance zones and clearances. Int J Adv Manuf Technol 13:630–636

Desrochers A, Ghie W, Laperrière L (2003) Application of a unified Jacobian-torsor model for tolerance analysis. J Comput Inf Sci Eng 3:2–14

Faerber PJ (1999) Tolerance analysis of assemblies using kinematically derived sensitivities. ADCATS report no. 99-3. http://adcats.et.byu.edu/reportsandpublications.php

Gao J, Chase KW, Magleby SP (1998) Generalized 3-D tolerance analysis of mechanical assemblies with small kinematic adjustments. IIE Trans 30:367–377

Gupta S, Turner JU (1993) Variational solid modelling for tolerance analysis. IEEE Comput Graph Appl 13:64–74

Hong YS, Chang TC (2002) A comprehensive review of tolerancing research. Int J Prod Res 40:2425–2459

ISO 8015 (1985) Fundamental tolerancing principle. International Organization for Standardization, Geneva

Laperrière L, Lafond P (1999) Modelling tolerances and dispersions of mechanical assemblies using virtual joints. In: Proceedings of ASME design engineering technical conferences, September 12–15, Las Vegas, Nevada, USA

Laperrière L, Kabore T (2001) Monte Carlo simulation of tolerance synthesis equations. Int J Prod Res 39:2395–2406

Laperriére L, Ghie W, Desrochers A (2002) Statistical and deterministic tolerance analysis and synthesis using a unified Jacobian-torsor model. CIRP Ann 51:417–420

Legoff O, Villeneuve F, Bourdet P (1999) Geometrical tolerancing in process planning: a tridimensional approach. Proc Inst Mech Eng Part B 213:635–640

Li B, Roy U (2001) Relative positioning of toleranced polyhedral parts in an assembly. IIE Trans 33:323–336

Marziale M, Polini W (2009a) Clearance joint modeling for tolerance analysis. In: Proceedings of the 11th CIRP international conference on CAT, Annecy, France, March 26–27

Marziale M, Polini W (2009b) A review of two models for tolerance analysis: vector loop and matrix. Int J Adv Manuf Technol 43:1106–1123

Nigam SD, Turner JU (1995) Review of statistical approaches to tolerance analysis. Comput Aided Des 27:6–15

Prisco U, Giorleo G (2002) Overview of current CAT systems. Integr Comput Aided Eng 9:373–397

Salomons OW, Haalboom FJ, Jonge Poerink HJ et al (1996) A computer aided tolerancing tool II: Tolerance analysis. Comput Ind 31:175–186

Shen Z, Ameta G, Shah JJ et al (2004) A comparative study of tolerance analysis methods. J Comput Inf Sci Eng 5(3):247–256

Teissandier D, Couétard Y, Gérard A (1999) A computer aided tolerancing model: proportioned assembly clearance volume. Comput Aided Des 31:805–817

Villeneuve F, Legoff O, Landon Y (2001) Tolerancing for manufacturing: a three-dimensional model. Int J Prod Res 39:1625–1648

Whitney DE (2004) Mechanical assemblies. Their design, manufacture and role in production development. Oxford University Press, New York

Whitney DE, Mantripragada R, Adams JD et al (1999) Toward a theory for design of kinematically constrained mechanical assemblies. Int J Robot Res 18:1235–1248

Part II
Impact on Product Quality Inspection

Chapter 3
Quality Inspection of Microtopographic Surface Features with Profilometers and Microscopes

Nicola Senin and Gianni Campatelli

Abstract With the increasingly widespread adoption of micromanufacturing solutions and with the production of a growing number of artifacts defined at the microscopic and submicroscopic scales, increasingly smaller geometries need to be verified for quality assurance. The study of precision at micro and submicro scales is gaining considerable interest: relevant issues pertain to how to define allowable geometric error on parts of such small sizes (*e.g.*, semiconductor products, microelectromechanical systems, other microcomponents) with proper dimensional and geometric tolerances, and how to measure them. This work addresses the specific problem of assessing geometric error associated with *micromanufactured surface features*. Three-dimensional digital microscopes and profilometers for microtopography analysis are increasingly being adopted for such a task, owing to their suitability to operate at very small scales. However, this raises several challenges, as three-dimensional microscopes and profilometers have traditionally been used in different application domains, and are mainly aimed at the inspection of surface finish; new modes of operation must be identified which take into consideration such peculiarities. Both families of instruments need to be closely investigated, and their main constraints and benefits dissected and analyzed to assess their adaptability to the new task of assessing geometric error on micromanufactured parts or surface features.

N. Senin
Dipartimento di Ingegneria Industriale, Università degli Studi di Perugia,
Via G. Duranti 67, 06125 Perugia, Italy,
e-mail: nsenin@unipg.it

G. Campatelli
Dipartimento di Meccanica e Tecnologie Industriali, Università degli Studi di Firenze,
Via S. Marta 3, 50139 Florence, Italy,
e-mail: gianni.campatelli@unifi.it

3.1 Introduction

The problem of defining and measuring geometric accuracy on micromanufactured parts and/or surface features is one of fundamental importance. The small sizes and the micromanufacturing processes currently adopted to fabricate these manufactured parts make it difficult to exert strict control on their geometries during fabrication. The conventional approaches adopted for standard-sized mechanical parts are difficult to scale down: for example, a dimensional tolerance interval of about 1% of the nominal dimension can be commonly found in the design of a standard-sized mechanical part; however, it can not be easily applied to a microelectromechanical system (MEMS), where it would result in a tolerance interval of a few nanometers, which is too demanding for the microfabrication processes currently adopted for realizing MEMS. Similar considerations apply to the definition of proper geometric tolerances for these devices. To make the problem even harder, an increasingly wider variety of geometries, materials, and architectures are being adopted for the production of semiconductors, MEMS, and several other types of microsystems: such a variety has led to the adoption of a multitude of different approaches for defining geometric precision requirements, and for verifying conformance, where each solution is often tailored to the specific product and/or to the operational habits and protocols of each specific manufacturing company.

Currently, no widespread consensus exists on an overall approach for defining maximum allowable geometric errors, and for measuring them on microfabricated parts, and consequently no general theory or industrially applicable guidelines are available (Hansen *et al.* 2006; Hansen 2007; Kurfess and Hodgson 2007; Nichols *et al.* 2008). However, the study of the metrology applied to microcomponents (sometimes referred to as *micrometrology*) is gaining relevance and awareness is gradually spreading that a simple adaptation of traditional "macro" approaches to the micro and submicro scales may not be as successful as hoped. Even further down the line is the development of a unified approach to metrology: a significant amount of research work is still required, both on the theoretical and on the industrial aspects of the problem, to bridge a gap that is widened also by the adoption of different measurement instrumentation depending on the scale of intervention.

In this work, an effort is made to collect useful information on a few specific classes of measurement instruments which are increasingly becoming routinely involved in geometric error assessment on products such as semiconductors, MEMS, and microcomponents in general, and also on standard-sized parts featuring microtopographic surface patterns. The attempt is to establish a basic knowledge foundation, which may be useful in understanding the main metrological issues related to the measurement of microtopographic surface features.

In this chapter, the focus is on profilometers for microtopography and 3D microscopes; in current industrial practice these instruments are often found applied to the inspection of microfabricated parts and microtopographic surface patterns in general. This is primarily due to the native capability of such instruments to operate at micrometric and submicrometric resolutions. The most notable dimensional metrology alternative, the *micro coordinate measuring machine* (CMM*)* (Weckenmann *et al.* 2004) is still largely confined to laboratory and research usage, although commercial implementations by several manufacturers are available (Carl Zeiss 2009; Mitutoyo 2009; Werth Messtechnik 2009). The application of the micro CMM to the inspection of microfabricated parts and surface features will not be dealt with in this chapter.

Profilometers for microtopography and 3D microscopes are not primarily designed for being applied to the metrological inspection of semiconductors, MEMS, and microsystems in general. Profilometers originate from the need for quantitative measures for *surface finish*, where quantitative information is mainly required in the form of synthetic indicators (*i.e.*, roughness, waviness, and form error parameters). Digital microscopes originally came from the domain of qualitative investigation of surface topography by means of visual inspection. Even though the class of digital microscopes analyzed in this chapter, 3D microscopes, is capable of producing quantitative height maps of surface topography, their application to quantitative metrology has not been fully analyzed and specified yet. Dimensional and geometric tolerance assessment on MEMS and other microfabricated systems requires these instruments to be carefully analyzed in terms of their intrinsic properties, mode of operation, capability, and limitations.

This chapter has two objectives. In the first part, the main classes of profilometers for microtopography analysis and 3D microscopes that are currently available are introduced and described. Architectures and implementation choices are justified from the viewpoint of their canonical applications. In the second part, the intrinsic advantages and limitations of these instruments are discussed from the perspective of applying them to micrometrology-focused domains, where the objective is to evaluate geometric errors on microfabricated parts, where "*geometric error*" refers to the overall deviation of a measured geometric entity with respect to its nominal counterpart. The main focus will be on how – and with what limitations – such instruments can be turned from their conventional tasks of surface finish assessment (profilometers) or support of visual inspection (microscopes) to the novel tasks of quantitative assessment of geometric error on microfabricated parts and surface features, and to how measurements obtained with such instruments can be used to verify conformance to geometric and/or dimensional tolerances.

A disclaimer should be made concerning quantitative performance values that will be provided throughout the text: the technology of surface microtopography measurement is rapidly evolving; manufacturers continuously renovate their offerings and provide a large array of solutions that cover a wide range of performances; even for instruments based on the same measurement principles, it is quite common to run into many different performance options, owing to the variety of possible combinations or interchangeable parts and accessories, or simply owing to different technological implementations of the same measurement principles. Whenever performance-related quantitative values are reported in this chapter, they should be interpreted as typical values for each specific instrument type;

these values were obtained by combining the information publicly available at the Web sites of several manufacturers, and should be representative of their current offerings, at least at the time of writing. While specific numbers may change, a comparison in terms of orders of magnitude should suffice in delineating the main performance differences among instrument families.

3.2 Profilometers and 3D Microscopes for Microtopography Analysis

In this section, the most common classes of profilometers and 3D microscopes are introduced and discussed. The focus will be on those instruments that are currently used for microtopography analysis and visual inspection, but that could be adapted to operate for the quantitative assessment of geometric error on microfabricated parts or surface features. The proposed list is not meant to be complete: the selection criterion was to highlight those instruments which are more successful and widespread, and therefore more commonly found in industrial practice.

Before proceeding with the illustration of the main instruments, it is convenient to provide a brief introduction concerning the main terms that are commonly used to refer to such instrument classes. In general, the term "*profile*" refers to some digital data aimed at depicting the outline/contour formed by the reliefs of a surface; in this sense, a *profilometer*, or *profiler*, is an instrument devoted to acquiring profiles. Profilometers acquire surface topography information by having a probe tracing the surface along a specified traversal trajectory. As the probe moves, surface height data are collected. The profilometers analyzed in this chapter are all *3D profilometers*, *i.e.*, they are devoted to acquiring *3D surface topography* information; this is generally done through the sequential acquisition of a series of profiles, conveniently placed on the original surface, which are then combined – via software – into a single coherent representation of surface topography.

The term "*microscope*" is commonly used to indicate an instrument that produces an *image*, where a small (hence the term "micro") surface region under scrutiny is magnified to aid visual observation (hence the term "scope"). When used in quality inspection of mechanical parts, traditionally microscopes have had the main role of supporting visual inspection for investigating surface finish and material microstructure (also in cross sections).

In this work, the term "*3D microscope*" is used to refer to a variation of the digital microscope, designed to acquire topography information as a height map, *i.e.*, the equivalent of a digital image, but whose pixels contain surface height data. Some 3D microscopes operate very similarly to 3D profilometers, and scan the surface as if they were acquiring a series of profiles; others acquire surface data points simultaneously, thus resembling more a conventional optical microscope, where the light needed to form the image reaches the observer simultaneously.

Some instruments that go under the name of microscopes, namely, optical instruments, may provide additional information in the form of color associated with the points of the height map, thus providing something closer to the results of visual observation.

In some industrial environments, the term “3D microscope” is also commonly used to describe instruments which are capable of providing *volume data*, *i.e.*, digital information concerning the entire 3D structure of the specimen (including the internal parts). An example of an instrument of this type is the confocal laser scanning microscope, which is included in the list of selected instruments and will be illustrated later. Instruments capable of providing volume data are generally also capable of providing surface topography data, as a subset of the measured data: this applies to the confocal laser scanning microscope as well, which is why it was included in the selection.

3.2.1 Stylus-based Profilometers

Stylus-based profilometers are based on a *stylus probe*, *i.e.*, a probe based on a stylus that slides in contact with the measured surface. The first example of a working instrument resembling the current design dates back to 1933 (Abbott and Firestone 1933a, b). As the stylus traverses the surface following its reliefs (see Figure 3.1), its displacements with respect to a reference (*z* direction) are captured by a transducer within the probe main body (linear variable displacement transducer or optical distance sensor), and are recorded along with the *x* position of the tip, which travels along a straight path. Typical stylus tips are 60° or 90° conical and terminate with 2-, 5- or 10-μm spherical radius, as defined in ISO 3274 (1998). Currently, a wide range of spherical radiuses are available, down to fractions of 1 μm.

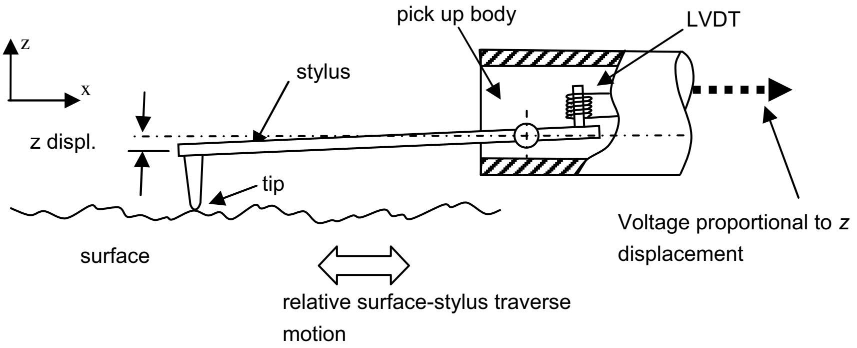

pick up body
LVDT
z
x
stylus
z displ.
tip
surface
relative surface-stylus traverse motion
Voltage proportional to z displacement

Figure 3.1 Architecture of a stylus probe. *LVDT* linear variable displacement transducer

Profile points are acquired sequentially, during traversal. Measurement can take place continuously, the acquisition being triggered at constant time intervals while the stylus travels at constant velocity (which gives rise to uniformly spaced points in the traversal direction, at least nominally) or it can take place in a step-by-step fashion, with the probe coming to a complete halt before z measurement takes place (allowing for points to be located anywhere and at any reciprocal distance along the probe traveling path). The actual implementation and the traveling strategy depend on the instrument type and the intended usage.

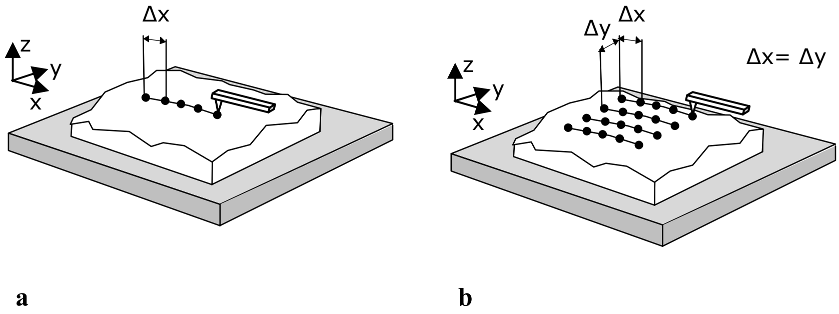


Figure 3.2 Profile scanning solutions for stylus-probe systems: **a** single-profile scanning with uniform point spacing, and **b** parallel-profile scanning with uniform point spacing (raster scanning)

With a single scan along a straight path a *2D profile* can be acquired (see Figure 3.2a). The term "2D profile" derives from the fact that such a profile can be completely defined in a 2D space. Traditional profilometers usually acquire a 2D profile by scanning the surface along a straight path. In this work they are referred to as *2D profilometers* to distinguish them from more complex solutions illustrated later. The 2D profilometer equipped with a stylus probe is the single most established and widespread solution for surface finish analysis; its main architectural elements are defined by international standards, see, in particular, ISO 3274 (1998). Measured data consist of a series of x, z coordinate pairs, or of a simpler series of z values with common x spacing (Δx) value, in the case of uniformly spaced acquisition along a straight path. The stylus-based 2D profilometer is used to assess surface finish properties related to form error, waviness, and roughness in the form of *quantitative parameters*, also specified in detail by international standards, in particular, ISO 4287 (1997), ISO 4288 (1996), ISO 12085 (1996), ISO 8785 (1998), ISO 13565-1 (1996), ISO 13565-2 (1996), and ISO 13565-3 (1998).

While acquiring surface topography information through one or multiple scans along straight paths may be sufficient for several quality inspection applications, it was highlighted that the reconstruction of a complete 3D representation of surface topography is often capable of providing superior results (Lonardo *et al.* 1996). Assuming 3D topography data are available, a significant research effort has been long under way to define form error, waviness, and roughness *areal parameters* to replace the 2D parameters in capturing the topographical properties of a 3D surface topography. For this purpose, an international standard has been published (ISO 25178-6:2010) and several others are currently being prepared: ISO/CD 25178-1, ISO/DIS 25178-2, ISO/DIS 25178-3, and ISO/DIS 25178-7.

Three-dimensional surface topography can be reconstructed from a series of 2D profiles (Figure 3.2b) with known relative displacement.

The most common measurement approach implemented by stylus-probe systems consists in doing parallel passes, which is compatible with the preferred mode of operation of the stylus probe. The simplest way to recombine the profiles into a single surface representation consists in having profiles equally spaced in the y direction (Δy spacing) and using the same spacing between points in each profile (Δx spacing), the overall result being a grid of equally spaced z coordinates; this is usually referred to as *raster scanning*. Profile data obtained by means of raster scanning are often referred to as a height map, or *image*, as it resembles a uniform grid of grayscale pixels, each z coordinate being equivalent to a gray level.

Stylus-based *3D profilometers* can be synthetically depicted as 2D profilometers with an additional y-axis drive. They are usually implemented as depicted in Figure 3.3 (the illustration does not include the control unit, usually a PC-based system).

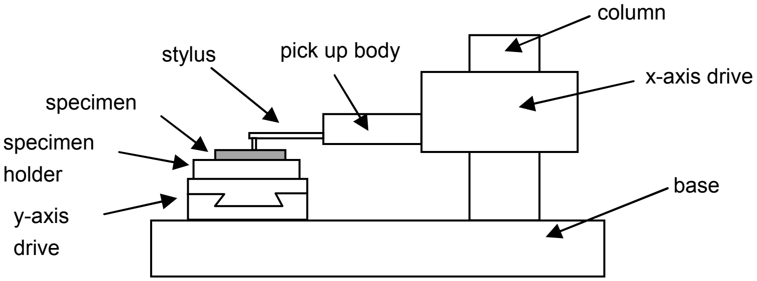


Figure 3.3 Architecture of a Cartesian, stylus-based 3D profilometer (does not include the controller unit)

The architecture was originally developed for 2D profilometers, and hence centered about the x-axis drive, whose predominant role is enforced by the very nature of operation of the stylus; 3D profilometers have an additional y-axis drive. A variant of this 3D architecture has the x-axis mounted under the specimen holder, together with the y-axis drive. Both architectures are very delicate in their implementations, as different effects related to the kinematics and dynamics of the axes have to be taken into account when designing and manufacturing the instruments, with significant influence on measurement error.

A problem shared by both 2D and 3D profilometers is the establishment of a *reliable datum* for the z measurement. Typical solutions for a 2D profilometer may include the use of an external reference (*e.g.*, optical flat) to retrieve the *z* error introduced by the probe while it translates along the scanning direction, and/or the use of an analogue high-pass filtering device, with behavior defined by international standards such as ISO 3274 (1998) and ISO 11562 (1996). to remove this type of error. These solutions are generally not viable for 3D profilometer architectures, because it is important to maintain the same datum for all measured profiles. The preferred strategy for 3D profilometers consists in not adopting any type of mechanical/analogue filter devoted to such a task (*i.e.*, acquisition of true profiles) and to proceed at the identification of a measurement datum at a later stage, after measurement, operating directly via software on the acquired digital data.

A particular type of mechanical filtering solution is worth analyzing in more detail: the *stylus with a skid*. Typical architectures of styluses with a skid are shown in Figure 3.4: the probe consists of a main body terminating with a skid that is in contact with the surface; the stylus is mounted within the main body. Both the main body and the stylus freely follow the surface reliefs during scanning since the probe body is connected to the instrument with a joint that allows it to follow surface reliefs; only stylus displacement (relative to the main body) is recorded, since the vertical position of the main body acts as the reference datum.

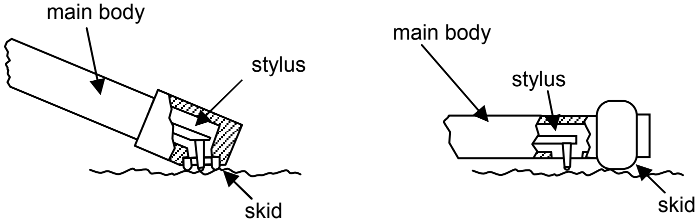


Figure 3.4 Two typical architectures for skidded styluses. Stylus displacement is recorded relative to the main body axis, whose orientation is determined by the skid sliding over the surface

The result is a mechanical high-pass filtering effect; therefore, this type of stylus is incapable of recording a complete profile shape, but can just record its highest-frequency components. For this reason, profilometers equipped with such a stylus are usually referred to as *roughness measurement instruments*, as the only information they are capable of acquiring pertains to the roughness (high-frequency components) of a surface. A skidded stylus has a more limited measurement range than a stylus without a skid, but has higher vertical resolution and accuracy; thus, it is preferentially adopted to measure surfaces of high nominal finish quality. The use of such a solution is quite common in 2D profilometers, because a surface with high surface finish requires a high-resolution probe, whose limited range may cause out-of-range measurements if the specimen surface is not perfectly leveled; the stylus with a skid solves this problem.

Nevertheless, the use of skidded styluses in 3D solutions should be avoided when possible, or at least it should be considered with care, as different geometric effects of the skid in the *x* and *y* directions may give rise to different filtering results, making it hard to identify a common datum for all the traces.

The last notable instrument belonging to the stylus-based class is the *portable roughness measurement system.* It is usually a simplified version of a Cartesian 2D profilometer where the *x*-axis drive, pickup body, and stylus are enclosed, together with a controller unit and minimal display, in a compact, portable unit. The probe is equipped with a stylus with a skid, so the reference datum for the single trace is provided directly, independently of the placement of the instrument. The system is usually configured to compute roughness and waviness parameters only.

3.2.2 Performance and Issues of Measuring with Stylus-based Profilometers

The measurement behavior of a stylus-based profilometer is largely ascribed to the physical interaction of the stylus tip with the surface, as shown in Figure 3.5.

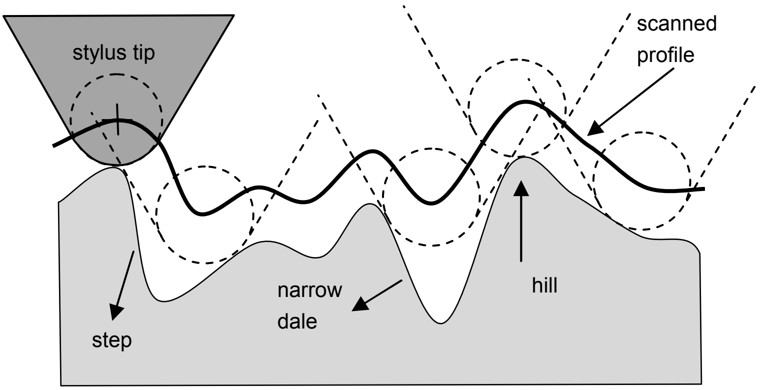


Figure 3.5 Interaction of the stylus tip with the measurand surface during scanning

As depicted in the figure, the interaction of the tip with the surface results in a mechanical filtering effect. Consequently, steep slopes, narrow cavities, and spikes may be smoothened, or remain undetected. The overall performance and issues related to sylus-based profile measurement, discussed in detail in the following paragraphs, are significantly influenced by the interaction of the stylus tip geometry (conicity and tip radius) and the profile *aspect ratio* (ratio of the length and height of the profile). The steeper the slopes (higher aspect ratios), the more difficult it is to acquire them correctly, for a given tip geometry.

3.2.2.1 Measurement Performance

When describing the measurement performance of a profilometer, the most commonly cited quantities are probably *resolution* and *range*. In simple terms, *resolution* is the smallest value difference that can be detected by the instrument, while *range* represents the interval of values that can be measured with acceptable error. For all profilometer instruments, a clear distinction is drawn between *vertical* resolution and range (*i.e.*, about the *z*-axis) and *lateral* resolution and range (*i.e.*, about the *x*- and *y*-axes). The distinction between vertical and lateral values reflects the traditional, overall conceptual approach that sees surface finish analysis as mainly targeted at studying *height deviations* from an ideal reference. Range and resolution vary greatly among stylus-based profilometers, as highlighted by a survey of several manufacturers (Ambios Technology 2009; Carl Zeiss 2009; KLA-Tencor 2009; Mahr Federal 2009; Taylor Hobson 2009; Veeco Instruments 2009; Werth Messtechnik 2009). *Vertical resolutions*, in particular, may be found for commercial instruments ranging from 0.1 nm, for very accurate instruments dedicated to high-precision measurement, to 0.5 μm, for instruments dedicated to coarser evaluations. Vertical range is related to resolution in the sense that usually the higher the resolution, the shorter the range: very accurate instruments may have vertical ranges as limited as 1 mm and thus are only suitable for very flat surfaces; to measure higher steps (or higher aspect ratios, in general), coarser-resolution instruments (larger ranges) must be adopted instead. With the proper probe, some instruments may get up to approximately 50 mm. Lateral resolution (*x, y* spacing of adjacent points) and range (overall *x, y* probe traverse length) are important as they determine the spatial wavelengths that are either filtered out or kept by the measurement process. In surface finish analysis, this determines what components of roughness, waviness, and form error are captured. Vertical resolution and range depend on the stylus probe (geometry, encoder, *etc.*); lateral resolution mainly depends on the actuators adopted for the *x, y* table. Stylus tip geometry may affect the maximum achievable lateral resolution (*i.e.*, minimum spacing between point measurements), since for small spacing envelope effects due to probe–surface interaction may become relevant (as hinted at in Figure 3.5). For a typical surface finish analysis application, lateral resolution is not as relevant as vertical resolution, and typical instruments have lateral resolutions ranging from 0.1 μm to several millimeters. Lateral ranges are typically in the 0.2–200 mm interval. The maximum lateral range is constrained by the vertical range: in fact, if the probe is calibrated to operate on a specific surface point, it may generate out-of-range measurements as soon as it moves too far away from such a point, owing to errors in specimen leveling that cannot be completely avoided. The vertical range compensates for specimen leveling errors, and thus greatly influences the extent of the actual region that can be acquired, regardless of the *x, y* table and actuator capabilities.

Resolution and range alone do not convey sufficient details about the measurement performance; some type of indication about the *measurement error* (difference between the true value of the measurand and the measured value) should be provided as well. It is now generally accepted that information about the measurement error should be provided in some rigorous statistical form, *e.g.*, as *measurement uncertainty* (*i.e.*, a characterization of the range of values within which the true value is asserted to lie with some level of confidence). Measurement uncertainty is considered a valid quantitative indicator of *measurement accuracy*, *i.e.*, the combination of *measurement trueness* (closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value – mainly related to systematic error), and *measurement precision* (closeness of agreement between measured quantity values obtained by replicate measurements – a measure of dispersion, mainly related to random error). Further details about the terminology and procedures for computing measurement uncertainty can be found in ISO/IEC Guide 99:2007(E/F) (2007), ISO/IEC Guide 98-3 (2008), ISO 5725-1,2,3,4 and 6 (1994), and ISO 5725-5 (1998).

The problem of obtaining reliable quantitative information concerning measurement uncertainty for a given instrument is complex, and is undermined by different interpretations concerning terminology and statistical procedures that have been adopted by the manufacturers and by the industrial community in general over the years. For example, the term "*accuracy*", now generally assumed as encompassing both *trueness* and *precision*, has been previously associated with the meaning of *trueness* alone; the term "*precision*" itself has sometimes been mistaken for accuracy, at least in colloquial terms, and used consequently. As a consequence of this, the procedures for calibrating the instruments, and thus ensuring *metrological traceability*, have sometimes suffered as well. Currently, the more traditional and well-established stylus-based 2D profilometers are the solution with the best metrological traceability among all the instrument classes that are discussed in this chapter. This is due to their widespread, long-term adoption, and to the existence of measurement standards and procedures dedicated to such instruments and their calibration; see, for example, ISO 5436-1 (2000), ISO 5436-2 (2001), and ISO 12179 (2000). Stylus-based 3D profilometers are a somewhat newer breed and two ISO standards have recently become available for them: ISO 25178-601:2010 and ISO 25178-701:2010.

3.2.2.2 Constraints on Material, Geometry, and Surface Topography of the Part To Be Measured

Except for portable roughness measurement instruments, which can almost always be brought directly onto the surface to be measured (unless specific impediments exist on the part), all other stylus-based profilometers are generally implemented as fixed installations, and require the specimen to be placed on its proper fixture within the instrument. Compared with noncontact solutions, stylus-based systems are generally more robust to environmental conditions since measurement originates from physical contact between the probe and the surface. Nevertheless, like in any other type of measurement solution, the higher the accuracy of the instrument, the more the instrument is sensitive to environmental factors such as temperature, vibrations, atmospheric conditions, and electromagnetic fields, stylus-based instruments being no exception. Typical 2D and 3D profilometers, including also portable roughness testers, usually operate without the need for enclosures: this means that often even large parts can be measured as long as a suitable orientation is found that allows the part to be placed without interfering with the instrument, while at the same time making the (usually small) region to be measured properly aligned with the stylus probe.

Very high resolution instruments may require a specifically sized specimen to be prepared. In this case the specimen may be physically extracted from the part (destructive testing), or may be manufactured to mimic the surface on the actual part. In general, as measurement resolutions increase, specifically prepared specimens need to be smaller and more accurately manufactured. Particular attention is usually dedicated to ensure that the specimen is manufactured so that the surface to be measured can be accurately aligned to the *x, y* plane defined by the instrument. Leveling errors can be compensated via software by acting on the coordinates of measured points, but only as long as out-of-range conditions are not encountered by the stylus probe.

As stated earlier, since the stylus is in physical contact with the surface during scanning, stylus-based profilometers are classified as *contact-based measurement instruments*. Contact techniques raise several concerns: specimen and/or stylus damage due to grooving, chipping, or wear, stick–slip interactions, stylus stuck between rough surface features, and bouncing behavior during traversal are all aspects that may introduce significant measurement error when not also damaging the inspected part, or the instrument as well. Key aspects to consider when evaluating the suitability of a material for being acquired by means of a stylus-based profilometer are the stylus contact force, typically in the range 9.81×10^{-8}–9.81×10^{-5} N for stylus-based instruments, traversal speed, friction, and the presence of topography features that may negatively affect the scanning process (*e.g.*, high-aspect-ratio features). The suitability ultimately depends also on the instrument and stylus type, the required performance, and the application domain. When the prerequisites are satisfied, stylus-based profilometers are the most robust solution currently available.

3.2.3 Optical Profilometers and Optical 3D Microscopes

Optical profilometers and *optical 3D microscopes* are surface topography measurement instruments adopting noncontact measurement techniques based on the use of an *optical probe*. Optical probes operate using reflected *visible light* to acquire topography information; their main differentiating aspect with respect to stylus-based profilometers is that they do not need to touch the surface for the measurement to take place.

Two main types of architectures are typically found for instruments equipped with optical probes. The single-point architecture is shown in Figure 3.6a: the probe acts as a distance measurement device for the point (actually a small region) targeted by the light spot. The distance to the probe can be turned into a point height with respect to a reference datum. Motorized *x, y* axes can be used to acquire multiple points at different *x, y* coordinates. Single and multiple profile scanning strategies can be implemented, the most common being raster scanning; as opposed to stylus-based profilometers, the main scanning direction is not constrained to be on the *x*-axis.

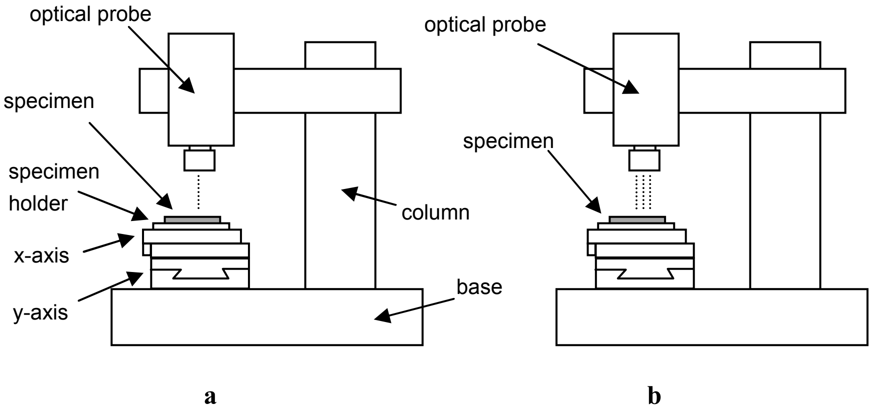


Figure 3.6 Example architectures for instruments equipped with optical probes: **a** single-point measurement system (raster scanning achieved by a motorized *x, y* table), and **b** wide-field measurement system (simultaneous acquisition of a surface region; a motorized *x, y* table can be used to acquire multiple regions sequentially, and collate them later)

The second type of architecture is shown in Figure 3.6b, and it is known as *wide-field architecture*. The term comes from the traditional optical microscope (the wide-field microscope), where image formation takes place by the optics, without scanning. In this case, it means that all the points of the height map are acquired simultaneously, thus mimicking the mode of operation of an eye observing through the ocular of an analog microscope.

As stated earlier, single-point architectures usually implement *raster scanning* to acquire rectangular surface regions. Scanning is either implemented by means of very fast *x, y* piezoelectric transducers (PZTs) or by a motorized *x, y* table. Wide-field architectures do not need raster scanning to acquire a single region, but may still have *x, y* actuators to translate the specimen (or the probe) so that multiple regions can be acquired in sequence, and collated into a single representation, through appropriate software programs (this is known as stitching).

3.2.3.1 Single-point Focus-detection Profilometers

Early attempts at developing profilometers adopting optical focus-based techniques can be traced back to the 1960s (Minsky 1961; Dupuy 1967). *Autofocusing* technology has been studied for quite some time in the digital imaging industry and, to date, a wide variety of sensoring techniques exist that are currently based on some form of focus detection. *Active autofocusing* solutions involve focusing a light beam (either a laser or not, visible or infrared; ultrasound may also be used) onto a point on the surface and using reflected light to obtain a distance measurement. A simplified, example architecture for a possible implementation of a focus-detection probe is illustrated in Figure 3.7. A light beam is focused on a plane at a specified distance from the probe (the *focal plane*). If the region to be measured is placed exactly on the focal plane, incident light will form a small, bright spot on it, and reflected light will form a small, bright spot at the detector as well. Conversely, out-of-focus conditions, due to the measured surface region being slightly above or below the focal plane, will result in a wider, weaker spot at the detector.

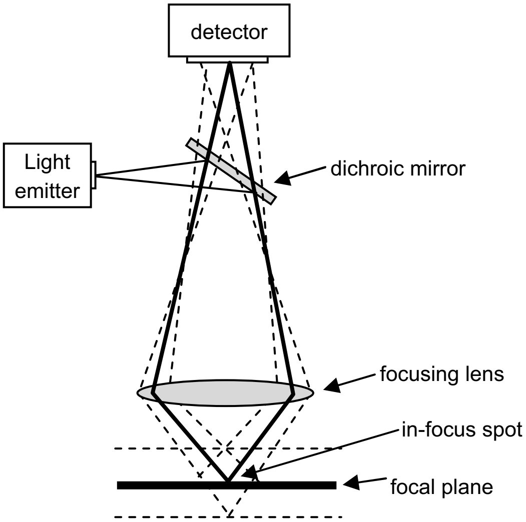


Figure 3.7 Structure of an optical system based on simple focus detection

As stated earlier, many types of focus-detection probes are available since there are many ways to analyze the reflected light pattern formed at the detector, and many ways to relate it to vertical displacement from the focal plane. Although one may measure the degree of defocusing at the detector and turn it into a vertical distance between the measured region and the focal plane, it is usually preferred to implement a controller-driven lens that is translated vertically until the optimal focus conditions are achieved at the detector. At that point, height information is computed directly from the vertical travel of the lens. Vertical translation can be assigned to the probe or to the specimen; traditionally actuated solutions (electric motors) are slower but allow for greater translations, while PZTs allow for faster movements but more limited ranges (better for measuring flatter surfaces).

Profilometers based on focus-detection probes such as the one described above are implemented as single-point measurement instruments (see Figure 3.6a); the raster scanning velocity is somewhat limited as the probe must be allowed to have enough time to achieve the optimal focus conditions at each *x, y* coordinate.

3.2.3.2 Wide-field Focus-detection 3D Microscopes

While solutions such as the one implemented by the simplified probe depicted in Figure 3.7 have been classified as active autofocusing solutions, since they involve the emission of a ray or light and the analysis of the light reflected back from the specimen surface in order to assess focus conditions, and thus distance, *passive autofocusing* solutions also exist, where the focus conditions are detected by analyzing information pertaining to the digital image formed at the detector: in *contrast measurement*, for example, a region of an image is in focus when there is there maximum intensity difference between two adjacent pixels; in *phase detection* two images are formed by dividing the incoming light – images are aligned and compared in terms of phase differences in order to identify in-focus regions.

Wide-field focus-detection 3D microscopes can be implemented that make use of passive autofocusing solutions: a sequence of images of the specimen are taken while the optics (or the specimen) are lowered/raised by means of a motorized *z*-axis, thus changing the focal plane at each image of the sequence. For each image, in-focus regions (pixels) are detected and associated with a reference height. Data extracted from the series of images are then combined into a single height map.

3.2.3.3 Confocal Laser Scanning Microscopes

Confocal laser scanning microscopes are microscopes whose probe implements a variation of the focus-detection technique already illustrated (Hamilton and Wilson 1982); the variation is summarized in Figure 3.8. A laser source is adopted to ensure that a sufficient amount of energy is received at the detector. The most important difference with respect to a simple focus-detection solution is that *pinholes* are placed on the light paths to prevent the out-of-focus light from reaching the detector. In this way the capability of discriminating in-focus and out-of-focus conditions is dramatically increased, as a signal will hit the detector only if the height of the surface point currently targeted by the incident light corresponds exactly to the focal distance.

Instruments adopting confocal probes are implemented as single-point instruments. Raster scanning is achieved by deflecting the beam toward each surface point through PZTs. Since the vertical movement of the focusing lens (needed to achieve optimal focus conditions at each surface point) is much slower than the PZT-actuated raster scanning, it is not convenient to stop at each *x, y* point and wait for focus detection. Instead a process is implemented where the focusing lens is kept at constant *z* height, and a complete *x, y* raster scanning is rapidly done; the result is a grid of on–off "pixel" values, where only the surface points lying on the focal plane have been detected. Then, the focal plane is shifted (by raising/lowering the focusing lens) and the entire scanning process is repeated, resulting in a new set of in-focus points. This process is also known as *z-slicing*. After a predefined number of *z* steps, the slices corresponding to each *z* position can be recombined into a single height map covering the entire measured surface.

The confocal laser scanning microscope was previously indicated as a microscope capable of providing also *volume data*. In fact, if the material is partially transparent to the laser light source, some in-focus information will hit the detector even if the actual surface point is above the focal plane; this information is related to the material that lies underneath the surface, at the correct focal distance. As a result, each *z* slice contains additional information that can be used to construct full 3D *volume data* representations (*voxel* data) of the specimen. In this case, surface topography information can be extracted by considering the point with the highest *z* coordinates at each *x, y* column.

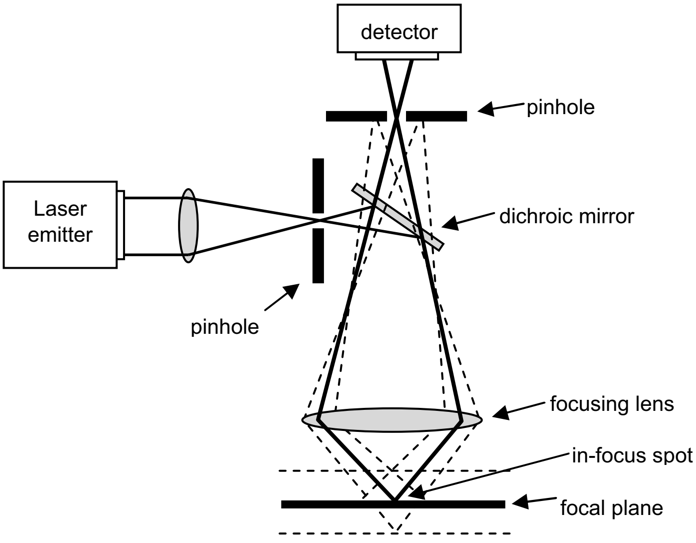


Figure 3.8 Architecture of a confocal probe

3.2.3.4 Chromatic Aberration Profilometers

Also based on a variation of the focus-detection principle, these profilometers are usually built as single-point measurement instruments and are equipped with a motorized *x, y* table for raster scanning. The technique implemented by the probe, also known as *chromatic distance measurement* (Molesini *et al.* 1984), is based on a focusing lens with high chromatic aberration, *i.e.*, having different, wavelength-dependent, refractive indices, which in turn results in wavelength-dependent focal lengths (see Figure 3.9). As white light traverses the lens, hits the surface, is reflected back, and is captured by a sensor, the spectrum of the signal contains a peak that indicates the in-focus wavelength; from such a peak wavelength it is possible to retrieve the distance measured by the probe.

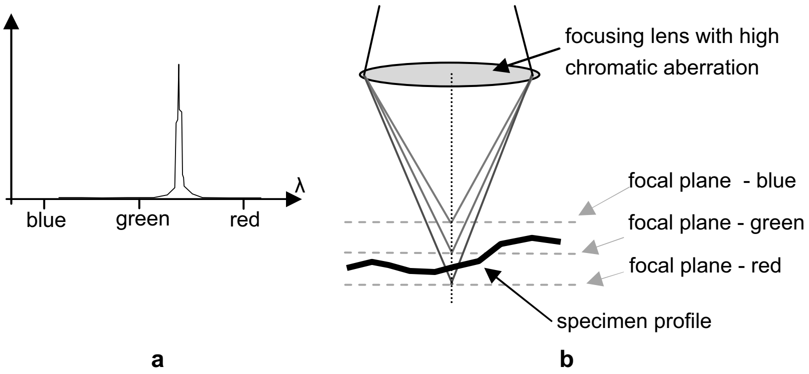


Figure 3.9 Chromatic aberration probe for point-based distance measurement: **a** example frequency spectrum of the signal received at the detector, and **b** position of the focal planes

3.2.3.5 Interferometric Profilometers

Interferometric profilometers are profilometers equipped with an optical probe whose principle of operation is based on interferometric principles (Hariharan 1985).

There are several ways interferometric effects can be applied to obtain profiles; the most widespread approach is based on implementing a *two-beam interferometry architecture*, directly derived from the Michelson interferometer, and illustrated in Figure 3.10. In the two-beam architecture, light emitted by a source and traveling as a parallel beam is split and sent toward the specimen surface and a reference mirror. Reflected by both, the light beams are recombined and finally sent to a detector, where a phase interference pattern (also called a *fringe image*, or *interferogram*) is formed. The constructive/destructive interference effect observed at each pattern point is due to the difference in the lengths of the paths followed by the corresponding light rays traveling within the two split beams (they are called the *reference path* and the *measurement path*). Such a difference in the path lengths can be converted into a height measure for the corresponding specimen surface point, as long as the position of the corresponding point on the reference mirror (*i.e.*, the length of the reference path) is known.

A common variation of the architecture depicted in Figure 3.10 specifically designed to operate at the micro and submicro scales is illustrated in Figure 3.11a and is known as the *Mirau interferometer*, or *Mirau objective* (Bhushan *et al.* 1985).

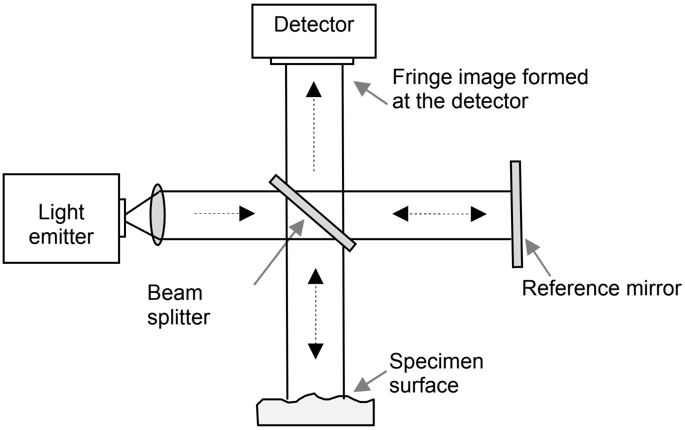


Figure 3.10 An interferometric profilometer implementing the two-beam architecture

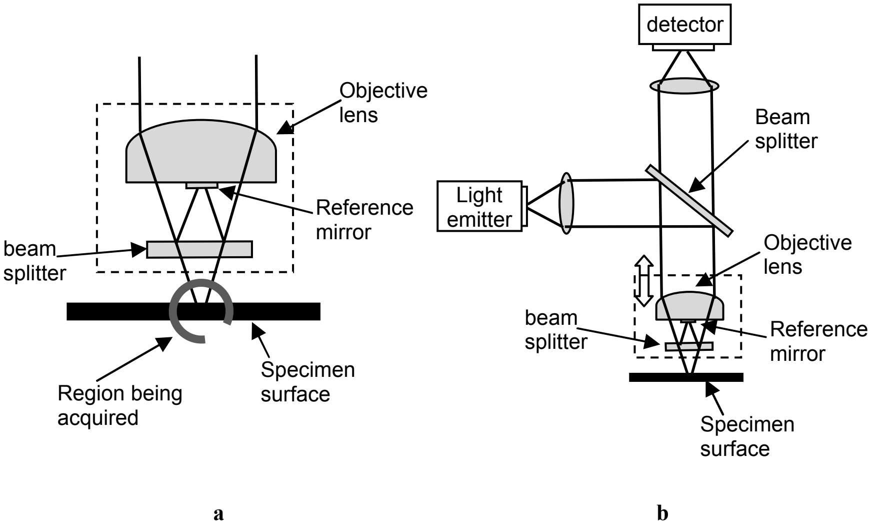


Figure 3.11 Mirau interferometric microscope architecture: **a** Mirau objective (Mirau interferometer), and **b** microscope assembly

Figure 3.11b shows the architecture of a profilometer instrument equipped with such an objective. In the Mirau objective (see Figure 3.11a), a lens is used to concentrate the beam onto a small region of the specimen; the reference mirror is mounted on the objective lens and the beam splitter is oriented so that the reference and measurement paths are aligned about the same vertical axis (which makes it easier to manufacture the objective with high precision).

Irrespective of whether the Mirau or the Michelson microscope architecture is adopted, the main issue resides in the interpretation of the interferogram and its translation into a topography height map. Interference effects are intrinsically periodic. They are related to the wave properties of the light signal, the range of wavelengths emitted by the light source, their coherence lengths, the overall energy of the emitted beam, the reflective properties of the surface, energy dispersion factors, as well as instrument manufacturing errors; all such things make the problem of reliably turning an interference pattern into a topography height map complicated.

Two approaches have become mainstream in recent years: *phase shifting interferometry* (PSI) (Greivenkamp and Bruning 1992) and *vertical scanning interferometry* (VSI) (Harasaki *et al.* 2000). In both cases the interferogram is acquired digitally at the detector, on an *x, y* grid where each grid cell corresponds to a single *x, y* point on the measured surface (the pixel spacing is related to the surface point spacing, depending on the magnification lenses). Both solutions start from the assumption that a single interferogram may not be enough to obtain a satisfactory reconstruction of surface topography; relying on a two-beam architecture, they operate by changing the length of one of the two paths (either the measurement or the reference one) while keeping the other fixed, and by using all the interferograms generated during the process to reconstruct the topography. The path lengths are changed by means of high-precision actuators (generally PZTs): actuators can be used to lower the probe toward the specimen (see Figure 3.11b for an example with a Mirau probe), raise the specimen toward the probe itself, or move the reference mirror, the overall effect being theoretically equivalent.

In PSI the key principle is that the *z* height of a single *x, y* surface point can be related, through a simple mathematical expression, to the change of intensity observed at the corresponding *x, y* point in the multiple interferograms obtained at different lengths. As few as three interferograms are needed to reconstruct the *z* height of each surface point. PSI solutions often privilege the use of *laser light sources*; laser light favors the formation of clear interferograms, owing to its high energy, good collimation, limited bandwidths and long coherence lengths; laser sources allow for achievement of a great surface height discrimination power, but at the price of a very limited *z* range. In fact, interference patterns are repeated at multiples of the light wavelength, and laser long coherence lengths make such repeating patterns almost indistinguishable from each other.

VSI starts from the assumption that for a given combination of measurement and reference path lengths (usually achieved by moving either the probe or the specimen) only those surface *x, y* points whose *z* height makes the two beam lengths exactly identical will result in maximum constructive interference at the interferogram. As the probe (or the surface) translates vertically by a step (hence *vertical scanning*), local maxima (constructive interference) in the interferogram allow for identification of those surface points whose height matches the above-mentioned conditions; their actual *z* height can be retrieved by combining the information obtained by the PZT with the known reference path length. The process continues until all the *x, y* pairs populating the detector grid have been assigned a *z* coordinate. Different algorithms exist for detecting with effectiveness and reliability maximum constructive interference points in an interferogram obtained at a given *z* position of the probe, for example *coherence correlation interferometry* (Taylor Hobson 2009). In terms of the light source, recent mainstream VSI solutions have privileged the use of white light; VSI using white light is often referred to as *white light scanning interferometry* or *scanning white light interferometry*. White light encompasses a wide array of wavelengths (theoretically, all of them; however, as a side note, some instruments are currently narrowing the bandwidth adopting colored light), and more importantly is characterized by short coherence lengths. A short coherence length means that a large positive interference effect will be visible only when the measurement and reference path lengths match with great accuracy, while it will decay significantly when the measurement length is equal to any multiple of the reference length. White light is therefore ideal for conditions where only the perfect match must be reliably identified, which is compatible with the mode of operation of VSI solutions.

Instruments adopting PSI or VSI usually acquire the points located within the measured region simultaneously, each point being associated with a specific *x, y* grid cell at the detector (*i.e.*, they can be classified as wide-field instruments); larger surface regions may be acquired by implementing stitching solutions paired with motorized *x, y* tables.

Other types of profilometers use similar interferometric principles to implement *single-point measurement* approaches (*i.e.*, they are equipped with interferometric probes for single-point distance measurement); in this case a PZT or motor is used for vertical translation of the probe (in the search for maximum constructive interference), while a motorized *x, y* table is used to implement single-profile or raster scanning.

3.2.3.6 Conoscopic Holography Profilometers

Conoscopic holography profilometers are a particular type of profilometer usually based on the single-point measurement architecture, and whose probe implements distance measurement through conoscopic holography (Sirat and Paz 1998).

The term "*holography*" refers to a particular interferometric technique for reconstructing information about a 3D shape. In ordinary holography, coherent light emanating from a source region is caused to interfere with a coherent reference beam in order to construct an interferogram in which the 3D characteristics of the source region are encoded.

Conoscopic holography is an original approach to holography aimed at obtaining a single-point distance measurement solution. The most notable variant introduced with respect to traditional holography consists in obtaining the holographic interferogram without an external reference beam; instead, the reference beam is created directly from the light emanated from the source region.

Figure 3.12 illustrates the basic architecture of a conoscopic holographic probe: laser light is directed toward the specimen surface and is focused onto a single spot; scattered light is reflected according to a spherical angle, and is collected through a lens into a parallel beam (hence the term "conoscopy": observe – scopy; through a cone – cono). After going through the beam splitter, the light passes through the main element of uniqueness of the architecture: an optical assembly made of a lens, two polarizers, and a uniaxial, birefringent crystal; the assembly is responsible for creating the reference beam from the original beam, and for constructing the holographic interference pattern.

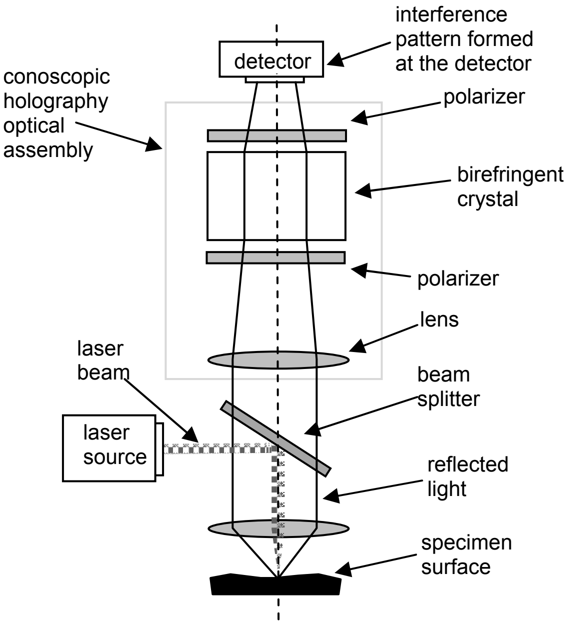


Figure 3.12 Architecture of a conoscopic holography probe

In detail, the first lens of the assembly is responsible for reproducing a conicity effect in the parallel beam, so that each ray is sent out at different angles. The polarizer splits each ray into two components of different polarization and sends them to the crystal. The crystal receives the two components and, since it is birefringent (*i.e.*, double refractive), slows down one of them, inducing a phase shift, which is proportional to the incidence angle of each ray hitting the uniaxial crystal. The phase shift creates the holographic interference effect between the two light components, one acting as the reference beam for the other, when the two phase-shifted components are recombined by the second polarizer. Finally, the interference ray exiting from the second polarizer is sent to the detector, where it forms a high-contrast interference pattern together with all the other rays. The interference pattern is analyzed by a computer to obtain a distance measurement for the surface point being targeted by the spot.

3.2.4 Performance and Issues of Measuring with Optical Profilometers and Microscopes

3.2.4.1 Measurement Performance

As for stylus-based instruments, also for optical instruments a wide array of ranges and resolutions are available; the performance depends essentially on the specific optical technique adopted by the instrument.

For focus-detection and confocal instruments, a survey of current offerings by several manufacturers (Carl Zeiss 2009; Fries Research & Technology 2009; Hirox 2009; Leica Mikrosysteme Vertrieb 2009; Olympus 2009; Sensofar Tech 2009; Solarius Development 2009; Veeco Instruments 2009) shows that the best performing instruments may achieve vertical resolutions smaller than 1 nm over a vertical range below 5 μm, while coarser instruments are found with vertical resolutions of 1–2 μm, operating over a vertical range of approximately 30 mm. The lateral resolution depends on several factors, and on whether the architecture is wide field or single point. For single-point architecture, significant constraints on the lateral resolution come from the probe spot size (the maximum resolution may be limited by measurement averaging effects within the spot region) and *x, y* table/probe positioning mechanisms; for wide-field instruments, probe components such as the magnification lens and the pixel spacing at the detector play a relevant role. For high-performance focus-detection and confocal instruments, a maximum lateral resolution below 1 μm can be achieved. Analogously to stylus-based systems, the maximum lateral range is constrained by the vertical range of the probe, since the likelihood of out-of-range measurement errors caused by specimen leveling errors increases with the width of the region measured.

Chromatic aberration sensors (Fries Research & Technology 2009) are characterized by vertical resolutions varying from approximately 3 nm (within a range below 300 μm) to 250 nm over 25 mm for coarser probes. The maximum horizontal resolution is within the 1–14-μm range (usually, the higher the vertical resolution, the higher the lateral resolution).

Interferometric instruments (Ambios Technology 2009; FOGALE nanotech 2009; Fries Research & Technology 2009; Novacam Technologies 2009; Sensofar Tech 2009; Solarius Development 2009; Taylor Hobson 2009; Veeco Instruments 2009; Zygo 2009) are characterized by the highest vertical resolutions among all optical instruments: vertical resolutions as low as 0.01 nm over a range of approximately 100 μm can be achieved by some configurations; the vertical range can be dramatically increased by means of *z* stitching. The maximum lateral resolutions are less than 1 μm, while the lateral range varies greatly in the 0.03–200-mm range, usually depending on the vertical resolution of the instrument and on the availability of *xy* stitching options.

Conoscopic holography probes (Optimet Optical Metrology 2009) are available with different vertical resolutions (depending on the lens assembly type): the highest resolutions can be less than 0.1 μm over ranges smaller than 1 mm, while the coarsest resolution is about 1.5 μm over a range below 200 mm. The maximum lateral resolutions are within the 5–100-μm interval, depending on the spot size/lens assembly.

Concerning accuracy and precision, the main concern for the optical techniques illustrated above is mainly related to traceability, which is not as good as for stylus-based instruments. Although several calibration activities can be carried out with the same traceable physical standards as are adopted for calibrating stylus-based instruments, not enough documentation is currently available concerning measurement protocols, architecture, and modes of operation of these classes of instruments. This is mainly due to the fact that the application of these noncontact techniques in industrial metrology is evolving very rapidly, especially for operating at small scales and high resolutions. Innovative solutions and new instruments are being produced at very fast rates. Given these premises, especially for some more recent instruments, their application is predominantly confined to laboratory and research work. Nevertheless, international standards concerning optical instruments for surface microtopography analysis have recently appeared for confocal instruments (ISO 25178-602:2010), and are under development for interferometric and point autofocusing instruments (ISO/DIS 25178-603, ISO/CD 25178-604 and ISO/CD 25178-605). The development of international standards is a fundamental premise toward a more significant introduction of these instruments in current industrial practice.

3.2.4.2 Constraints on Material Geometry and Surface Topography of the Part To Be Measured

Although several classes of optical instruments are quite similar in performance to stylus-based instruments, the two classes are hardly interchangeable. The physics involved in measurement is very different between contact and noncontact techniques. Optical techniques are intrinsically less robust than contact techniques. The performance of an optical probe is greatly influenced by the optical properties of the surface to be measured. Reflectivity of the specimen surface is a key property for the successful application of focus-detection and interferometric techniques as it affects the overall amount of light which is captured at the detector. Low-reflectivity materials are particularly problematic; even for highly reflective surfaces, sharp features (*e.g.*, vertical steps) may alter the reflectivity locally, and thus result in measurement errors. Small features (with respect to lateral resolution power/focus spot size), for example, small holes and pins, may also result in errors, or be averaged out during measurement. One of the key performance aspects that must be considered when assessing the applicability of an optical measurement technique is referred to as the *maximum detectable slope*: information taken from a recent review paper (Hansen *et al.* 2006) indicates that focus-detection instruments can detect a maximum slope of approximately 15°, while confocal microscopes achieve up to 85°; interferometric instruments (white light scanning interferometry in particular) lie somewhere in between, with detectable slopes of approximately 30°. Conoscopic holography probes may achieve detectable slopes up to 85° (Sirat and Paz 1998).

Heterogeneity of the measured material may cause problems as well, again leading to local variations of reflectivity and optical constants. Impurities (dust, liquids, and particles) may also cause local measurement errors, like for contact-based measurement, and often require the surface to be cleaned before measurement. For the same reasons, optical instruments are in general more sensitive to environmental conditions, including atmosphere composition, and ambient illumination.

3.2.5 Nonoptical Microscopes

The scanning electron microscope (SEM) is type of electron microscope that acquires an image of the surface by scanning it with a high-energy beam of electrons, in a raster scan pattern (Thornton 1968). It is classified as a nonoptical microscope since light is replaced by a beam of electrons. The SEM is not capable of acquiring a 3D profile from a specimen surface topography *per se*; however, if properly equipped and configured, a SEM can produce a 3D topography representation by means of a tilting specimen holder (tilting table) and the application of *stereophotogrammetry*, giving rise to a *stereoscopic SEM* (stereo SEM).

The basic architecture of a SEM fit for 3D stereo scanning is illustrated in Figure 3.13. A high-energy beam of electrons, known as the *primary beam*, is generated by an electron gun acting as the cathode and is accelerated toward the specimen surface, which acts as the anode. The beam is guided by magnetic fields and is focused onto a single small point of the specimen surface (the spot being usually a few nanometers in diameter). Magnetic scanning coils control the beam deflection and drive the beam spot over the surface according to an *x, y* raster scanning pattern.

As the primary beam hits the surface, different types of emissions occur owing to the interaction between the high-energy electrons of the beam with the atoms of the specimen. Three main types of emissions can usually be captured by equipping the microscope with suitable detectors: *primary electrons* (electrons belonging to the primary beam), which are reflected by the specimen surface (this is known as elastic scattering and these electrons are also called *backscattered electrons*); *secondary electrons*, which are produced by the surface itself thanks to the energy received from the primary beam (this is known as inelastic scattering); and *X-rays* generated as a consequence of the interaction. Each main emission type carries different types of information. Primary electrons and X-rays provide indications on the chemical composition of the specimen; secondary electrons can be used to obtain indications of the local slope, as the intensity of their emission is related to the incidence angle between the primary beam and the specimen topography at the spot. Given this latter point, the preferred approach for obtaining surface topography information is to record the emissions of secondary electrons during a raster scanning process, and use the information to produce a grayscale image that resembles the topography, as if it were observed by optical means.

The maximum lateral resolution that can be achieved with a raster scanning process depends mainly on the energy of the electron beam: compromises must be found between the desired resolution and the risk of damaging the specimen.

Scanning electron microscopy requires the specimen to be conductive (since it acts as the anode), and measurement must take place in a vacuum, to avoid degrading the electron beam. Both aspects pose constraints on the specimen: non-conductive surfaces can be coated, but the coating layer introduces changes in the topography which may be undesirable, especially when measuring at such small scales; specimens with risk of releasing vapors need to be dried or frozen. A variation of a traditional SEM, the *environmental SEM* (ESEM), allows for operation at lower pressures and humid atmosphere (Danilatos and Postle 1982) by adopting more sophisticated detectors.

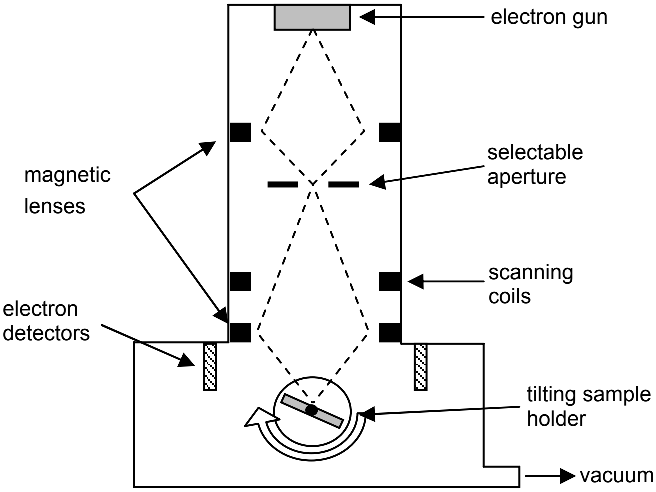


Figure 3.13 Architecture of a stereo scanning electron microscope

To reconstruct a 3D model of surface topography, the *stereo SEM* allows at least two images to be acquired from the same specimen oriented at two different tilt angles (usually 1–7° apart) with respect to the direction of observation. Once the two images are available, 3D reconstruction can take place by means of stereophotogrammetry algorithms (Hudson 1973). The basics of triangulation state that once the *x, y* position of the *same* surface point is known in two different images taken from the same specimen oriented at two different tilt angles, then its height can be computed through simple trigonometric relations. In the practical application of the approach to stereo images obtained through use of the SEM, the main concern is to make sure the same topography point has been located in both images. Point localization is typically the weakest point of the approach: it must be robust and extremely accurate, since the tilt angle is small, and so are the typical topography height variations to be trigonometrically reconstructed. In some cases, point correspondences between images cannot be found: this happens, for example, when tilting results in a point obstructed by other surface features (which imposes a limitation on surface topographies that can be analyzed and on maximum tilting angles). Even when points are not obstructed, correspondence localization may be a daunting task: except for where the specimen topography shows clear reference marks that can help locate specific surface points, the robust and effective identification of point correspondences relies heavily on pattern identification algorithms, and is still an active subject of considerable research work.

3.2.6 Performance and Issues of Measuring with Nonoptical Microscopes

3.2.6.1 Measurement Performance

The only type of nonoptical microscope discussed in this work is the SEM, which by itself is only a 2D imaging device and thus can only be described in terms of lateral resolution. As stated earlier, the main factor in determining lateral resolution is electron beam energy. An overview of the offerings of several manufacturers indicates maximum lateral resolutions in some cases smaller than 1 nm, or more often of a few nanometers depending on the beam energy (Aspex 2009; CamScan Electron Optics 2009; FEI 2009; Hitachi High Technologies America 2009; JEOL 2009; Nanonics Imaging 2009; VisiTec Microtechnik 2009). The actual performance may be limited in some real-life application scenarios for materials not withstanding the energies required to achieve the highest resolutions. The lateral range for typical SEMs is a few hundred micrometers. ESEM performance varies compared with that of conventional SEM: the energy of the electron beam is dissipated more in ESEM instruments owing to the atmosphere, thus potentially resulting in lower contrast. Lateral resolution in ESEMs can be lower than, approximately equal to, or higher than in a conventional SEM, depending on beam energy.

Vertical resolutions for SEM devices producing 3D output by means of the stereo-pair technique have been reported varying from a few nanometers to a few tens of nanometers. However, the performance may be significantly affected by several factors: pixel pair-matching may not work as desired, and even assuming that an ideal pixel pair-matching performance could be achieved, the smallest measurable height difference would be related to the smallest measurable slope, and thus to a combination of pixel color resolution, spacing, and the topographic properties of the surface measured. Additional issues that may degrade the resolution are related to different focus conditions associated with the image pair, which in turn may lead to slightly different magnifications. ESEM instruments usually produce noisier images than conventional SEM instruments, thus affecting also the quality of 3D reconstruction.

Some information is available concerning the accuracy and precision of conventional SEMs, and nominal data are available from most manufacturers. Calibration artifacts are commercially available for SEMs operating as 2D imaging devices, which reproduce various types of 2D patterns. However, scarce information exists for stereo SEM solutions, and – as stated earlier – the performance remains strongly related to the surface type (material and texture).

3.2.6.2 Constraints on Material, Geometry, and Surface Topography of the Part To Be Measured

The two main drawbacks of conventional scanning electron microscopy are that the specimen must be conductive (since it acts as the anode), and that measurement takes place in a high vacuum, to avoid degrading the electron beam. Both aspects pose constraints on the specimen: nonconductive surfaces can be coated, but the coating layer introduces changes in the topography which may not be desirable, especially when measuring at such small scales; specimens with risk of releasing vapors need to be dried or cryogenically frozen. ESEM could be used to reduce the need for vacuum conditions; the price to pay is a more significant dissipation of electron beam energy. High-energy beams are necessary to achieve the highest resolutions. Again, not all materials may withstand such energies without experiencing degradation effects. Three-dimensional surface topographies obtained by stereo pairs through photogrammetry are seldom reliable enough for extended quantitative inspection tasks, owing to the number of issues related to the identification of correct pixel pairs and the significant reconstruction errors associated with wrong matches. For this reason, 3D surface topographies reconstructed through stereophotogrammetry on SEM images are mainly confined to laboratory usage and research work. However, qualitative inspection of surface topography by SEM 2D imaging remains one of the strongest and most useful approaches and often provides invaluable support to other 3D surface topography measurement techniques.

3.2.7 Scanning Probe Microscopes

The term "*scanning probe microscope*" (SPM) encompasses a class of instruments for measuring surface topography at submicrometric scales. In terms of general architecture, SPMs are essentially Cartesian instruments equipped with a single-point distance measurement probe. Their peculiarity is the probe itself, which makes them capable of acquiring topography information at very small scales, typically down to the atomic level, albeit with limited *z* ranges and on surface regions of limited *xy* sizes.

Topography is usually acquired through raster scanning with uniform point *xy* spacing; the resulting topography height map can be equivalently handled as a set of points in 3D space, or as a grayscale digital image.

The two most popular SPMs are the *scanning tunneling microscope* (STM) (Binnig and Rohrer 1983) and the *atomic force microscope* (AFM) (Binnig *et al.* 1986). In both cases the probe consists of an atomically sharp, tiny tip, placed very close to the measurement surface.

3.2.7.1 Scanning Tunneling Microscopes

The working principle of a STM is illustrated in Figure 3.14. By means of computer-controlled PZTs, the tip, which is made of conductive material, is kept in close proximity (a few angstroms) to the specimen surface, which must also be conductive. A voltage is applied to the tip and to the surface; since the tip and the surface are very close, a small tunneling current is created that flows through the gap dividing the two. The tunneling current is very small, but it is very sensitive to small variations of the gap. The STM can operate in two ways: by keeping the probe at a constant height and measuring surface height changes in terms of changes of the tunneling current (*constant height mode*), or it can follow surface reliefs by moving vertically with the PZTs, in order to keep the tunneling current constant (*constant current mode*). The critical aspect of a good STM is the sharpness of the stylus tip: ideally it should be made of a single atom, but in practice this is very difficult to manufacture, and tip geometry greatly affects the measurement results. Secondly, the capability of bringing the tip very close to the surface is critical. Commercial instruments adopt a coarse translation mechanism to bring the tip quite close to the surface, and a fine translation solution implemented by means of the PZTs mentioned above; when the tip is at the right distance, PZTs are also used to implement raster scanning on the *x, y* plane.

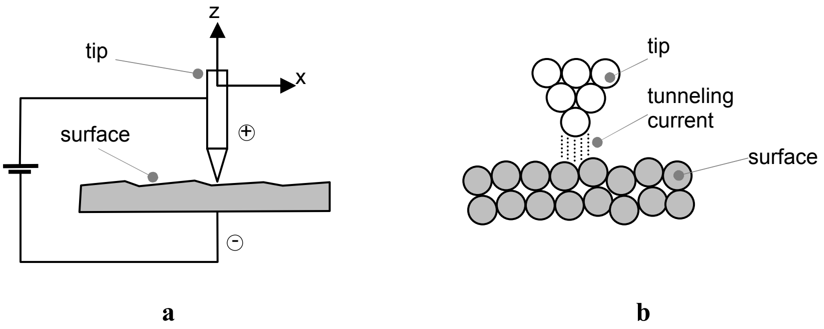


Figure 3.14 Scanning tunneling microscope: **a** overall schema, and **b** close-up view of the tip–surface interaction

3.2.7.2 Atomic Force Microscopes

The most widespread SPM alternative to the STM is the AFM, which is essentially a derivation of the STM itself. In the AFM (see Figure 3.15), the probe consists of a microscopic cantilever terminating with a tip, with a radius of a few nanometers. The tip is brought very close to the surface, so it is affected by the interatomic forces existing between the atoms of the surface and the tip itself. As the tip approaches the surface, the atomic forces are initially attractive, and then become repulsive, which is basically "contact" at such small scales. Atomic forces can be indirectly measured from the deflection of the cantilever, since its stiffness is known. The deflection, in turn, can be measured in several ways, the most common being by means of a laser beam reflected by the cantilever and hitting an array of photodiodes (see Figure 3.15a).

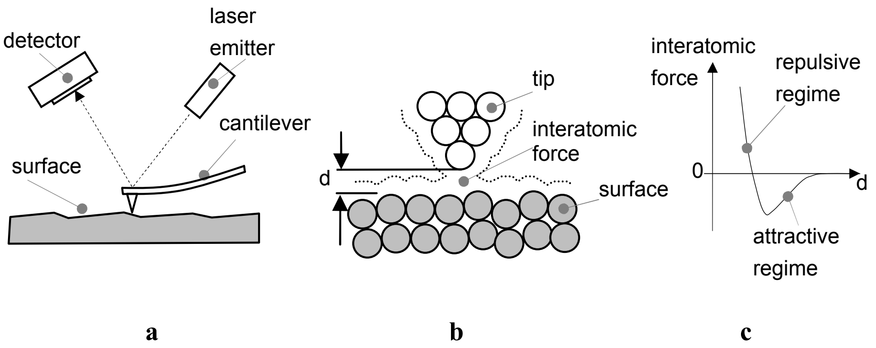


Figure 3.15 Atomic force microscope: **a** overall schema, **b** close-up view of the tip–surface interaction, and **c** interatomic force curve

Several approaches exist for acquiring a surface topography. Recall the traditional stylus-based profilometer illustrated in Section 3.2.1. In the stylus-based profilometer, deflection is measured in an open-loop manner while the tip is translated at constant reference *z* height over the surface; this is typically not applicable in AFMs, given the increased risk of collision and damage of the delicate tip at such small scales. Instead, a wide variety of closed-loop solutions are available, the most important being listed in the following.

In *contact mode* (or *static mode*), the tip is so close to the surface that the atomic force is repulsive (see Figure 3.15c) and deflects the cantilever upward. A PZT raises/lowers the cantilever with respect to the measured surface point so that the deflection force is kept constant. The vertical displacements generated by the transducer during traversal are recorded as height points, and are used to reconstruct the surface topography.

In *noncontact mode*, the tip is subjected to forced oscillation and is kept at a distance from the surface so that the atomic forces are mainly attractive; changes in the resonant frequency or amplitude in the cantilever oscillation can be related to changes of material and topographic properties of the underlying surface point.

In *tapping mode*, the tip is subjected to forced oscillation and is placed at a distance from the surface so that, while oscillating, it is partially subjected to attractive forces, and partially to repulsive forces (this is known as intermittent contact); again the oscillation can be studied to obtain surface material and topographic properties. With respect to the contact mode, the tapping mode eliminates lateral forces, such as drag, and reduces the risk of damaging the surface.

Additional approaches exist that apply variations of the illustrated AFM techniques for obtaining surface information that go beyond simple topography (such as friction properties and elasticity); the most widely known include lateral force microscopy, force modulation microscopy, magnetic force microscopy, and electrostatic force microscopy. Since most such techniques can be obtained from the same basic SPM structure by simply changing the probe, several commercial multimode SPM instruments are currently available.

3.2.8 Performance and Issues of Measuring with Scanning Probe Microscopes

3.2.8.1 Measurement Performance

STMs have the highest resolutions amongst all the instruments illustrated in this chapter. A survey of the offerings of several STM and AFM manufacturers (Agilent Technologies 2009; AIST-NT 2009; Ambios Technology 2009; Asylum Research 2009; Bruker AXS 2009; Fries Research & Technology 2009; JEOL 2009; Micro Photonics 2009; Nanonics Imaging 2009; Nanosurf 2009; Novascan Technologies 2009; NT-MDT 2009; Park Systems 2009; Veeco Instruments 2009) shows that the maximum vertical resolution of a STM is approximately 1 pm, over a vertical range of 200 nm, and the maximum lateral resolutions are less than 10 pm. AFMs have slightly poorer performance, but still can achieve vertical resolutions of 0.1 nm, over a vertical range of 10–20 µm, with lateral resolutions of a few nanometers.

Information on accuracy and precision is available from several manufacturers; some high-resolution, traceable calibration artifacts are also commercially available that reproduce specific patterns that can be used to determine SPM performance experimentally. Once again, however, performance tends to vary significantly depending on materials and surface properties, as discussed also in the next section.

3.2.8.2 Constraints on Material, Geometry and Surface Topography of the Parts To Be Measured

As for the SEM, the main drawback of the STM is that the surface to be measured must be of conductive or semiconductive material. This greatly limits the applicability of the instrument. Nonconductive surfaces can be coated, but the coating layer introduces changes in the topography which may not be desirable, especially when measuring at such small scales. The AFM mode of operation makes the AFM appear very similar to a stylus-based instrument; thus, it may appear more robust and generally applicable, although at slightly inferior resolutions. However, the nature of the contact at such small scales makes the application of the AFM for quantitative measurement of surface topography slightly more complex, as the signal detected by the probe may be influenced by factors other than topography, such as material properties.

SPMs such as the AFM and STM are usually built so that measurement takes place in a confined enclosure, to achieve a great degree of separation from the surroundings; as a result, specimens are usually very small, and must be specifically prepared to be compatible with the instrument.

3.3 Application to the Inspection of Microfabricated Parts and Surface Features

The classes of profilometers and microscopes that have been illustrated so far have been conceived to fulfill specific functional roles in conventional quality control. Profilometers come from the domain of surface finish analysis at micro and submicro scales, where topography is seen as a combination of wavelike height variations describing roughness, waviness, and form error, and the main objective is to provide a quantitative measurement of such height variations. Microscopes are conceived to support visual inspection from a mainly qualitative standpoint, while limited quantitative measurement capabilities, if present, are usually implemented through photogrammetry. Three-dimensional microscope solutions are an improvement in this sense, as they are intrinsically more capable of quantitative measurements, since they produce height maps.

In order to assess the applicability of the illustrated classes of instruments to the quantitative evaluation of geometric error on microfabricated parts and surface features, some peculiar aspects, usually common to the instruments, must be analyzed first. Conceptual parallels to typical issues related to geometric error inspection on standard-sized parts are drawn afterward.

3.3.1 Aspects and Issues Peculiar to the Application of Profilometers and Microscopes

A selection of topics has been identified in this work, each constituting a relevant subject to be investigated when assessing the possibility to use profilometers and 3D microscopes for evaluating geometric error on microfabricated parts and surface features. These topics can be synthetically ascribed to the following terms:

- different measurement performance in *x, y* and *z*;
- unidirectional probing;
- raster scanning; and
- image-inspired data processing.

Each one raises specific issues, as discussed in the following sections.

3.3.1.1 Different Measurement Performance in *x, y* and *z*

It was shown that there is usually a significant difference in nominal performance between lateral measurement (*x, y*) and vertical measurement (*z*). This is usually by design, given the aforementioned relevance assigned to height measurement. This is particularly true for profilometers, where often *z* resolutions are 1 order of magnitude higher than lateral resolutions. Higher resolutions usually come at a price of more limited ranges: when measuring the geometries of several semiconductor products and microfabricated surface features this may not be an issue; however for surface features characterized by a *high aspect ratio* (*e.g.*, in MEMS), these differences between axes may become a serious problem.

Performance differences between axes can usually be ascribed to the different technologies adopted for acquiring point coordinates about *z* and about *x, y.* While vertical coordinates are acquired through the probe (thus the performance depends on probe technology), lateral coordinates are computed from lateral displacement of the probe/table. Coarser solutions for measuring lateral displacement make use of encoders that perform an indirect measurement of the displacement generated by PZT or electric motors. More accurate solutions are based on direct measurement of displacement and adopt additional distance measurement probes (*e.g.*, interferometric). Since different solutions are adopted for acquiring the *x, y* and *z* coordinates, the accuracy, precision, resolution, range, and any other property related to measurement performance for each coordinate are potentially different. Since traditional surface finish analysis is mainly concerned with the study of deviations of surface heights, it is more likely to have information about the measurement performance about the *z*-axis when using a profilometer, while information about the *x*- and *y*-axes may be lacking. Conversely, when a SEM or a similar device rooted in the 2D imaging realm is used, it is often more likely to provide metrological information about lateral accuracy and precision, rather than about height measurement performance.

3.3.1.2 Unidirectional Probing

Since profilometers and 3D microscopes are designed with the main goal of acquiring vertical distance (height) measurements, they are designed so that the probe approaches the surface from the same direction. This has several important consequences.

The first is a limitation on the types of geometric features that can be successfully measured, as *geometric undercuts cannot be acquired.* At any *x, y* position, the probe can usually measure only the distance to the closest surface point aligned on the vertical (*z*) axis. A notable exception is 3D microscopes capable of acquiring volume data (*e.g.*, confocal laser scanning microscopes), but only if the material properties are compatible with volume data acquisition.

The second consequence of unidirectional probing – actually, strictly related to the first – is that measurement performance is strongly related to the orientation of the surface being measured. Data on vertical or high-slope surfaces are usually difficult to acquire and measurement performance is degraded with inclination to the point that factors such as the *maximum detectable slope* are key for preferring a measurement technique over another. It is often problematic, if not impossible, to acquire data on complex multifaceted surfaces in a single measurement session (*i.e.*, without repositioning the probe or the specimen), or even data on single surfaces with high curvatures.

The third and last consequence of unidirectional probing is related to the nature of the geometric data available as a result of the measurement, which is often colloquially referred to as *2.5D geometry*, *i.e.*, not fully 3D geometry. Geometric models result from height maps and, again, the only way to obtain a full 3D geometric model of a specimen (*e.g.*, a microfabricated part) is to reconstruct it from multiple measurements (*i.e.*, multiple probe/workpiece orientations) and through stitching techniques in the attempt to compensate for the accuracy losses due to repositioning.

All the limitations of unidirectional probing have been long accepted in the domain of surface finish analysis, but less so in the new domain of quality inspection of microfabricated parts and surface features, where the relevance of such issues depends on the type of feature investigated and on the manufacturing process that was used to fabricate it. For manufacturing processes that are essentially 2.5D themselves (*e.g.*, *LIGA*, *etching*, *lithography*), unidirectional probing may be tolerated as long as the functional requirements and types of tolerances to be inspected can be expressed in a 2.5D space as well. The same considerations apply to the maximum measurable angle, as with some manufacturing processes high-slope surfaces cannot be obtained anyway. However, for some micromanufacturing processes that make it easier to obtain full 3D geometries (*e.g.*, *micromilling*, *laser micromachining*), the limitations of unidirectional probing may be unacceptable and more complex measurement strategies may need to be developed.

3.3.1.3 Raster Scanning

In the conceptual paradigm underlying surface finish analysis, the nominal geometry is assumed to be smooth and regular (*e.g.*, a horizontal plane) and error is assumed as a composition of height deviation components represented as harmonic oscillations of different amplitudes and wavelengths. Assuming such a paradigm exists, it is perfectly reasonable to implement a raster scanning process where the surface is sampled with uniform spacing according to a regular grid. Spacing allows for determining the range of wavelengths captured by the measurement process. When form error on an average-sized manufactured part is inspected, in contrast, the nominal geometry is usually provided in the form of a multifaceted geometric model, and edges, regions with high curvatures, and other significant features can be used to plan a measurement process optimized for the geometry at hand.

If profilometers and 3D microscopes are to be used to inspect microfabricated parts and surface features, this is like being in the condition where possibly complex geometries are to be measured with raster scanning, *i.e.*, rather suboptimal, strategies. An example of this type of problem is illustrated in Chapter 5. Trade-off solutions must be pursued to identify optimal sampling resolutions that take into account instrument capability, cost of the measurement process, representativeness of the result, measurement uncertainty.

3.3.1.4 Image-inspired Data Processing

Unidirectional probing and the overall differentiation introduced between the lateral (*x, y*) axes and the vertical (*z*) axis are also evident in the design and implementation of the software applications currently available for processing geometric data (height maps) resulting from measurement. Height map data processing is basically digital image processing adapted to operate on height values instead of grayscale values. While, on one hand, this is positive, as a great number of the techniques available from the literature on digital images have been ported to surface topography analysis, on the other hand, it implicitly drives the analysis to consider geometric data as images, and to treat any transform as a transform whose output is an image. To give a simple example, while alignment between geometries in a 3D space implies rigid roto-translation transforms, on images any translation and/or rotation takes place in the *x, y* plane, and usually implies pixel remapping, and potentially, resampling. In the domain of surface finish analysis, this image-inspired approach has usually been considered acceptable; when dealing with the assessment of geometric error on microfabricated parts, this may no longer be the case.

3.3.2 Aspects and Issues That Are Shared with Quality Inspection of Average-sized Mechanical Parts with Conventional Instruments

Some of the problems that must be faced and solved when applying profilometers and microscopes to the measurement of micromanufactured surface features and parts are actually also common to more traditional metrology applications, such as the measurement of an average-sized mechanical part by means of a CMM. In some cases, the solution is scale- and instrument-independent, in other cases, specific aspects must be taken into account.

3.3.2.1 Registration with Nominal Geometry

A shared problem when assessing geometric error on any part and with any instrument is being able to spatially relate (register) the measurement to a reference (nominal) geometry. The reference geometry could be provided as an explicit model (*e.g.*, CAD model for quality inspection of an average-sized mechanical part) or as a set of mathematical equations (a plane, a cylinder, a parametric surface, *etc.*). In conventional surface topography analysis at the micro and submicro scales, which is the domain from which profilometers and microscopes come, the reference geometry is usually assumed to be a plane, or similarly a simple surface (*e.g.*, second-order or third-order polynomial). Once properly placed with respect to measurement points, height differences are used to compute roughness, waviness, and form error parameters.

Regardless of the scale, application, and model type, the registration of the reference geometry and the measured points is almost invariably determined through some sort of point fitting. The way fitting is accomplished may change depending on the application. For example, in roundness evaluation least-squares fitting and minimum-zone fitting are the two most popular approaches; in surface finish analysis, since most of time the reference geometry is a plane with limitless extensions, least-squares fitting is used to properly place the reference surface in its correct vertical position, and no lateral alignment is necessary – this is mainly because in surface finish analysis we are mostly interested in deviations of surface heights.

When moving toward the inspection of micromanufactured surface features, more complex geometric models may be provided as the nominal reference, and z-only alignment may not be enough; the methods and approaches conventionally developed for geometric inspection with a CMM and similar equipment could be adapted to operate on profilometer- or microscope-generated data, provided that the image-inspired approach for processing geometric data is abandoned in favor of a fully 3D approach to data manipulation.

3.3.2.2 Data Stitching

As mentioned several times earlier, data stitching is a process that is used to register data from different measurements so that wider portions of a given geometry can be obtained even from instruments with more limited ranges. Stitching is based on the alignment of those portions of each data set which are known to be derived from the same region of the measured geometry, thus implying that the measurement itself must be planned so that it provides some degree of overlapping. Stitching techniques have always been popular in geometric reconstruction, as they allow a fully 3D representation of a manufactured part to be obtained by collating measurements taken from different directions. In the domain of surface microtopography analysis by means of profilometers and microscopes, the application of stitching techniques is more recent, especially in noncontact instruments, and is mainly aimed at achieving high resolutions together with high measurement ranges.

When dealing with the inspection of microfabricated parts and surface features by means of profilometers and microscopes, stitching techniques may play a fundamental role, especially when considering the aforementioned issues related to undercuts and maximum detectable slope, and when considering also the overall issues related to resolution and range for high-aspect-ratio surface features. Similar to the issues of registering the nominal and measured geometry, the problem of collating measurements together in a successful way is shared with conventional inspection of mechanical parts, and most of the algorithmic solutions developed are scale-independent.

3.4 Conclusions

Profilometers for microtopography analysis and 3D microscopes are increasingly being used for assessing geometric error on semiconductor products, MEMS, and other microsystems, and microcomponents in general, owing to their intrinsic capabilities of operating at micrometric and submicrometric scales. However, as their domain of application is shifted from the assessment of surface finish/visual inspection to more quantitative metrological tasks, their architectures, performance characteristics, and applicability constraints must be carefully investigated. In this work, the main types of profilometers and 3D microscopes which are commercially available and routinely applied in typical industrial scenarios were reviewed and critically analyzed. The analysis highlighted a series of fundamental issues, partly specific to these types of instruments, and partly shared with the instruments routinely involved in assessing geometric error on standard-sized mechanical parts.

A careful analysis of such issues and the identification of proper strategies to handle them are seen as fundamental steps toward the development of successful micrometrology solutions to the problem of assessing geometric error on micromanufactured parts and surface features.

References

ISO 3274 (1998) Geometric product specifications (GPS) – surface texture: profile method – nominal characteristics of contact (stylus) instruments. International Organization for Standardization, Geneva

ISO 4287 (1997) Geometrical product specifications (GPS) – surface texture: profile method – terms, definitions and surface texture parameters. International Organization for Standardization, Geneva

ISO 4288 (1996) Geometrical product specifications (GPS) – surface texture: profile method – rules and procedures for the assessment of surface texture. International Organization for Standardization, Geneva

ISO 5436-1 (2000) Geometrical product specifications (GPS) – surface texture: profile method; measurement standards – part 1: material measures. International Organization for Standardization, Geneva

ISO 5436-2 (2001) Geometrical product specifications (GPS) – surface texture: profile method; measurement standards – part 2: software measurement standards. International Organization for Standardization, Geneva

ISO 5725-1,2,3,4 and 6 (1994) Accuracy (trueness and precision) of measurement methods and results. Parts: 1,2,3,4 and 6. International Organization for Standardization, Geneva

ISO 5725-5 (1998) Accuracy (trueness and precision) of measurement methods and results – part 5: alternative methods for the determination of the precision of a standard measurement method. International Organization for Standardization, Geneva

ISO 8785 (1998) Geometrical Product Specifications (GPS) – surface imperfections – terms definitions and parameters. International Organization for Standardization, Geneva

ISO 11562 (1996) Geometrical product specifications (GPS) – surface texture: profile method – metrological characteristics of phase correct filters. International Organization for Standardization, Geneva

ISO 12085 (1996) Geometrical product specifications (GPS) – surface texture: profile method – motif parameters. International Organization for Standardization, Geneva

ISO 12179 (2000) Geometrical product specifications (GPS) – surface texture: profile method – calibration of contact (stylus) instruments. International Organization for Standardization, Geneva

ISO 13565-1 (1996) Geometrical product specifications (GPS) – surface texture: profile method – surfaces having stratified functional properties – part 1: filtering and general measurement conditions. International Organization for Standardization, Geneva

ISO 13565-2 (1996) Geometrical product specifications (GPS) – surface texture: profile method – surfaces having stratified functional properties – part 2: height characterization using the linear material ratio curve. International Organization for Standardization, Geneva

ISO 13565-3 (1998) Geometrical product specifications (GPS) – surface texture: profile method – surfaces having stratified functional properties – part 3: height characterization using the material probability curve. International Organization for Standardization, Geneva

ISO 25178-6:2010 Geometrical product specifications (GPS) – Surface texture: Areal – Part 6: Classification of methods for measuring surface texture, International Organization for Standardization, Geneva

ISO 25178-601:2010 Geometrical product specifications (GPS) – Surface texture: Areal – Part 601: Nominal characteristics of contact (stylus) instruments, International Organization for Standardization, Geneva

ISO 25178-602:2010 Geometrical product specifications (GPS) – Surface texture: Areal – Part 602: Nominal characteristics of non-contact (confocal chromatic probe) instruments, International Organization for Standardization, Geneva

ISO 25178-701:2010 Geometrical product specifications (GPS) – Surface texture: Areal – Part 701: Calibration and measurement standards for contact (stylus) instruments, International Organization for Standardization, Geneva

ISO/IEC Guide 98-3 (2008) Uncertainty of measurement – part 3: guide to the expression of uncertainty in measurement (GUM:1995), 1st edn. International Organization for Standardization, Geneva

ISO/IEC Guide 99:2007(E/F) (2007) International vocabulary of basic and general terms in Metrology (VIM). International Organization for Standardization, Geneva

Abbott EJ, Firestone FA (1933a) A new profilograph measures roughness of finely finished and ground surfaces. Autom Ind 204

Abbott EJ, Firestone FA (1933b) Specifying surface quality: a method based on accurate measurement and comparison. Mech Eng 55: 569–572

Agilent Technologies (2009) Agilent Technologies – home page. http://www.home.agilent.com. Accessed 30 Sep 2009

AIST-NT (2009) AIST-NT – home page. http://www.aist-nt.com. Accessed 30 Sep 2009

Ambios Technology (2009) Ambios Technology – home page. http://www.ambiostech.com/index.html. Accessed 30 Sep 2009

Aspex (2009) Aspex Corporation – home page. http://www.aspexcorp.com. Accessed 30 Sep 2009

Asylum Research (2009) Asylum Research – home page. http://www.asylumresearch.com.

Bhushan B, Wyant JC, Koliopoulos CL (1985) Measurement of surface topography of magnetic tapes by Mirau interferometry. Appl Opt 24(10):1489–1497

Binnig G, Rohrer H (1983) Scanning tunnelling microscopy. Surf Sci 126:236–244

Binnig G, Quate CF, Gerber C (1986) Atomic force microscope. Phys Rev Lett 56(9):930–933

Bruker AXS (2009) Bruker AXS – home page. http://www.bruker-axs.com. Accessed 30 Sep 2009

CamScan Electron Optics (2009) CamScan Electron Optics – home page. http://www.camscan.com. Accessed 30 Sep 2009

Carl Zeiss (2009) Carl Zeiss International – home page. http://www.zeiss.com/explore. Accessed 29 Sep 2009

Danilatos GD, Postle R (1982) The environmental scanning electron microscope and its applications. Scanning Electron Microsc 1:1–16

Dupuy MO (1967) High-precision optical profilometer for the study of micro-geometrical surface defects. In: Proceedings of the Institute of Mechanical Engineers, vol 182, part 3k, pp 255–259

FEI (2009) FEI – home page. http://www.fei.com. Accessed 30 Sep 2009

FOGALE nanotech (2009) FOGALE nanotech – home page. http://www.fogale.fr. Accessed 30 Sep 2009

Fries Research & Technology (2009) FRT – home page. http://http://www.frt-gmbh.com/en/

Greivenkamp JE, Bruning JH (1992) Phase shifting interferometry. Opt Shop Test 501–599

Hamilton DK, Wilson T (1982) 3D surface measurement using confocal scanning microscopes. Appl Phys B 27:211–213

Hansen HN (2007) Dimensional metrology in micro manufacturing. Whittles/CRC, Botovets, pp 29–35

Hansen HN, Carneiro K, Haitjema H et al (2006) Dimensional micro and nano metrology. CIRP Ann Manuf Technol 55(2):721–743

Harasaki A, Schmit J, Wyant JC (2000) Improved vertical-scanning interferometry. Appl Opt 39(13):2107–2115

Hariharan P (1985) Optical interferometry. Academic, New York

Hirox (2009) Hirox Co Ltd – home page. http://http://www.hirox.com/global_home.html. Accessed 29 Sep 2009

Hitachi High Technologies America (2009) Hitachi High Technology – home page. http://www.hitachi-hta.com. Accessed 30 Sep 2009

Hudson B (1973) The application of stereo-techniques to electron micrographs. J Microsc 98: 396-401

JEOL (2009) Jeol Ltd. – home page. http://www.jeol.com. Accessed 30 Sep 2009

KLA-Tencor (2009) KLA-Tencor – home page. http://www-kla-tencor.com. Accessed 30 Sep 2009

Kurfess T, Hodgson TJ (2007) Metrology, sensors and control. In: Ehmann KF, Bourell D, Culpepper ML et al (eds) Micromanufacturing – international research and development. Springer, Dordrecht

Leica Mikrosysteme Vertrieb (2009) Global home: Leica Microsystems. http://www.leica-microsystems.com. Accessed 24 Sep 2009

Lonardo PM, Trumpold H, De Chiffre L (1996) Progress in 3D surface microtopography characterization. Ann CIRP 42(2):589–598

Mahr Federal (2009) Home USA – Mahr metrology. http://www.mahr.com. Accessed 29 Sep 2009

Micro Photonics (2009) Micro Photonics Inc. – home page. http://http://www.microphotonics.com/. Accessed 30 Sep 2009

Minsky M (1961) Microscopy apparatus. US Patent 3,013,467, 19 Dec 1961

Mitutoyo (2009) Mitutoyo – home page. http://http://www.mitutoyo.co.jp/index.html. Accessed 30 Sep 2009

Molesini G, Pedrini G, Poggi P et al (1984) Focus-wavelength encoded optical profilometer. Opt Commun 49:229–233

Nanonics Imaging (2009) Nanonics Imaging – home page. http://www.nanonics.co.il. Accessed 30 Sep 2009

Nanosurf (2009) Nanosurf – home page. http://www.nanosurf.com. Accessed 30 Sep 2009

Nichols JF, Shilling M, Kurfess TR (2008) Review of MEMS metrology solutions. Int J Manuf Technol Manag 13(2–4):344–359

Novacam Technologies (2009) Novacam – home page. http://www.novacam.com. Accessed 30 Sep 2009

Novascan Technologies (2009) Novascan Technologies – home page. http://www.novascan.com. Accessed 30 Sep 2009

NT-MDT (2009) NT-MDT – home page. http://www.ntmdt.com. Accessed 30 Sep 2009

Olympus (2009) Olympus – home page. http://http://www.olympus-global.com/en/global/. Accessed 30 Sep 2009

Optimet Optical Metrology (2009) Optimet – home page. http://www.optimet.co.il. Accessed 3 Oct 2009

Park Systems (2009) Park Systems – home page. http://www.parkafm.com. Accessed 30 Sep 2009

Sensofar Tech (2009) Sensofar Tech – home page. http://www.sensofar.com. Accessed 29 Sep 2009

Sirat G, Paz F (1998) Conoscopic probes are set to transform industrial metrology. Sens Rev 18(2):108–110

Solarius Development (2009) Solarius – home page. http://www.solarius-inc.com/index.html. Accessed 29 Sep 2009

Taylor Hobson (2009) Taylor Hobson – home. http://www.taylor-hobson.com. Accessed 30 Sep 2009

Thornton PR (1968) Scanning electron microscopy. Chapman and Hall, London

Veeco Instruments (2009) Veeco – solutions for a nanoscale world. http://www.veeco.com

VisiTec Microtechnik (2009) VisiTec – home page. http://http://www.visitec-em.de/. Accessed 30 Sep 2009

Weckenmann A, Estler T, Peggs G et al (2004) Probing systems in dimensional metrology. CIRP Ann Manuf Technol 53(2):657–684

Werth Messtechnik (2009) Werth Messtechnik – home page. http://http://www.werth.de. Accessed 17 Dec 2009

Zygo (2009) Zygo Corporation – home page. http://www.zygo.com. Accessed 28 Sep 2009

Standard under Development

ISO/CD 25178-1 Geometrical product specifications (GPS) – surface texture: areal – part 1: indication of surface texture. International Organization for Standardization, Geneva

ISO/CD 25178-604 Geometrical product specifications (GPS) – surface texture: areal – part 604: nominal characteristics of non-contact (coherence scanning interferometry) instruments. International Organization for Standardization, Geneva

ISO/CD 25178-605 Geometrical product specifications (GPS) – surface texture: areal – part 605: nominal characteristics of non-contact (point autofocusing) instruments. International Organization for Standardization, Geneva

ISO/DIS 25178-2 Geometrical product specifications (GPS) – surface texture: areal – part 2: terms, definitions and surface texture parameters. International Organization for Standardization, Geneva

ISO/DIS 25178-3 Geometrical product specifications (GPS) – surface texture: areal – part 3: specification operators. International Organization for Standardization, Geneva

ISO/DIS 25178-7 Geometrical product specifications (GPS) – surface texture: areal – part 7: software measurement standards. International Organization for Standardization, Geneva

ISO/DIS 25178-603 Geometrical product specifications (GPS) – surface texture: areal – part 603: nominal characteristics of non-contact (phase-shifting interferometric microscopy) instruments. International Organization for Standardization, Geneva

Chapter 4
Coordinate Measuring Machine Measurement Planning

Giovanni Moroni and Stefano Petrò

Abstract Once a measuring instrument has been chosen for the control of the quality of a part of known design and specifications, the measurement process must be planned. For coordinate measuring instruments implementing point-based measurement, planning implies appropriately choosing the number and placement (pattern) of the points to be measured. In fact, coordinate measuring instruments sample points on features to be measured, but how should points be located on the feature itself? This problem is particularly relevant with measuring instruments which require a long time to sample dense clouds of points, *e.g.*, most coordinate measuring machines (CMMs). In this chapter the problem of planning the inspection strategy, *i.e.*, defining the number and pattern of sampling points, provided the measuring systems allow the operator to define the inspection strategy, will be addressed, with particular reference to CMMs. Sample size planning will be approached as an economic problem, because as the sample size increases, uncertainty is reduced and measurement cost rises, and a trade-off has to be searched for. Then, a few different criteria for defining the sampling pattern are proposed; these differ in terms of the accuracy and the information required for their application. These criteria can be categorized as blind, adaptive, and process-based sampling strategies. A few examples are proposed, outlining the effectiveness of different approaches to sampling strategy planning. In order to better understand the problem of strategy planning, a brief description of the main CMM features is provided.

G. Moroni
Department of Mechanical Engineering, Politecnico di Milano,
Via La Masa 1, 20156, Milan, Italy,
e-mail: giovanni.moroni@mecc.polimi.it

S. Petrò
Department of Mechanical Engineering, Politecnico di Milano,
Via La Masa 1, 20156, Milan, Italy,
e-mail: stefano.petro@polimi.it

4.1 Introduction

According to the ISO 10360-1 (2000) standard, a “coordinate measuring machine” (CMM) is a “measuring system with the means to move a probing system and capability to determine spatial coordinates on a workpiece surface”. Therefore, a CMM is a (most of time the Cartesian) robot able to move in a 3D space a generic sensor which can probe the coordinate of a cloud of points belonging to the workpiece under measurement. As in any coordinate measuring system, the cloud of points is then represented (ISO 14660-1 1999) by a nominal model of the geometry of the measured part. Finally, the geometric error or the dimension is calculated from the associated geometry.

4.1.1 What Is a CMM?

The CMM concept may be associated with a very large family of measuring systems, so there will be CMMs with contacting or noncontacting probes, CMMs with bridge structures, CMMs with a horizontal arm, manual or motorized CMMs, and small or large CMMs. Anyway, regardless of the implementation of the specific CMM concept, CMMs are flexible. Their ability to sample points in a 3D space allows them to check any geometric or dimensional tolerance. Of course, different CMMs are more or less suitable for different measurement tasks, but broadly speaking any CMM can measure any part subject to geometric specifications, provided the CMM measuring volume is adequate (the part is not too large).

A CMM is essentially constituted by the following parts (Figure 4.1) (Bosch 1995):

- mechanical setup (machine axes and transducers);
- sensor;
- control unit; and
- computer with software for data processing.

4.1.1.1 Mechanical Setup

The CMM configuration describes the overall structure of the machine. Even though the commercial solutions are very different, they can be reduced to five basic configurations: bridge, cantilever, horizontal arm, gantry, and non-Cartesian. ISO 10360-1 (2000) further details these structures, distinguishing machines with a fixed or moving table, column-type CMM, *etc.*, which are variations of the previously mentioned configurations.

Differing measurement tasks require different characteristics from CMMs. Therefore, different CMM configurations have been proposed, each one with its advantages and disadvantages.

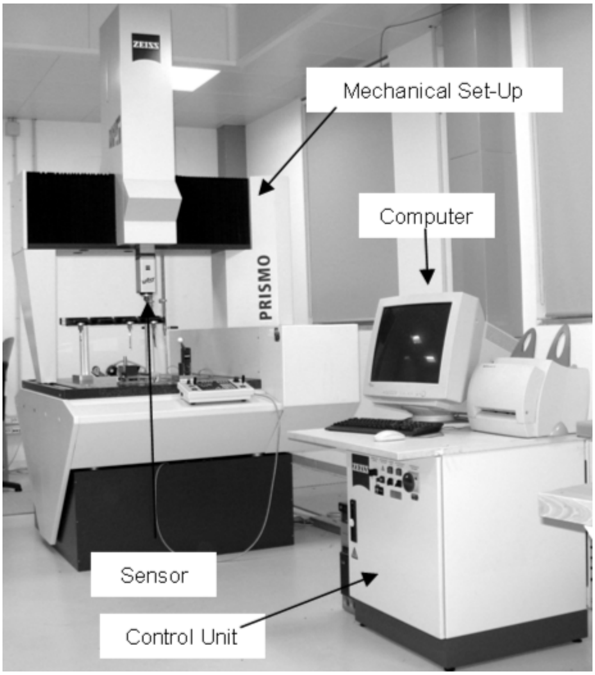


Figure 4.1 Coordinate measuring machine components (Geometrical Metrology Laboratory in the Department of Mechanical Engineering at the Politecnico di Milano)

A cantilever CMM allows for good accessibility of the measuring volume, but the cantilever arm reduces the structure stiffness, thus reducing the overall accuracy. This may be avoided by adopting a bridge structure, which improves stiffness, even though accessibility is reduced. Gantry and horizontal arm CMMs are very effective for large volume measurement, owing to their inherent structure characteristics. However, because in these kinds of CMMs the fixture supporting the part and the structure moving the measuring sensor are often not directly connected, accuracy is reduced with respect to other solutions. Finally non-Cartesian CMMs are usually less accurate than any other configuration. However, because of their ability to move "around" the part, this kind of CMM may be really flexible from the accessibility point of view and really suitable to sample large clouds of points in a short time, as is required for reverse modeling purposes, when coupled to a sensor like a laser stripe.

4.1.1.2 CMM Sensors

A CMM sensor, or probe, is that part of the machine which "touches" (physically, or optically, or with any other principle) the part to be inspected. Most of the CMM accuracy relies on sensor accuracy, which strongly influences the overall

CMM performance. Therefore, manufacturers have proposed a great variety of CMM sensors, each one with its particular advantages.

Three main categories of CMM sensors can be identified: contact sensors, noncontact sensors, and hybrid sensors.

A contact sensor (Weckenmann *et al.* 2004) is a sensor which physically touches the part to be inspected. Contact sensors represent the most diffused CMM sensors, because of their higher accuracy, the availability of specific international standards, and their adoption for a long time. However, because of the interaction between the probe and the measured part, contact sensors are not suitable for measuring soft parts. Contact sensors can usually perform discrete-point or scanning measurements, or both, depending on their specific characteristics.

Conversely, noncontact sensors (Schwenke *et al.* 2002; Savio *et al.* 2007) do not touch the part to be inspected. A wide variety of measuring principles have been proposed for noncontact probes, including laser triangulation, various focalization technologies, confocal holography, and vision systems. Currently, noncontact probes suffer from more uncertainty sources than contact probes, and this makes their adoption harder. Moreover, they lack the ability to measure internal features and undercuts, *e.g.*, they cannot measure holes. Finally, standardization for noncontact sensors is still insufficient. However, their higher measuring speed and the ability to efficiently measure some characteristics which are impossible to measure with contact probes, such as soft parts and printed circuits, are making noncontact probes more and more diffused.

Hybrid sensors include both a contact and a noncontact sensor. The two sensors can sometimes work independently: in this situation, the noncontact sensor is adopted to roughly identify the location of the feature to be inspected, *e.g.*, a hole, and then the feature is accurately measured by the contact sensor. Otherwise, the two sensors may work only coupled, like in the "fiber probe" (Schwenke *et al.* 2001). CMMs equipped with hybrid sensors are further discussed in another chapter.

4.1.1.3 Control Unit

The control unit of a CMM coordinates the various parts of the CMM itself. Different kinds of controls are possible. The simplest control is manually driven, both with or without axis motorization. However, most of the present CMMs are controlled by a computerized numerical control (CNC) system. The CNC system may control the machine in discrete-point probing, or scanning (ISO 10360-1 2000). In discrete-point probing, after probing a point, the CMM probe leaves the surface and moves toward the next point, and so on. In scanning, the probe is in continuous contact with the surface, so a line is sampled. Scanning may be performed both on a predefined path or on an unknown path. In this last situation, the control system has to able to adapt the probing path in order to move from a point on the surface to be inspected to another point without losing contact with the surface itself.

The control unit is also responsible for software compensation (Sartori and Zhang 1995; Schwenke *et al.* 2008). In recent years, CMM manufacturers have pointed out that further increasing CMM mechanical accuracy is not economic. The solution has been to "map" the measurement errors in the CMM measuring volume, based on repeated measurement of specifically designed, calibrated or uncalibrated, artifacts. Then, because the errors due to mechanical inaccuracy are known, they may be compensated in successive measurement tasks, thus reducing measurement uncertainty.

4.1.1.4 Computer and Software for Data Processing

The output of the CMM itself is just a cloud of points. In order to evaluate dimensions and geometric errors, the cloud of points has to be further analyzed. This analysis is performed by specific software. The principal function of the software is then to fit the clouds of points in order to extrapolate measurements.

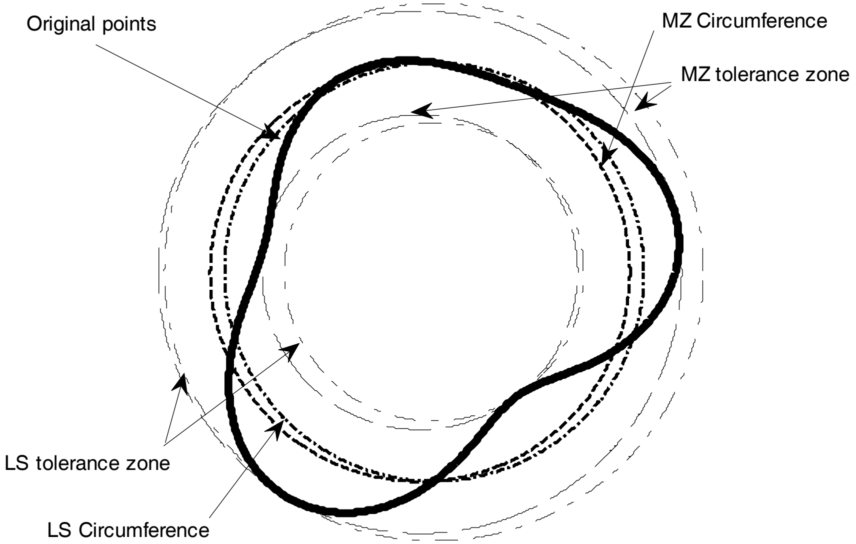


Figure 4.2 Least-squares (*LS*) and minimum-zone (*MZ*) fitting for a circle

Two fitting principles (Anthony *et al.* 1996) are mainly used for analyzing clouds of points: least-squares (also known as Gaussian) and minimum-zone (also known as Chebyshev) fitting (Figure 4.2).

Least-squares fitting consists in solving the problem

$$\min_{\mathbf{p}} \sum_{i=1}^{n} d_i^2(\mathbf{p}), \tag{4.1}$$

where **p** is a vector of parameters defining the fitting geometry, $d_i(\mathbf{p})$ is the *signed* distance of the *i*th point from the fitting geometry, and is, of course, a function of **p**, and n is the number of sampling points.

Minimum-zone fitting is defined by the solution of the following minimum–maximum problem:

$$\min_{\mathbf{p}} \max_{i} \left| d_i \left(\mathbf{p} \right) \right| . \qquad (4.2)$$

Least-squares fitting is considered to be more robust to the presence of anomalous points characterized by a large measurement error, and is therefore preferred for fitting when dimensional characteristics have to be inspected. Minimum-zone fitting best interprets the definition of "tolerance zone" given in the ISO 1101 (2004) standard for geometric tolerances, and therefore is preferred for geometric error evaluation.

4.1.2 Traceability of CMMs

In order to ensure metrological traceability of measurements, measurement uncertainty evaluation is required. However, when dealing with CMMs, experience has shown that uncertainty is strongly affected by the specific measurement task considered (Wilhelm *et al.* 2001). This makes the definition of CMM performance in terms of uncertainty difficult. Moreover, CMM calibration is not sufficient to define uncertainty. Therefore, the problem of CMM traceability consists in performance verification, which is addressed by the ISO 10360 standards series, and uncertainty evaluation, addressed by the ISO 15530 standards series.

4.1.2.1 Performance Verification

Performance verification consists in a series of tests which, if passed, ensure the CMM currently works in its nominal performance condition. In order to ensure international validity, tests are described in ISO 10360-1 (2000), ISO 10360-2 (2001), ISO 10360-3 (2000), ISO 10360-4 (2000), ISO 10360-5 (2000), and ISO/DIS 10360-7 (2008). Performance tests are based on the measurement of reference artifacts, and then comparison of measurement results and calibrated values. Among the others, two kinds of performance are usually considered most relevant: size measurement performance, and probing performance.

Size measurement performance is defined by the "maximum permissible error of indication of a CMM for size measurement" (MPE_E) (ISO 10360-1 2000). This is the maximum measurement error allowed when measuring a specific length standard. MPE_E is usually defined in a form such as $\mathrm{MPE}_E = \pm\, (A + L/K)$, where L is the length of the measured length standard, and A and K are constants specific for the CMM considered. Currently, the ISO 10360-2 (2001) standard allows gauge blocks and step gauges to be adopted as length standards, but the standard is under revision, in order to allow the use of instruments such as laser interferometers, which are required for large-volume CMMs. Three repetitions of a set of five length standards have to be performed, in seven different positions within the measuring volume of the CMM.

Probing performance is defined by MPE_P, the "extreme value of the probing error *P* permitted by specifications, regulations, *etc.*", where *P* is the "error of indication within which the range of radii of a spherical material standard of size can be determined by a CMM" (ISO 10360-1 2000). Its test is performed by measuring a reference (calibrated) sphere according to a pattern of 25 points specified by the ISO 10360-2 (2001) standard. Similar performance indicators have been proposed for scanning CMMs (ISO 10360-4 2000) and multistylus or articulating probe CMMs (ISO 10360-5 2000).

Currently, these procedures are suitable only for CMMs equipped with contacting probes. The ISO 10360-7 standard is currently under development for CMMs with imaging probes, and at present it exists only as a draft (ISO/DIS 10360-7). This lack of a reference standard makes noncontact CMMs not completely traceable.

4.1.2.2 Uncertainty Evaluation

There exist a large number of uncertainty sources in CMM measurement, so the problem of uncertainty evaluation is quite complex. Uncertainty is therefore considered to be specific for every measurement task (Wilhelm *et al.* 2001). Usually, uncertainty sources in CMMs are:

- hardware uncertainty sources, that is, anything related to the structure (sensor, mechanical structure, *etc.*) of the CMM;
- workpiece uncertainty sources, related to properties of the workpiece and measurement interaction with the workpiece;
- sampling strategy, including inadequate sampling or datums, and interaction between the sampling strategy and the actual geometric error;
- fitting and evaluation algorithms; and
- extrinsic factors, such as temperature, operator, and dirt.

This variety of uncertainty sources has led to a few uncertainty evaluation procedures, summarized in the ISO 15530 series of standards. Currently, only parts three (ISO/TS 15530-3 2004) and four (ISO/TS 15530-4 2008) have been published.

ISO/TS 15530-3 (2004) proposes a way of evaluating uncertainty which is based on repeated measurement of one or more calibrated artifacts in the same conditions in which the following measurements will be performed. Even though the procedure is quite simple, for each uncertainty evaluation a calibrated artifact, which has to be as similar as possible to the real parts, is required. Therefore, this procedure is suitable only for large-scale production, and it is too expensive in other contexts.

To avoid this limitation, ISO/TS 15530-4 (2008) proposes simulation to evaluate uncertainty. In this procedure, measurements are only simulated, so it is applicable to any measurement task without requiring a specific artifact. However, the standard does not define the simulation procedure: it just defines simulation validation. Therefore, the development of simulation software for CMMs is demanded of CMM manufacturers and software developers.

Currently, part two of ISO 15530 is under development (ISO/CD TS 15530-2). This standard is aimed at developing a strategy for artifact calibration, and should be based on multiple measurement strategies in order to eliminate, or at least reduce, measurement bias. In fact, if a single strategy is adopted, the measurement result could be biased because of bad interaction between the form error and the strategy, or local errors not taken into account by the software compensation.

4.1.3 CMM Inspection Planning

After a CMM has been selected to perform a specific measurement task, the inspection has to be planned. Inspection planning consists in defining the single aspects of the measurement:

- fixturing definition;
- sensor configuration;
- sampling strategy definition; and
- path planning.

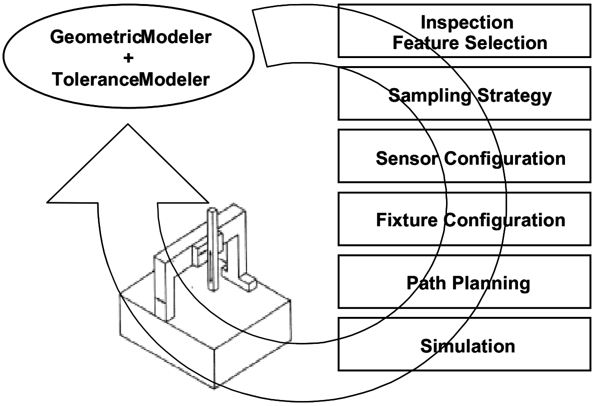


Figure 4.3 An automatic inspection planning system for a coordinate measuring machine

CMM inspection planning is an activity performed by well-trained operators, but different measurement techniques, using the same data analysis algorithms, yield different measurement results. This is a well-recognized source of uncertainty in coordinate measurement. A CMM equipped with an automatic inspection planning (CAIP) system (Figure 4.3), permits one to implement more accurate and efficient operating procedures, and to employ higher quality assurance standards and tighter production timings.

A CAIP system should be able to perform the following activities:

- Select the features to be inspected (Fiorentini *et al.* 1992); in fact, the number of features to be inspected in a single part is often very large. This could lead to a long measurement time. Therefore, methods are required to choose which features have to be inspected.
- Configure the sensor and the fixture (Moroni *et al.* 1998), that is, define how the fixture and the sensor should be designed in order to minimize part deformation, ensure surface accessibility, and minimize risk of collision between the probe and the part.
- Plan the probe path (Yau and Menq 1995), avoiding collision between the part/fixture and the probe.
- Plan the sampling, or measurement, strategy. Strategy planning consists in defining how many sampling points should be taken, and in which locations.

Because this last factor is responsible for most of the measurement uncertainty, the rest of this chapter will illustrate methods of measurement strategy planning.

4.2 Measurement Strategy Planning

Under the assumption that a coordinate measuring system has been set up, *i.e.*, a suitable fixture for the part has been defined, allowing for easy access to every feature to be inspected, the probe is correctly configured and qualified, and so on, the next problem to solve is the choice of where to sample the part feature in order to obtain a description of the surface complete enough to estimate substitute features with sufficient accuracy for the measurement task, but at the same time without exceeding the sample size, because the measurement time and cost tend to increase as the sample size increases; this is in part true even for standstill scanners. In fact, if the lateral resolution of the scanner is not adequate in the current setup, the operator will have to take multiple scans of the part and then register them – of course, more scans means a longer measuring time. This problem is known as "defining a sampling strategy", where the sampling strategy is the complete description of locations of the sampling points.

When dealing with geometric error estimates, the sampling strategy is one of the most relevant contributors to measurement uncertainty. Oversimplifying, it could be stated that, according to definitions in ISO 1101 (2004), geometric error is defined by the point that deviates the most from the nominal geometry; however, in a feature sampling process the actual measurement is not a complete description of the feature itself, it is just an approximation. Therefore, a sampling strategy is effective if it is able to catch the most deviating point of the feature most of the times. In fact, according to the actual definitions of tolerance zones, only these points influence compliance or noncompliance of parts, that is, if these points fall inside the tolerance zone, then the part is compliant. Of course, if the most deviating point is not caught by the sampling, some measurement error will be present. Figure 4.4 exemplifies how this error originates. First of all, it may be interesting to point out that this measurement error will be present even if the sampling process does not generate any measurement error, that is, *e.g.*, for a CMM equipped with a touching probe, when the measured coordinates of the contacting point coincide with the coordinates of the real point for any sampling point. In this case, the measurement error will consist of an underestimate of the geometric error (see Figure 4.4a, which shows what happens if a nominally circular, truly elliptical profile is sampled with an insufficient density of points), and the underestimate is inversely related to the sample size. However, because any real sampling process generates some sampling error, it is possible to overestimate the geometric error. Consider, for example, the sampling of a perfect geometry, *i.e.*, a manufactured geometry which is identical to its nominal counterpart. Of course, the geometric error should be null, but because of probing error, the sampling points will probably not belong to the perfect geometry, so the geometric error estimate will be greater than zero, and will tend to increase as the sample size increases, because measurement error dispersion will tend to form a sort of envelope around the real geometry. Similarly, if a nonperfect feature is inspected, the sampling error tends to inflate the geometric error estimate (Figure 4.4d). Unfortunately, in most practical application scenarios involving coordinate measuring systems the underestimation due to undersampling is more relevant than the overestimation due to probing error, so geometric error estimates usually include some residual systematic measurement error.

Therefore, in order to obtain geometric error estimates as accurate as possible, a correct sampling strategy has to be defined, able to identify those areas of the surface which have the maximum local geometric deviation, that is, the maximum deviation of the real feature from the nominal feature (ISO 14660-1 1999). This can be obtained by sampling a large number of points uniformly distributed across the feature to be inspected. If the measuring system adopted is fast (that is, it is able to sample a large number of points in a short time), this is probably the best way to sample, because there is no risk of missing some relevant areas of the feature itself. However, measuring instruments characterized by the minimum probing error are often slow, so the problem of choosing the correct sampling strategy may become relevant, because a complete description of the feature would require an unaffordable measurement time.

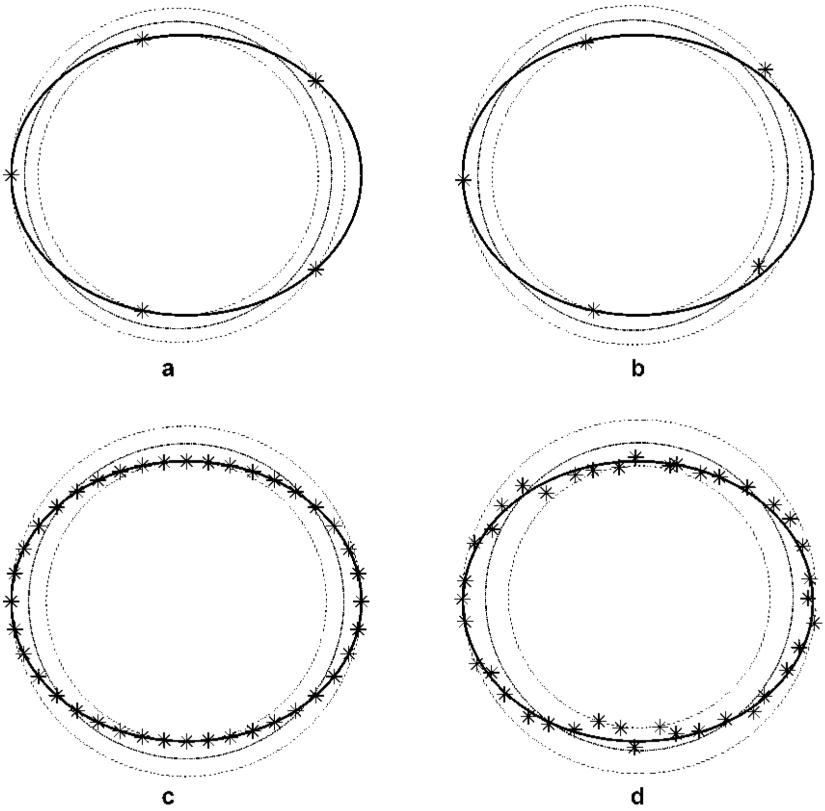


Figure 4.4 Effects of measurement error and sample size on the estimated geometric error (the real geometric error is 2): **a** no probing error, five points, estimated geometric error 1.523, **b** with probing error, five points, estimated geometric error 1.6396, **c** no probing errors, 40 points, estimated geometric error 2, and **d** with probing error, 40 points, estimated geometric error 2.6303

Moreover, it should be pointed out that often with typical coordinate measuring systems the definition of the sampling strategy is not completely free. Consider, for example, measurement by means of a structured light scanner. The number and the pattern of the sampling points will depend on the relative position of the structured light projector, camera(s), and the object to be measured, and the sampling points will be measured more or less simultaneously. If more sampling points were to be measured, one would have to modify this setup, but this may imply – among other things – an increase of probing error or difficulties in registering clouds of points. Measurements could be repeated within the same setup to average the results. Averaging will lead to a reduction of probing error, and improve accuracy; but if the overall coverage of the sampling strategy does not change, that is, points are located in the same positions in every measurement repetition, it is still possible to miss critical areas. Therefore, the sampling strategy may tend to the situation depicted in Figure 4.4a: some very accurate points, missing the areas of maximum geometric deviation, thus introducing a measurement bias (*a priori* impossible to evaluate). However, when adopting these measuring systems, the sample size obtained in a single acquisition is often enough to describe the whole surface with a sufficient point density. Differently, if a profiler or a scanning CMM (Weckenmann *et al.* 2004) is adopted as the measuring instrument, the sampling points will be grouped in profiles, even when the inspected feature is a surface. The measurement of a single point will be quite fast, but covering the whole surface with the same density of points along profiles and in any other direction may be very time-consuming. Finally, if a CMM is adopted which is capable of point-to-point measuring, significant freedom is granted to define the sampling strategy; unfortunately, these measuring systems are usually the slowest. Therefore, the problem of choosing the correct sampling strategy is most critical in this situation.

Summarizing, with fast surface scanners there should be no significant problems in planning the sampling strategy, and measurement uncertainty is mainly influenced by probing error; with profile scanners or CMMs equipped with scanning probing systems, even though it is possible to sample a large number of points, the number and the placement of profiles to be acquired must be carefully selected, in order to ensure a good coverage of the surface; finally, with point-to-point measuring systems, densely measuring the whole feature will often require an unaffordable measuring time. Because the last problem is related to CMMs and is the most critical situation for sampling strategy planning, it will be considered as the reference situation in the chapter.

The problem of choosing the correct sampling strategy may be split into two subproblems: sample size choice and sampling pattern choice.

The first problem is economic: measurement accuracy and measurement cost are, most of the time, directly related to sample size, so a trade-off between accuracy and cost has to be identified. In the last part of this chapter the problem of economically evaluating accuracy will be addressed, and an economic model for inspecting cost will be proposed that, when optimized, defines the optimal sample size.

The choice of the sampling pattern is less straightforward. The effectiveness of a sampling strategy has been defined as its ability in identifying the most deviating point. Provided one is completely free to choose the sampling strategy, in most situations the only information available about the measurement to be made pertains to nominal geometry, tolerances, and fixtures. A sampling strategy based on this information alone is defined as a “blind sampling strategy”, because it does not consider any information on the actual geometric error of the feature. However, it is possible to develop a sampling strategy based on the actual geometric deviation of the part. If the sampling strategy is based on information on the geometric deviation acquired during measurement, that is, a set of points is sampled, then depending on such points, one or more additional points are sampled according to some criterion, and so on, until some terminating condition is reached, the strategy is an “adaptive sampling strategy”, which tries to “adapt” to the actual geometric error of the inspected part. Finally, if the information available on the geometric deviation relies on a preliminary study of the manufacturing process, and the measurement strategy is planned on the basis of this information, then this is a “process-based” sampling strategy. Consistency of process-based strategies is guaranteed by the fact that geometric deviations generated by a particular manufacturing process tend to be similar in each manufactured part.

Because the choice of the sample size is strongly affected by the criterion chosen for planning the pattern, blind, adaptive, and process-based strategies will be introduced before addressing the problem of choosing the sample size.

4.3 Sampling Patterns

Several kinds of sampling pattern definitions have been proposed in the literature. They may be categorized according to the following classification:

1. blind sampling strategies;
2. adaptive sampling strategies; and
3. process-based sampling strategies.

4.3.1 Blind Sampling Strategies

A sampling strategy is blind if the only information required to define it is the nominal geometry and tolerances of the part to be inspected, the metrological characteristics of the measuring system adopted for the inspection, and the number of sampling points is chosen once and applied to all parts to be inspected. The sampling pattern is defined *before* starting the measurement, and does not change from part to part.

Because blind sampling strategies consider only the nominal geometry of a part, most of them tend to spread sampling points with a constant or nearly constant density throughout the surface. Therefore, the most important parameter that must be dealt with to define a blind sampling strategy is the density of sampling points, which is directly linked to the sample size. In particular, a blind sampling strategy is deemed accurate if the sample size is large enough to sample every part of the surface with an adequate sampling point density.

The uniform coverage of the surface guarantees robustness of the sampling strategy, that is, provided the sample size is adequate, it is unlikely that critical areas of the surface will be missed; regardless of the actual shape of the geometric error, measurement accuracy is ensured. Therefore, if it is possible to adopt an adequate sample size (*i.e.*, not excessively expensive or time-consuming), blind sampling strategies may be considered an adequate choice because the whole profile/surface is sampled, thus ensuring robustness.

Moreover, sampling patterns for blind sampling strategies are easy to define. Often, blind sampling strategies are completely defined by the sample size only, that is, given the sample size, the sampling point pattern is completely defined, or at most depends only on random parameters (*e.g.*, if the distribution of sampling points is random, then having chosen the sample size to be *n*, there will be *n* randomly spread sampling points throughout the surface). Blind sampling strategies are therefore easy to implement in CMM control systems, and often strategies based on such a pattern, like a uniform strategy, may be automatically generated by CMM software.

The most common blind sampling strategies, including those proposed in the international standards, are now briefly illustrated.

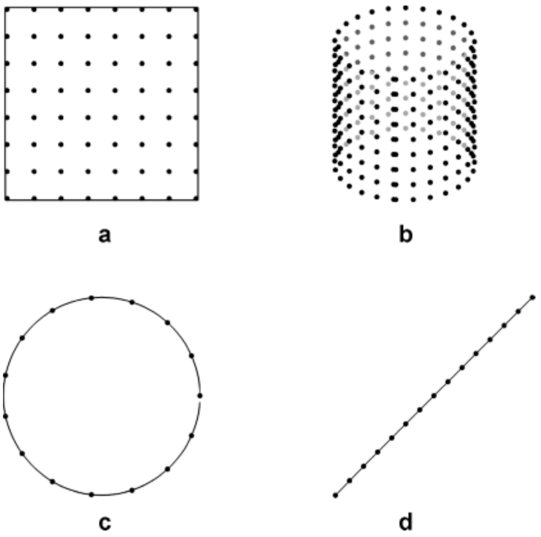


Figure 4.5 Uniform sampling strategies for **a** a plane, **b** a cylinder, **c** a circumference, and **d** a straight profile

A first way to distribute sampling points is to place them evenly spaced on a grid, for surfaces, or along the profile to be inspected (Figure 4.5): because the points are uniformly spaced, this is a *uniform strategy* (Dowling *et al.* 1997). A uniform strategy is strictly linked to Nyquist's theorem (Bracewell 2000). Any geometric feature can be described as the sum of sine waves. Provided the number of waves (harmonic content) required to completely describe a geometry is finite, Nyquist's theorem states that if a feature is sampled by a uniform strategy with a frequency more than double the maximum frequency of the sinusoids, the sampling points contain all the information concerning the feature, *i.e.*, the feature can be reconstructed exactly based on the sample. In contrast, if the sample size is not sufficient, the feature geometry cannot be completely reconstructed, and becomes indistinguishable from another feature which in turn is completely described by the same sample (a phenomenon known as “aliasing”, see Figure 4.6). Real features are seldom entirely defined by a finite harmonic content; however, in most situations, the first few harmonics of higher amplitude will be sufficient to reconstruct the geometry to a sufficient degree of approximation. If a threshold has been defined beyond which the amplitude of the harmonics is negligible, this threshold can be considered for application of Nyquist’s criterion when choosing the density of the sampling points.

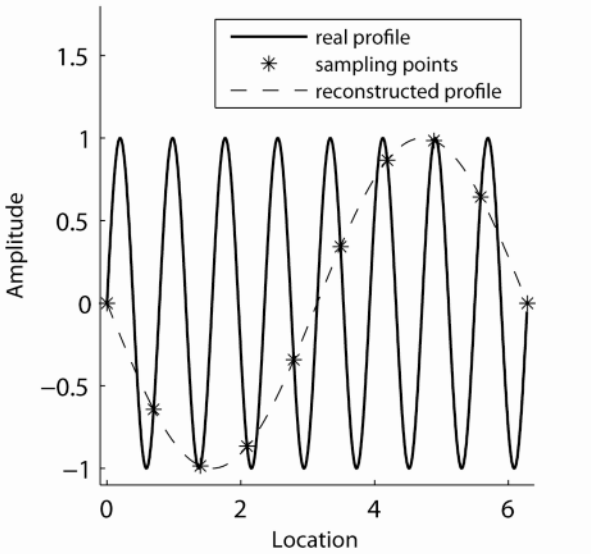


Figure 4.6 Effects of undersampling (real frequency 16, sampling frequency 10)

However, when dealing with surfaces, applying a uniform strategy which respects Nyquist’s criterion (even after having chosen a proper threshold) tends to lead to huge sample sizes; therefore, a uniform sampling strategy may be affordable only for fast measuring systems. On the other hand, the use of a uniform strategy that does not respect Nyquist’s criterion may again lead to aliasing-related problems. For example, if the surface is characterized by a dominant wave and the sampling point frequency is not adequate, sampling points may end up being located on wave crests, without any point in valleys, thus resulting in an incorrect assessment of the underlying geometry.

If a uniform strategy respecting Nyquist’s criterion is unaffordable owing to the excessive sample size required, a few strategies have been proposed that are capable of avoiding problems related to the adoption of a uniform grid of points. One is the *random strategy* (Dowling *et al.* 1997). In a random strategy a given number of sampling points are randomly scattered throughout the feature to be inspected. This avoids the generation of any pattern, so no bad interaction should be created between the geometry of the real feature (ISO 14660-1 1999) and the sampling strategy; this reduction of systematic error is compensated by an increase of random measurement error, because if sample size is not so large, a completely random strategy may leave some areas of the feature scarcely covered, whereas useless concentration of sampling points is found in other areas (Figure 4.7). To avoid inhomogeneous concentration of sampling points, a *stratified strategy* may be chosen. In a stratified strategy the feature is split in some (usually equal) areas, and then a portion of the sampling points (proportional to the extent of the related area) is randomly distributed in each area (Figure 4.7). Because of the complete randomness of the pattern, random and stratified sampling strategies are slightly harder to implement with respect to the uniform strategy.

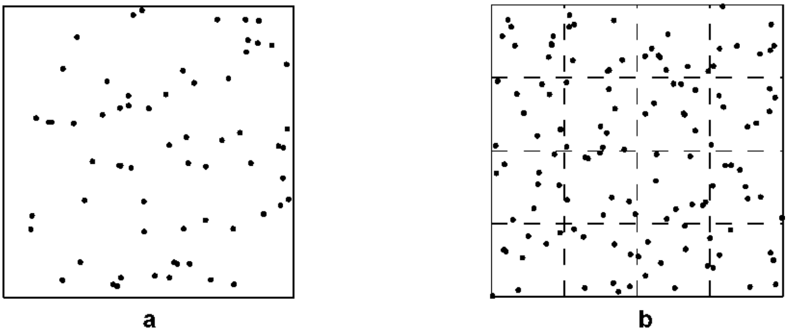


Figure 4.7 **a** Random sampling strategy, and **b** stratified sampling strategy

Some strategies which avoid defects of regular patterns without placing points in random locations come from the Monte Carlo method and are based on the so-called quasi-random sequences (Hammersley and Handscomb 1964). Probably, the most famous strategy of this kind (for surfaces) is the *Hammersley strategy*. The Hammersley sequence sampling strategy is mathematically defined as follows. If n points have to be sampled, then define

$$u_i = \frac{i}{n} \quad v_i = \sum_{j=0}^{k-1} b_{ij} 2^{-j-1}, \qquad i \in \{0,1,\ldots,n-1\}, \tag{4.3}$$

where u_i and v_i are normalized coordinates of the ith point (*i.e.*, they are defined in the [0,1] interval and must be rescaled to retrieve the actual surface coordinates), b_i is the binary representation of i, and b_{ij} is the jth bit of b_i (so $b_{ij} \in \{0,1\}$), and $\mathrm{k} = \lceil \log_2 n \rceil$, where $\lceil x \rceil$ is the smallest integer greater than or equal to x. As highlighted in Equation 4.3, once the sample size n is defined, the Hammersley strategy is univocally defined. However, as Figure 4.8 shows, points are spread throughout the whole surface without any apparent pattern, and this should avoid any interaction with any harmonic content in the surface, or any other systematic behavior. Lee *et al.* (1997) have shown that, given the sample size, the Hammersley sequence based sampling strategies outperforms both the uniform and the random sampling strategies for a plane, cylinder, cone, and dome (sphere). The authors claim that the Hammersley sequence has a nearly quadratic reduction in the number of points needed by a uniform sampling for the same level of accuracy.

Similar to the Hammersley sampling strategy is the *Halton–Zeremba strategy* (Figure 4.8), which is defined only if the sample size is a power of 2. The mathematical definition is as follows:

$$u_i = \sum_{j=0}^{k-1} b_{ij} 2^{-k-j}, \qquad v_i = \sum_{j=0}^{k-1} b'_{ij} 2^{-j-1}, \qquad i \in \{0,1,\ldots,n-1\}, \tag{4.4}$$

where b'_{ij} is equal to 1-b_{ij} if j is odd, otherwise b'_{ij} is equal to b_{ij} if j is even. Kim and Raman (2000) claim the Halton–Zeremba strategy not only outperforms the uniform and random strategies, but also outperforms the Hammersley strategy for flatness.

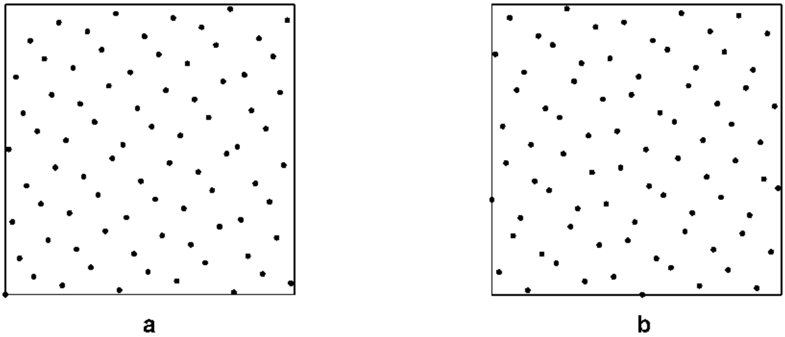


Figure 4.8 **a** Hammersley sampling strategy, and **b** Halton–Zeremba sampling strategy

Finally, some ISO technical specifications, *i.e.*, ISO/TS 12180-2 (2003), ISO/TS 12181-2 (2003), ISO/TS 12780-2 (2003), and ISO/TS 12781-2 (2003), deal with the problem of sampling respecting Nyquist's criterion when form tolerances are inspected (roundness, straightness, flatness, cylindricity); currently no suggestion is given concerning other geometric tolerances. Standards require that a Gaussian filter is always adopted to eliminate the influence of roughness. Then, in particular, if the toleranced feature is a profile (roundness or straightness), the standards suggest that a uniform strategy is adopted that samples *at least* with a frequency 7 times higher than the frequency of the filter adopted to eliminate the contribution of high-frequency components of geometric error (roughness). This choice ensures that only components with a wavelength *less than 0.02%* of the filter cutoff frequency can be significantly affected by aliasing.

When dealing with surfaces, the standards suggest adopting *profile extraction strategies*. In a profile extraction strategy, points do not uniformly cover the surface, but are grouped in profiles (usually scattered on the surface according to some pattern). Profile strategies are suitable for measuring instruments which naturally sample profiles, *e.g.*, profilers or CMMs equipped with scanning technologies. Profile strategies may be adopted even with systems which measure single points, simply by grouping single sampling points along profiles. Several patterns for profiles are proposed in standards for the plane and the cylinder, and are illustrated in Figures 4.9 and 4.10.

In order to choose how many points have to be sampled in each profile, the same criterion proposed for roundness and straightness is adopted (*i.e.*, a uniform

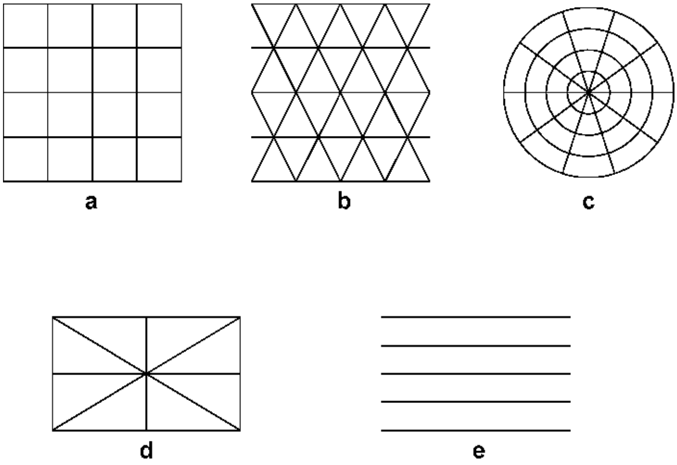


Figure 4.9 Profile extraction strategies proposed in ISO technical specifications for flatness: **a** rectangular grid, **b** triangular grid, **c** polar grid, **d** Union Jack, and **e** parallel profiles

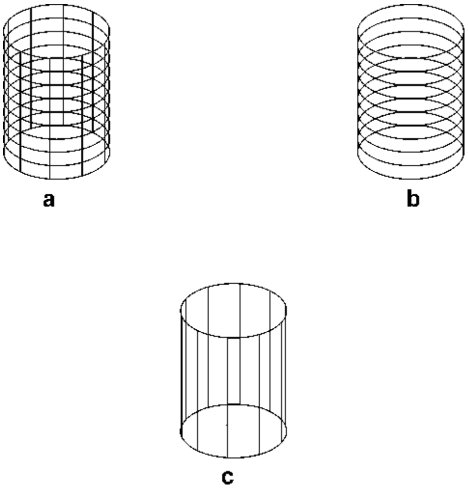


Figure 4.10 Profile extraction strategies in ISO technical specifications for cylindricity: **a** bird cage, **b** roundness profiles, and **c** generatrix

strategy characterized by 7 times the cutoff frequency of the Gaussian filter adopted, the filter being applied to each profile). Moreover, it is suggested that rectangular, triangular, and polar grids should be usually adopted for flatness, and the bird cage for cylindricity; the Union Jack, parallel profiles, roundness profiles, and generatrix strategies should be adopted only if one is mainly interested in geometric errors exhibiting some particular behavior (*e.g.*, if one is interested in conical deviation for a nominally cylindrical surface, a generatrix extraction strategy should be chosen).

Differently from low-density uniform strategies, because of the higher point density along each profile, the risk that sampling points are sampled only on crests or valleys of an undulated profile is reduced and, in general, the interaction between sampling strategy and actual geometric error is less relevant as a source of uncertainty. However, the ISO technical specifications do not give any indication about how many profiles should be sampled.

The ISO technical specifications group sampling strategies, consisting in just single points, patterned or not (*e.g.*, random or Hammersley strategies), under the name “points extraction strategies”. The technical specifications state that because of the usually small sample size adopted, these strategies are not as able to describe the geometric feature as profile extraction strategies, and present problems when filtering, so they suggest adopting them only when an approximate evaluation of geometric error is required.

Finally, regardless of the sampling strategy chosen, international standards assert that, with actual CMMs, it is hard to obtain a complete description of a surface, and therefore only specific information may be obtained on a surface. This means the inspection strategy should be carefully planned for the specific measurement task that is being performed.

A general limitation of any sampling strategy is some difficulty in defining it for a complex geometry. Blind strategies may be easily defined for features with a simple geometry, *e.g.*, straight profiles, rectangular planes, and disk planes, but they are rather complex when the feature to be measured is, *e.g.*, the flat surface of a plate in which holes have been drilled or a cylindrical surface with a slot. A possible, straightforward solution for such cases is to approximate the feature with a regular one, define the blind strategy for this feature, and then discard points which cannot be sampled. However, this solution cannot be considered optimal, and it may reduce the effectiveness of the sampling strategy. Further investigation on this subject is still required.

4.3.2 Adaptive Sampling Strategies

An adaptive sampling strategy is a strategy which does not adopt a predefined pattern for sampling points, but *adapts* itself to the real feature while measuring it. An adaptive strategy requires information pertaining to nominal geometry, tolerance values, the metrological characteristics of the measuring system adopted for the inspection, and the part to be measured. An adaptive sampling strategy starts from a few sampling points scattered throughout the surface to be inspected; then, depending on the information acquired through such points, more sampling points are added. Point addition may go on until some criterion is met (*e.g.*, the computed geometric deviation does not vary significantly anymore, or some evaluation of the lack of information in unsampled zones is sufficiently small, or a given maximum sample size has been reached).

Adaptive sampling is therefore a technique which aims to find critical areas of the surface to be inspected with a reduced sample size (if compared with blind sampling strategies). Reduction is made possible by not taking all points at one time, but by varying the sampling pattern according to the information that may be extracted from sampling points as they are sampled. This kind of strategy may be effective for slow measuring systems, which require every effort to reduce the sample size to control the measurement cost. Another advantage of an adaptive sampling strategy is that no planning of the pattern is required, and in some implementations not even the sample size is required – the operator chooses only the target accuracy for the measurement task and then the measuring system will automatically choose the measuring strategy. Adaptive sampling may then be considered as a fully flexible automatic way of planning the measurement strategy.

However, flexibility is the Achilles' heel of adaptive sampling; in fact, even though some methods for the automatic planning of the probe path have been proposed (Moroni *et al.* 2002), current CMM software applications are not able to automatically define a path for the probe which is guaranteed to be collision-free, owing to the presence of fixtures and other features of the part itself. Therefore, for commercial instruments, the adoption of fixed patterns for sampling points is justified by the need to avoid damage risks for the measuring system, risks that may be solved once and for all for static sampling patterns by correctly defining just one sequence of CMM movements. Adaptive sampling strategies may therefore be considered as the most promising strategies for the future; however, the integration of automatic probe path planning methods in current CMM control systems is required to allow adaptive sampling. If collision detection problems are eventually solved, adaptive strategies could be able to combine the advantages of both blind (suitability for inspection of a few parts and ease of definition) and manufacturing-signature-based (reduced uncertainty, given the sample size) strategies; however, more research is required to make adaptive sampling feasible.

An adaptive sampling strategy is defined essentially by three elements: a starting (blind) sample set, a criterion aimed at choosing the next sampling point(s), and a stopping rule.

The starting sample set is usually generated by means of a simple uniform sampling strategy, and it is characterized by a few points (at most ten or 20). Only in Badar *et al.* (2005) is it suggested that the starting set pattern should be based on the manufacturing process, the pattern itself being chosen empirically, by operator choice.

Since when verifying compliance to tolerances, only points showing the maximum local deviation are relevant, the criterion for choosing the next sampling point usually looks for such points. In particular, three ways have been proposed:

1. Sampling point deviations from the nominal geometry are fitted by means of a spline (Edgeworth and Wilhelm 1999). Then, this spline is adopted to predict deviations in those areas of the feature which have not been sampled yet. The area (or areas) that shows the largest deviation is chosen for sampling in the next step of the algorithm. However, it should be pointed out that a spline fitting a set of unevenly spaced sampling points may tend to introduce unwanted undulation, which could lead to a wrong choice of the location of the next sampling point.
2. Focusing on roundness, Rossi (2001) proposed that lobes on the roundness profile are identified from the starting set. Then, if lobing is significant, sampling points are concentrated in crests and valleys of the profile; otherwise, a uniform strategy is adopted. Because lobing is often found in roundness profiles, the criterion is efficient for finding maximum and minimum geometric deviations. Some limitations of this technique include its scarce adaptability to tolerances different from roundness; furthermore, if the initial sample size is small, some undulations may not be detected;
3. Direct search techniques are adopted (Badar *et al.* 2003, 2005). Most adaptive sampling techniques involve fitting the surface in order to choose the next sampling point(s) as the most deviating one. They may therefore be assimilated to numerical optimization techniques, aimed at identifying maxima and minima of a function, which is nothing other than the actual deviation of the geometric function as a function of the sampling coordinates. Differently from the previously indicated techniques, direct search optimization techniques do not require any surface fitting in order to find maxima (*i.e.*, choose the next point or points). Owing to the nature of most inspected surfaces, which often have sharp variations or noise present, this may be an interesting feature.
4. Barbato *et al.* (2008), adopting kriging interpolation (Cressie 1993) to fit sampled points, proposed a different criterion: because kriging interpolation allows for an evaluation of the sample variance of fits, the location characterized by the maximum variance is chosen. This approach is intended to ensure that the amount of information pertaining to the behavior of the surface is as uniform as possible in every area of the surface itself.

A few stopping criteria have been proposed. The easiest one is to stop when, after performing some iterations of the algorithm, a fixed number of sampling points is reached, leading to strategies characterized by a constant sample size. A more interesting criterion was proposed by Edgeworth and Wilhelm (1999), which evaluates the uncertainty for each iteration (*i.e.*, points addition), and stops when the uncertainty, which tends to decrease as the sample size grows, is sufficiently low.

Regardless of the method adopted, the research works mentioned above show that adaptive sampling is capable of reducing the number of sampling points required to achieve a given uncertainty by at least of an order of magnitude with respect to blind strategies. This result is really encouraging; however, the problem of online path planning has to be solved before adaptive sampling can be widely adopted.

4.3.3 Manufacturing-signature-based Strategies

Under the assumption that the manufacturing process leaves a "manufacturing signature" (Figure 4.11) on the part, that is, a typical pattern of geometric deviations, since only those areas of the surface/profile which deviate the most from the nominal geometry significantly affect conformance or nonconformance to geometric tolerances, knowledge of the manufacturing process may lead to strategies concentrating sampling points in such areas which are repeatable throughout the whole production.

Therefore, manufacturing-signature-based strategies can be defined as sampling strategies which, based on the nominal geometry of the part, tolerance values, metrological characteristics of the measuring system, and *some knowledge of the manufacturing process*, optimize the sampling point pattern in order to increase accuracy for *measuring one or more features manufactured by means of that particular manufacturing process.*

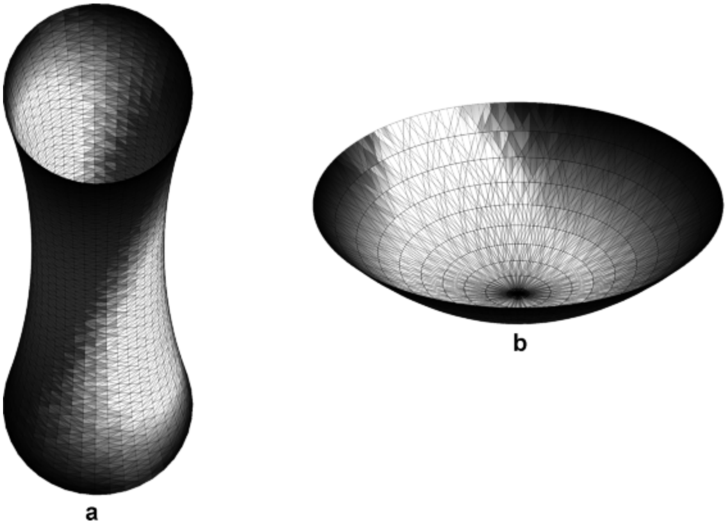


Figure 4.11 Signature examples: **a** shaft turned between centers, and **b** face-turned plane

Differently from adaptive strategies and similar to blind sampling strategies, the sampling pattern and the sample size are fixed once and for all. However, while blind sampling strategies tend to spread sampling points throughout the whole surface, signature-based strategies concentrate sampling points in those areas where, because of the repeatability of the manufacturing process, it may be expected that most deviating points are to be found. In order to do so, some knowledge of the manufacturing process is required. This knowledge may be implicit, *e.g.*, the only information available is a set of dense samplings of features manufactured by the same manufacturing process, or explicit, in which case a model (numerical or analytical) of the manufacturing signature has been defined, and the strategy is based on that model.

Given the sample size, signature-based strategies may allow one to improve accuracy with respect to blind sampling strategies; therefore, like adaptive strategies, they are particularly suitable when the cost of sampling a single point is high. Since the sampling point pattern is defined once and for all, the probe path may be defined once and for all as well, thus avoiding any risk of collisions between the probe and the part/fixture. Finally, implementation of process-based sampling strategies in a CMM control system is usually not very difficult, so they may be adopted with current measuring systems.

However, signature-based sampling strategies have a few disadvantages. First of all, since a process-based sampling strategy is specific to a particular manufacturing process, if the process changes, in order to keep uncertainty low the strategy should change as well. If the manufacturing signature changes because of undetected process failures, and the sampling strategy is not changed consistently, since sampling is not uniform, it is probable that critical areas of the surface will be missed. It follows that signature-based sampling strategies are usually less robust than blind or adaptive strategies. In order to adopt process-based strategies that minimize the risk of incurring relevant systematic measurement errors because of process modifications, appropriate statistical quality monitoring of geometric tolerances has to be adopted. Moreover, differently from blind and adaptive strategies, some information is required on the manufacturing process. If, as usual, this information comes from dense samplings of parts, the cost for performing these measurements should be considered. Therefore, care must be taken to evaluate whether the advantage of reducing the sample size (and then the measurement time and cost) justifies the effort needed to gather this information. Usually, this cost is not relevant if a medium- to large-scale production is considered, but it may become relevant for small-scale production. For small-scale production, solutions may come from models for predicting the process geometric deviations, *e.g.*, based on cutting parameters, which could replace the preliminary study of the process, but still have to be further developed. Finally, the adoption of a process-based sampling strategy requires a measuring system which leaves the operator completely free to define the sampling strategy; if a structured light scanner or a profiler is adopted, it will probably not be possible to freely concentrate sampling points in critical areas, thus reducing the effectiveness of the strategy. However, as has already been mentioned, since these instruments are usually fast enough to densely sample the whole surface, the adoption of a signature-based sampling strategy is somewhat meaningless.

Finally, it should be recalled that signature-based sampling strategies are effective if and only if the manufacturing process produces at least a partially systematic geometric deviation pattern; on the other hand, if the manufacturing process produces parts whose geometric errors are absolutely random, concentrating sampling points instead of spreading them throughout the feature to inspect is nonsense.

In order to improve measurement accuracy, if the typical behavior of the feature is known, then it is possible to try to predict the behavior of the measured feature even at those points which have not been sampled. This process is called "reconstruction", and may allow one to obtain a better approximation of critical areas of the surface by fitting them even if they have not been sampled really densely. This process is defined as "feature reconstruction".

Process-based sampling strategy may be split into two subclasses:

1. strategies based on process raw data, in which the geometric deviation pattern of the manufacturing process is assumed as available implicitly from a set of raw measurement data obtained from a collection of parts deemed representative of the manufacturing process; and
2. strategies based on a manufacturing signature model, in which the manufacturing signature is assumed as explicitly available as a (analytical, statistical, numerical) model.

4.3.3.1 Sampling Strategies Based on Process Raw Data

A strategy based on *manufacturing process raw data* is a signature-based sampling strategy which considers, as inputs, the nominal geometry and tolerances for the part, metrological characteristics of the measuring system, plus a set of *raw measurement data* obtained from a collection of parts deemed representative of the manufacturing process, and thus describing the manufacturing process itself. On the basis of this information, a strategy is proposed whose pattern is defined once and for all, and which is optimized for the manufacturing process considered. Raw data are measurement data *as given by the measuring system adopted*; they usually consist of clouds of points. The definition of a strategy based on process raw data starts with the measurement of a preliminary set of parts manufactured by the process for which the sampling strategy is intended. In order to ensure the parts are characterized by a stable geometric deviation pattern, the process has to be stable itself, *i.e.*, preproduction parts or parts produced during system ramp-up should be avoided. Ideally, the process should be under statistical control conditions. The number of parts in the preliminary set should be large enough to capture the overall manufacturing variability (usually ten to 20 parts are sufficient). For some sampling strategy methods, further specific information on these parts may be required, *e.g.*, they may have to be calibrated. Usually, the sampling strategy adopted for this preliminary data acquisition will consist of a blind strategy characterized by a point density as high as possible, to ensure that the amount of information collected is adequate. Most algorithms for defining strategies based on process raw data require the initial sampling strategy to be the same throughout the preliminary set. Because of these requirements, a dense uniform sampling strategy is adopted most of the time. In most cases, the same measuring instrument is selected to perform the preliminary measurements and for successive inspection. This choice ensures that the accuracy of the measurement system is implicitly considered in the definition of the sampling strategy; if different measuring systems are considered, different performances should be taken into account in the definition of the strategy.

Once all the required information is available, the method for defining the sampling strategy is directly applied to it, without any complex preliminary elaboration. The methods for setting up a measurement strategy based on raw data usually do not require significant human intervention, apart from data collection, thus making this kind of sampling strategy easy to plan, even for inexperienced operators. These strategies provide the typical accuracy improvements which are to be expected from a process-based strategy, while only adding the cost of acquiring raw data.

As previously pointed out, the performance of process-based sampling strategies is sensitive to manufacturing process instability; therefore, the success of such techniques is strongly dependent on successful monitoring of the process itself. However, the most efficient process control techniques available from the literature are all based on the availability of a model of the manufacturing process signature (Colosimo *et al.* 2008b, 2010), and therefore cannot be applied to this case, where a signature model is not available. If such a model were available, sampling strategies based on the manufacturing signature model should be adopted instead.

A few strategies based on manufacturing process raw data have been proposed in the literature for the evaluation of form error For several geometric features (*e.g.*, plane and straight line) the points that deviate the most from the nominal geometry always belong to the convex hull of the point cloud; given this, some strategies have been proposed that choose sampling points among those belonging to the convex hull of raw data. In Raghunandan and Rao (2007), it is proposed that sampling points are randomly chosen from these convex hull points; in order to ensure efficiency, it is also suggested that the strategy is tested on a few more parts before it is adopted. Buonadonna *et al.* (2007) proposed a hybrid approach where point selection is based on using the information provided by a signature model (hence the hybrid nature of the approach), with the application of the D-method (Montgomery 2004) (*i.e.*, minimizing the variance of the model parameters), and combining such information with the extraction of some more sampling points from the convex hull. This further addition is driven by point ranking, which in turn is based on the distance from the fitted model. The presence of a signature model in the method could suggest it should be listed as a manufacturing-signature-based approach; however, the authors suggest one “avoid a complete signature description” to enhance performance; therefore, not requiring a full signature model identification, the approach is commonly considered as belonging to the techniques based on raw data.

Colosimo *et al.* (2008a) proposed two different raw-data-based strategies. The first adapts a multivariate statistical technique known as the "principal component variables" technique (Cadima and Jolliffe 2001), in order to define the sampling strategy for geometric tolerances inspection. The input for the principal component variables technique is a series of vectors, each containing the local geometric deviations of a measured geometric feature and generated from a cloud of points (raw data). Vectors are then grouped into a matrix, which is numerically analyzed to highlight statistical correlation between deviations found at different coordinates. Finally, the pattern definition criterion is designed so that it favors the selection of that subset of sampling points which can retain *the largest fraction of the overall* variance; in other words, a point whose geometric deviation is correlated to the deviation of an already selected point, will not be selected because its selection would not add significant information to the overall explained variance.

The second method proposed by Colosimo *et al.* (2008a) is referred to as the "extreme points selection" method. It assumes that, given a point cloud, only a small subset of points actually define the geometric error; for example, after having fitted a plane to a cloud of points by means of the "minimum zone" principle (Anthony *et al.* 1996), only four points are sufficient the define the minimum zone itself, any other point showing a smaller local form deviation. These points are defined "extreme points". The extreme points selection method extracts extreme points from every cloud of points constituting raw data. Because of process repeatability, extreme points tend to concentrate in those zones of the surface which repeatedly show the maximum deviation. Sampling points will then concentrate in these areas.

Finally, if an uncertainty evaluation method is found which depends only on raw data and on the sampling pattern, it will be possible to try to directly optimize the estimated uncertainty value. This approach was proposed by Moroni and Petrò (2008); it is based on the application of optimization techniques, and will be further discussed in the next sections.

4.3.3.2 Sampling Strategies Based on the Manufacturing Signature Model

If a model of the manufacturing signature is available for the actual process, then a sampling strategy based on this model may be proposed. In other words, these are sampling strategies that take as input the nominal geometry and tolerances, metrological characteristics of the measuring system, and – in addition – a *model of the manufacturing signature*. On the basis of this information, strategies are proposed whose pattern is defined once and for all, optimized for the manufacturing signature. The model required by these methods may be either experimentally determined or generated through the application of some predicting technique that links the manufacturing signature to manufacturing process parameters. If the model is derived experimentally, the same considerations that were made for raw-data-based strategies apply here as well for the selection of sampling pattern and the size of the preliminary set. Since the manufacturing signature model usually does not describe the measurement error of the measuring system adopted, the measuring instrument adopted to collect data from which to derive the signature model does not necessarily have to be the same as that adopted for inspection, and therefore a high-accuracy measuring system should be adopted in order to generate models that are as accurate as possible. However, regardless of the origin of the signature model, the metrological characteristics of the measuring system adopted for successive inspection should be taken into account when dealing with the definition of the sampling pattern.

Most of the methods for planning signature-based strategies require particular kinds of signature models. As an example, the extended zone criterion, which will be introduced later, requires an ordinary least-squares regression model to describe the signature. In general, the choice of the applicable planning method is restricted to those methods suitable for the signature model available, or, if the method has been chosen, the correct kind of model has to be selected. Anyway, the signature models adopted will be mostly of statistical nature because they have to be able to capture the variability of the manufacturing signature. Moreover, models such as pure time series models are not suitable for modeling the signature because the description of a feature, even when considering correlation, does not consider systematic behavior. Regression models (Draper and Smith 1998) are the most common choice.

The identification of a manufacturing signature model requires greater effort than does the simple acquisition of raw data. However, one should remember that knowledge of the manufacturing signature leads to several advantages (Dowling *et al.* 1997; Colosimo *et al.* 2008b), in addition to allowing for optimization of the inspection strategy.

Although the problem of modeling the signature will not be directly addressed in this work, it is important to recall the three parts that make a signature model:

1. a *structure*, which defines the kind of behavior geometric deviations show (*e.g.*, roundness profiles are usually characterized by the presence of lobes, flat surfaces often show a polynomial surface, *etc.*);
2. an *average amplitude*, which defines, *on average*, the geometric deviation induced by the structure; and
3. a *random noise* quantification, which characterizes residuals not explained by the rest of the model.

Summerhayes *et al.* (2002) suggested the application of the V-optimality criterion (Montgomery 2004) to signatures described by regression models. If the manufacturing signature is described by a regression model, geometric deviation at those locations of the geometric feature which have not been inspected can be predicted. Moreover, a statistical variance evaluation can be associated with the prediction. The V-method consists in choosing the pattern of sampling points that minimizes the average variance of the predictions. This “extended zone” criterion has the drawback of considering only the *structure* of the signature, and not its *average amplitude* or *random noise*.

To overcome the limitations of the extended zone method, one of the authors of this work (Petrò 2008) proposed developing a sampling strategy based on tolerance intervals for a regression model describing the signature. Regression intervals, given the coordinates, define the upper and the lower bounds within which a given fraction of future observations at those coordinates will fall. The amplitude of the tolerance intervals depends on the concentration of the sampling points around the coordinates, on residual dispersion of *random noise*, and on the model *structure*, while the values of the upper and lower bounds of the interval are influenced also by the *average amplitude* at the coordinates considered. Therefore, it is proposed to choose the pattern of sampling points by minimizing the difference between *the maximum value of the upper bound, and the minimum value of the lower bound.* This approach has proven capable of concentrating the sampling points in those areas of the feature that typically deviate the most from the nominal behavior, taking into account every part of the signature.

Similarly to raw-data-based approaches, if a technique for evaluating the uncertainty associated with a manufacturing signature model is available, then it is possible to directly optimize the sampling strategy by minimizing such uncertainty. This method will be further discussed later.

4.3.3.3 Reconstruction Strategies

As introduced earlier, "reconstruction" is the process of obtaining a complete geometric model of a surface feature even when its sampling does not cover it entirely, or it is not dense enough.

Reconstruction may lead to some reduction of uncertainty, as it tries to predict feature behavior where it has not been sampled. Reconstruction is accomplished through fitting from available points. Understandably, some knowledge of a signature model would improve the fitting process; however, Yang and Jackman suggested applying reconstruction with generic models, such as kriging reconstruction (Yang and Jackman 2000) and Shannon reconstruction (Yang and Jackman 2002), on points sampled according to some blind sampling strategy (necessarily a uniform strategy for Shannon reconstruction). Even though this approach does not apparently require any assumption regarding the signature, it may be categorized as belonging to the manufacturing-signature-based strategies because its effectiveness is based on the presence of a manufacturing signature, since at least some spatial correlation must be present to ensure that the method is effective.

The method proposed by Yang and Jackman (2000) proposes "universal kriging" (Cressie 1993) as a method to fit the inspected surface. Universal kriging requires some modeling of the surface in order to be effective. The authors adopted generic kriging models to fit the surface; however, they admitted that a wrong choice of the model could badly influence the error estimate. Therefore, in a subsequent paper (Yang and Jackman 2002), "Shannon reconstruction" (Zayed 1993) was proposed, which is a fitting technique completely independent of the actual geometry, which solves the problem of the model choice. For kriging reconstruction, Yang and Jackman suggested a random sampling strategy, and for Shannon reconstruction, a uniform sampling strategy was chosen.

Some signature-based strategies have been proposed that make use of reconstruction techniques. Summerhayes *et al.* (2002) adopted the extended zone criterion, reconstructing the signature by means of *ordinary least-squares* regression. Moroni and Pacella (2008) proposed a reconstruction approach that is very similar to Yang and Jackman's (2002), as both of them adopt Shannon reconstruction in order to fit the model. The two approaches differ in the greater attention Moroni and Pacella gave to the harmonic content in the feature: Yang and Jackman adopted Shannon reconstruction "blindly", without considering the real behavior of the feature; Moroni and Pacella analyzed the influence of the real behavior of the feature on the effectiveness of reconstruction.

4.3.4 Effectiveness of Different Sampling Patterns: Case Studies

To better understand the effectiveness of different sampling strategies, let us introduce some examples. Sampling strategies will be compared in terms of measurement uncertainty (given the sample size). Adaptive sampling strategies will not be considered as they are not supported by current commercial CMM control systems and therefore cannot be adopted in production environments. The case studies will therefore aim to compare blind strategies and manufacturing-signature-based strategies.

Two case studies will be considered. The first one pertains to the assessment of flatness on face-milled planar surfaces: the strategies compared include the uniform strategy, Hammersley's strategy, and a raw-data-based strategy which will be described later. The second case study is related to roundness on shafts manufactured by turning, and will compare a uniform strategy and a model-based sampling strategy.

4.3.4.1 Strategies for Flatness: Face-milled Planes

The first case study considers face-milled planes, as defined in the ISO 10791-7 (1998) standard. A series of nine 160 mm × 160 mm planar surfaces have been machined on an MCM Synthesis machining center, as illustrated by Figure 4.12. The material chosen was a 6082-T6 aluminum anticordal alloy. The cutting parameters were as follows:

- spindle speed 3,000 rpm;
- feed rate per tooth 0.12 mm;
- depth of cut 0.5 mm;
- diameter of the mill 100 mm; and
- number of teeth 7.

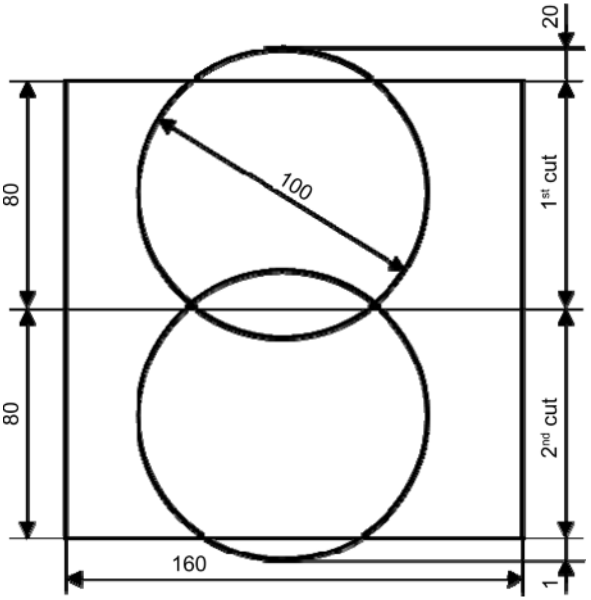


Figure 4.12 Milling path superimposition in face milling. Tool path segments are supposed to be horizontal

Machined surfaces were then measured with a Zeiss Prismo Vast HTG CMM, adopting a uniform strategy. A square 157×157 grid of points (a point per millimeter, leaving 2 mm of clearance from the edges) was sampled on each workpiece.

Figure 4.13 shows the mean surface obtained by averaging the *z* coordinates of the nine acquired grids, aligned in the *xy* plane. The mean surface gives an idea of the manufacturing signature: the surface is characterized by an overall smooth saddle geometry, a region with steeper variations and an abrupt discontinuity which is imputable to the superimposition of the two milling path segments (as illustrated in Figure 4.12).

To obtain the raw data needed to investigate this family of strategy planning approaches, a larger number of points is needed: to generate such additional data, a simulation-based solution is introduced. Of course, for real applications of the method, such points would be better obtained experimentally.

The simulation solution is based on generating additional points affected by random error. To assess the random error, the geometric deviation evaluated on 24,649 sampling points belonging to the experimental measurement of a single surface specimen has been taken as the reference value, with 1.5 μm standard uncertainty, as defined in ISO/IEC Guide 99:2007(E/F) (2007). Therefore, to generate a cloud of points constituting raw data, a random error, uniformly distributed in the ±2-μm interval, was added to the measured cloud of points. Ten clouds of points for each surface were simulated this way. The result of this simulation constitutes the raw data feeding the strategy planning method.

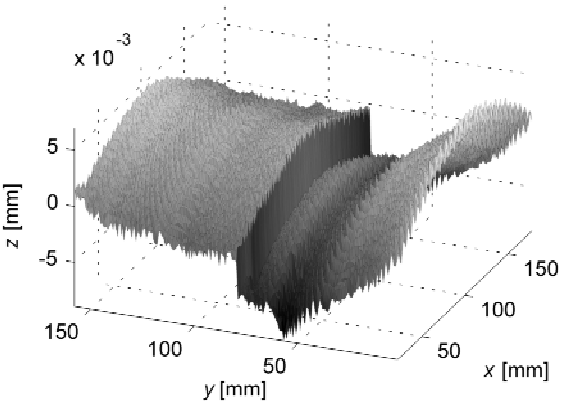


Figure 4.13 Mean surface for the nine face-milled planes

4.3.4.2 Process-raw-data-based Sampling Strategy

The comparison between sampling strategies will be in terms of measurement uncertainty. To be coherent with the choice of a raw-data-based sampling strategy, a method capable of evaluating uncertainty based on raw data only is introduced; it does not require a model for the behavior of the measuring system or of the signature. This uncertainty evaluation procedure will be the core of the strategy planning method: given the sample size, the pattern of points that minimizes the measurement uncertainty will be identified by means of a suitable minimization algorithm. This approach was first proposed by Moroni and Petrò (2008) and is here referred to as the "minimum *U*" approach, where *U* represents the measurement uncertainty as illustrated in the following.

The ISO/TS 15530-3 (2004) technical specification proposes a procedure to evaluate measurement uncertainty which is based on raw data obtained from repeated measurements of a calibrated artifact. From raw data, some terms are estimated, such as u_{cal} (uncertainty contribution due to the calibration uncertainty), u_{p} (uncertainty contribution due to the measurement procedure), u_{W} (uncertainty contribution due to the variability of the manufacturing process), and b (measurement bias). The term U [expanded uncertainty, see ISO/IEC Guide 98-3 2008 and ISO/IEC Guide 99:2007(E/F) 2007] is evaluated as

$$U = k\sqrt{u_{\mathrm{cal}}^2 + u_{\mathrm{p}}^2 + u_{\mathrm{W}}^2} + |b| , \tag{4.5}$$

where k is the expansion factor. The ISO 15530-3 specification allows one to compensate for the b term, thus reducing its influence; anyway, in the following discussion it will be supposed not to be compensated.

Mathematical expressions found in the technical specification allow for uncertainty evaluation if a single calibrated artifact is adopted in the procedure. However, a single artifact is not sufficient to describe the manufacturing signature variability. Moreover, if the sampling pattern optimization algorithm is applied to a single artifact, the resulting strategy will probably be very specific for that particular artifact, thus generating a strategy completely lacking in robustness. Therefore, a slight modification of the standard is proposed, so that more than one calibrated artifact may be used. In particular, the *b* term (the average bias) should be evaluated as

$$b = \frac{\sum_{j=1}^{m}\sum_{i=1}^{r_m}\left(y_{i,j} - x_{\mathrm{cal},j}\right)}{mr_m}. \tag{4.6}$$

In Equation 4.6 m is the number of calibrated artifacts adopted, r_m is the measurement repetitions number for each artifact, $y_{i,j}$ is the measurement result (estimated geometric error) of the ith measurement repetition of the jth artifact, and $x_{\mathrm{cal},j}$ is the reference value for the jth artifact (calibrated geometric error). It is supposed that each calibrated workpiece is measured the same number of times; to be as similar to the ISO 15330-3 standard as possible, it $r_m \geq 10$ is suggested.

Then, to estimate u_{p}, a pooled standard deviation could be used:

$$u_{\mathrm{p}} = \sqrt{\frac{\sum_{j=1}^{m}\sum_{i=1}^{r_m}\left(y_{i,j} - \bar{y}_j\right)^2}{m\left(r_m - 1\right)}}, \quad \bar{y}_j = \sum_{i=1}^{r_m}\frac{y_{i,j}}{r_m}. \tag{4.7}$$

Substituting Equations 4.6 and 4.7 into Equation 4.5, the required evaluation of *U* results.

A final note on u_{W}: ISO 15530-3 introduces it to take into account the “variability of the production”, that is, part-to-part differences in local form deviations; however, if more than one calibrated workpiece is measured, then u_{p} should contain this uncertainty contribution, and then $u_{\mathrm{W}} = 0$, as the international standard itself suggests.

Having identified a method to evaluate uncertainty based on raw data, the next step is to choose a sampling strategy that minimizes U (given the sample size *n*). Suppose that the raw data consist of r_m dense measurements of *m* calibrated parts, and that the sampling strategy is the same for every measurement. Then, sampling points corresponding to any sampling strategy may be extracted from these clouds of points. Extracted subsets of points can be introduced in the measurement uncertainty estimation procedure. The uncertainty evaluation obtained is, of course, influenced by the interaction between the sampling pattern considered and any typical geometric error left by the manufacturing process. If the sampling pattern is effective, i.e., it is able to detect regions of the feature that deviate the most, then the uncertainty will be low. The identification of an optimal pattern can be seen as an optimization problem where several different alternative patterns are compared. For the search of the optimal strategy, an optimization algorithm may be suggested. Given the discrete nature of the problem, genetic algorithms (Holland 1992) and simulated annealing algorithms (Kirkpatrick *et al* 1983) are suitable for the task.

4.3.4.3 An Example: Sampling Strategy for Face-milled Planes

A simulated annealing algorithm was applied to the experimental data described earlier in this paragraph, having chosen the number of sampling points *n* such that $n \in \{4, 9, 16, 25, 36, 49, 64, 81, 100, 121, 144, 169, 196, 225, 256, 289\}$.

In order to speed up the convergence of the simulated annealing algorithm, sampling points in the initial subset were chosen according to the "extreme points selection" criterion (Colosimo *et al.* 2008a).

The results from the uncertainty optimized strategy were compared with Hammersley and uniform strategy results. The resulting patterns are shown in Figure 4.14. As expected, the Hammersley and uniform sampling strategies do not show any area of the surface with a particular concentration of sampling points, and the Hammersley strategy does not show any pattern too. In contrast, by looking at the optimized sampling strategy with $n = 16$ (Figure 4.14c), it appears that the sampling points are completely concentrated in two small areas near the left and right edges of the surface, and along the lower and upper edges. Considering Figure 4.13, it is apparent that these are the areas of the surface which deviate the most from an ideal flat plane. When the sample size increases to 100, a similar concentration of sampling points is not apparent anymore. A higher concentration of sampling points may be still seen at the surface borders; however, most of the points appear uniformly scattered throughout the surface. Therefore, the method acts by trying to diffuse sampling points uniformly over the surface, while avoiding excessive concentrations of points at the borders. These points diffused uniformly on the surface will usually not be critical because they belong to regions which do not show the maximum deviations from the nominal behavior; however,

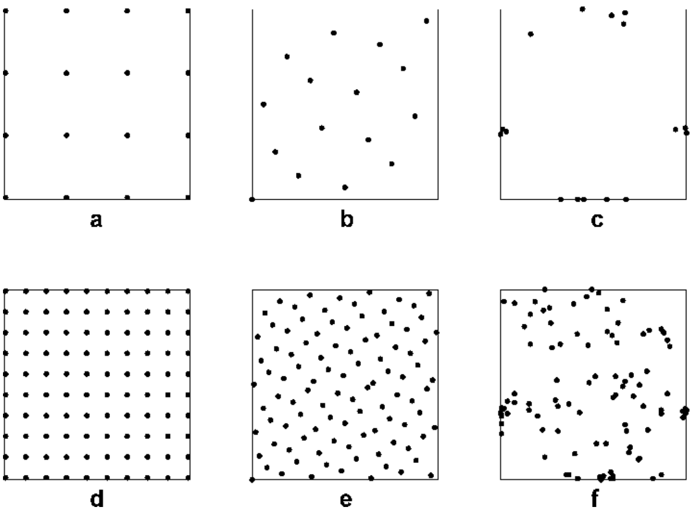


Figure 4.14 Strategies considered in the comparison: **a** uniform sampling, $n = 16$, **b** Hammersley sampling, $n = 16$, **c** optimized sampling, $n = 16$, **d** uniform sampling, $n = 100$, **e** Hammersley sampling, $n = 100$, and **f** optimized sampling, $n = 100$

they may be relevant if some anomaly happens in the manufacturing process, thus providing better robustness for the strategy.

Now, consider Figure 4.15, which shows the expanded uncertainty *U* for the uniform, the Hammersley, and the proposed strategy (with a coverage factor $k = 2$). First of all, it can be noted that, in contrast with what Lee *et al.* (1997) stated, the Hammersley strategy does not outperform the uniform one in terms of expanded uncertainty. Moreover, the uncertainty does not monotonically decrease with increase of the sample size. This is an example of negative interaction between the part form error and the sampling strategy: since extreme points are located in the same regions of the surface throughout the whole set of workpieces, if the (blind) sampling strategy adopted does not place any points in these regions, the uncertainty will be high, regardless of the sample size. On the other hand, being optimized for the particular manufacturing process being considered, the raw-data-based sampling strategy shows a significantly lower uncertainty. Moreover, for a sample size larger than 100 points, the uncertainty is nearly constant and mainly depends on the workpiece calibration uncertainty.

The different influences of bias and measurement procedure repeatability on the overall uncertainty are depicted in Figures 4.16 and 4.17. As is evident, the raw-data-based sampling strategy outperforms the uniform and Hammersley strategies in that it is both less biased and less uncertain. As highlighted in Figure 4.16, the bias term for the Hammersley and uniform strategies is about 5 times larger than the measurement procedure repeatability. This proves that blind sampling strategies, if the sample size is not large enough, tend to fail to detect extreme points, thus underestimating (biasing) the geometric deviation,. Differently, an optimized sampling strategy, if the sample size is large enough, can make the bias negligible. Finally, it can be pointed out that Figures 4.16 and 4.17 show really similar plots, as long as the sign is

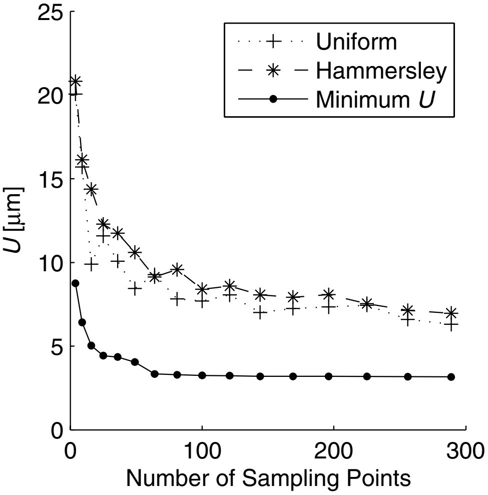


Figure 4.15 Expanded uncertainty for the case study considered and uniform, Hammersley, and minimum *U* strategies

ignored. Therefore, for an optimized sampling strategy, the uncertainty depends mainly on the bias; however, as just pointed out, the bias can be made negligible by a suitable raw-data based sampling strategy, thus making the uncertainty only dependent on the calibration uncertainty and the repeatability. In this particular case, the calibration uncertainty appears to be much more relevant than the repeatability.

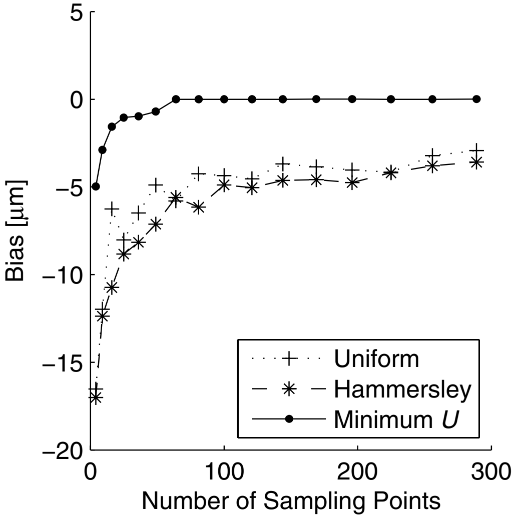


Figure 4.16 Bias for the case study considered and uniform, Hammersley, and minimum *U* strategies

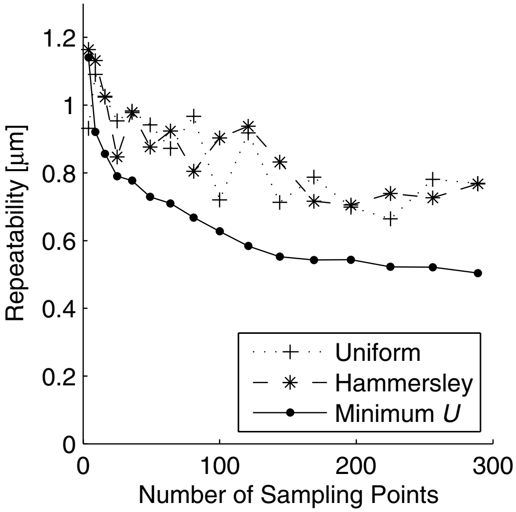


Figure 4.17 Measurement procedure uncertainty for the case study considered and uniform, Hammersley, and minimum *U* strategies

4.3.4.4 Strategies for Roundness

Usually, the standard strategy for profile measurement consists in uniform sampling. Consistently, roundness profiles are usually sampled at evenly spaced points. However, if a manufacturing process signature model is available, different strategies may be considered.

In the following, a signature model for turned profiles will be considered, and a signature based strategy will be proposed.

Signature-model-based Sampling Strategy

Along the lines of the raw data based sampling strategy proposed for a flat surface, the main idea here is to minimize the measurement uncertainty, once the sample size has been chosen. Uncertainty evaluation can be performed with the support of a signature model.

A similar method was proposed by the authors of the present work (Moroni and Petrò 2009), and is based on the "virtual CMM" concept. In particular, this strategy planning method is based on the idea of a virtual CMM which takes into account the presence of the manufacturing signature and its interaction with the actual sampling strategy when evaluating the measurement uncertainty.

A classic virtual CMM (Wilhelm *et al.* 2001; Balsamo *et al.* 1999) is based on the simulation of ideal, but not necessarily geometric-error-free, geometric features for which the geometric error x is known; a sampling error is simulated according to a model of the real behavior of the CMM for which the uncertainty is being evaluated, and is added to the ideal feature (which is referred to as "feature perturbation"). Measurement uncertainty is evaluated by comparing geometric errors evaluated on the perturbed features and the known real geometric errors of ideal features. The overall method may be regarded as a Monte Carlo simulation whose aim is to obtain values for the real geometric error x and the measured geometric error y, which are then subtracted to yield the measurement error x-y. A Monte Carlo simulation based virtual CMM able to take into account signature–sampling strategy interaction is not difficult to obtain: it is sufficient to simulate ideal features according to some signature model. If the simulated ideal features are generated according to some real signature model instead of "perfect features", then the uncertainty evaluation will implicitly consider the presence of the signature.

Various error sources should be considered in the simulation, including measurement strategy, environmental conditions, and CMM volumetric errors; a complete list of these sources may be found in the ISO/TS 15530-4 (2008) technical specification. This recently published standard deals with the problem of validating virtual CMM models, proposing four validation methods.

Several methods have been proposed to extrapolate uncertainty from simulation results. Here, the approach proposed by Schwenke *et al.* (2000) was adopted. This approach does not allow one to explicitly calculate a standard uncertainty u, but only an expanded uncertainty U characterized by some coverage probability p

(ISO/IEC Guide 99:2007(E/F) 2007). Suppose a Monte Carlo simulation of several (thousand) measurement errors *x*-*y* is available. From these data a Monte Carlo evaluation of the statistical distribution of *x*-*y* is derived. Let us define G (*x*-*y*) as the cumulative distribution of *x*-*y*. Therefore, an evaluation of the expanded uncertainty *U* characterized by the coverage probability *p* may be obtained by

$$\mathrm{G}(U)-\mathrm{G}(-U)=p\,. \tag{4.8}$$

Note that the resulting evaluation of *U* is coherent with the definition of the coverage probability, and that any uncorrected bias is considered (Schwenke *et al.* 2000).

To simulate geometric errors, an approach based on results obtained by van Dorp *et al.* (2001) was considered. The model developed is based on the frequency content of the error signal, and may be applied to a Zeiss Prismo CMM.

Example Strategies for Roundness

As case study for a manufacturing-signature-based sampling strategy, the roundness model proposed by Colosimo *et al.* (2008b), was considered. It consists of a "spatial error model of the second degree" (see Cressie 1993 for further details):

$$\begin{aligned}
&r(\theta_i)=\beta_1\cos(2\theta_i)+\beta_2\sin(2\theta_i)+\beta_3\cos(3\theta_i)+\beta_4\sin(3\theta_i)+u_i,\\
&\mathbf{u}=\sum_{i=1}^{2}\rho_i\mathbf{W}^{(i)}\mathbf{u}+\boldsymbol{\varepsilon},\\
&\boldsymbol{\varepsilon}\sim N_{748}\left(\mathbf{0},\sigma^2\mathbf{I}\right),
\end{aligned} \tag{4.9}$$

which, for a roundness profile, describes the deviation $r(\theta_i)$ from the average radius as a function of the angle θ_i; in the model, β_j are coefficients, $\mathbf{W}^{(i)}$ is the neighborhood matrix of the *i*th order, ρ_i is the *i*th spatial correlation coefficient, and σ^2 is the residuals variance.

The results obtained by applying a virtual CMM to estimate the measurement uncertainty and adopting a simulated annealing algorithm to minimize this estimated uncertainty by considering different sampling point patterns are similar to those proposed for the planar surface sampling strategy. The comparison is outlined in Figure 4.18, which shows the decrease of the expanded uncertainty as the sample size increases. Even though the uncertainty decreased for both the uniform and the optimized sampling strategy, it is evident that the uncertainty is consistently smaller for the optimized strategy.

Finally, one could think that because the same sets of 1,000 simulated profiles and 1,000 perturbed profiles was adopted to optimize the sampling strategy, the sampling strategy itself may be effective only for these specific simulated profiles. In order to prove the general effectiveness of the strategy, a new, independent

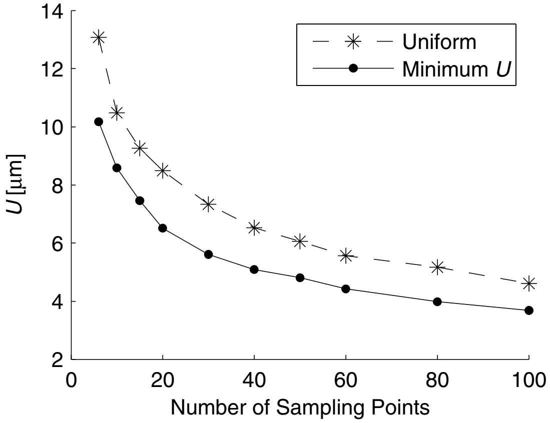


Figure 4.18 Expanded uncertainty for the case study considered and uniform and minimum *U* strategies

virtual CMM uncertainty evaluation was performed for every strategy proposed. The results are not distinguishable from those illustrated here, thus proving the general validity of the result.

The case studies show that process-based sampling strategies may outperform traditional, blind sampling strategies. Of course, if there is no need to reduce the sample size, and if the measurement uncertainty that can be obtained with a blind sampling strategy is adequate, there is no need to adopt a process-based sampling strategy. However, when the number of parts to be inspected is very large, or the tolerance is critical for the functional characteristics of the part, care should be taken to choose the correct sampling pattern.

4.4 Sample Size Definition

The criteria for planning a sampling strategy described so far allow one to define a sampling pattern having defined the sample size. However, the problem of choosing how many points to sample has not been addressed yet. This problem is known as *sample size definition*.

As may be gathered from Figures 4.15 and 4.18, as the sample size increases, the measurement uncertainty *U* tends to decrease. This progressive reduction of the measurement uncertainty may be explained by considering the more complete description of the measured feature that can be achieved with a larger sample size. Sometimes (*e.g.*, see Figure 4.15), when the sample size is small, the interaction between the manufacturing form error and the blind sampling strategy may make the decrease nonmonotonous; however, for large sample sizes, the decreasing behavior is restored. Moreover, even though the uncertainty decreases as the sample size increases regardless of the sampling strategy chosen (blind, signature-based *etc.*), the connection between the sample size and the uncertainty depends on the criterion adopted for pattern generation. For example, given the sample size, process-based sampling strategies may show a lower uncertainty with respect to blind sampling strategies.

Measurement uncertainty definition, and then sample size definition, is a problem of agreement between the customer and the manufacturer. This, in turn, is related to economic aspects of measurement, and, more generally, of inspection. Low uncertainty measurement procedures are usually more expensive. This is evident for coordinate metrology. It has already been pointed out that uncertainty is inversely proportional to sample size: for coordinate measuring systems which do not sample points simultaneously, such as 3D scanners, but sequentially (point-to-point such as a touch-trigger CMM or in a set of profiles such as profilers), a larger sample size often implies a longer measurement time, and a consequently higher measurement cost. However, uncertainty quantifies the dispersion of the values which may be attributed to a measurand (ISO/IEC Guide 99:2007(E/F) 2007), so larger uncertainty usually implies more inspection errors, and excessive inspection errors may lead to unexpected costs. When dealing with inspection of any product, two kinds of inspection errors are possible (Burdick *et al.* 2003): rejecting a conforming product ("false failure") and accepting a nonconforming error ("missed faults"). Both errors may generate an "inspection error cost" for the manufacturer and the customer. False failure generates a cost because a part that could be sold is discarded, or has to be reworked. Missed faults cost is usually indirect: the customer may reject a batch because of an excessive fraction of non-conforming parts in the batch itself or, if the part has to be assembled, it can make the final product defective. Of course, the probabilities of both false failure and missed faults increase with measurement uncertainty, so the inspection error cost is directly proportional to uncertainty. Measurement uncertainty definition is therefore the search for the optimal trade-off between the measurement cost and the inspection error cost, that is, minimization of the inspection cost.

The ISO/TS 14253-2 (1999) and ISO/TS 14253-3 (2002) technical specifications deal with the problem of finding an agreement between the customer and the manufacturer. After an agreement has been reached, the "Procedure for Uncertainty Management" (PUMA) method proposed in ISO/TS 14253-2 may be adopted to evaluate the uncertainty and define the overall measurement procedure, including the choice of the measuring system, part fixturing, identification of uncertainty sources, and so on (for further details on the PUMA method see the ISO/TS 14253-2 technical specification). Since the sample size definition is part of the measurement procedure definition, it is implicitly defined during the application of the PUMA method.

Anyway, when choosing the sample size, complete freedom is normally guaranteed only by point-to-point measuring systems, for example, a CMM equipped with a digital probe. Unfortunately, this kind of measuring system is the slowest, and thus the one with the higher measurement cost (even though it is usually accurate, thus reducing the inspection error cost). Another motivation for freedom reduction when defining the sample size is the presence of a time limit. If every manufactured part has to be inspected, the measurement time will have to be adequate with respect to the cycle time. Because an increase in the sample size causes an increase in the measurement time, a time limit creates an upper bound to the sample size. In this situation it may be important to adopt a nonblind strategy in order to reduce measurement uncertainty.

In the literature the problem of sample size definition has been addressed mainly for blind sampling strategies. A technique to estimate the expected evaluation error (essentially systematic measurement error) for the straightness error evaluation has been proposed: Namboothiri and Shunmugam (1999) proposed a method that is capable of choosing the right sample size once an upper bound for the measurement error has been defined. Some objections may be suggested: the criterion considers only systematic errors, but random error in measurements are at least as important as systematic ones, such as usual uncertainty evaluations show; the sampling strategies analyzed were limited to random sampling; and an arbitrary choice of the error level may be suitable when calibrating a single artifact, but the choice of the sample size for quality check of mass production should be based on agreement between the manufacturer and the customer. Anyway, an interesting consideration found in this article is:

> "Furthermore, this study clearly underlines the importance of the sampling pattern to be followed on the surface during form error measurement. The guiding principle should be to catch that point that has got maximum error. If we get this point at the initial stages of measurement then further prolonging the measurement process is not necessary. Hence the measurement time can be considerably reduced."

This consideration guides one toward adaptive or process-based sampling strategies, which have already been discussed.

Two approaches were proposed by Lin and Lin (2001a, b) dealing with the use of "gray theory" (Deng 1982) (which is typically applied in control systems) for evaluating the right sample size. A gray model, which could be considered as an evolution of a time series in which slope is considered, is adopted to predict the geometric error in the next inspected part. Then, if the predicted value is critical, *i.e.* it is very large, or significantly different from the previous one, the sample size is changed (increased). If the estimated values for the geometric error are stable, then the sample size is reduced. The sampling strategy was supposed to be uniform in these works, but the really interesting subject is the attention paid to the possibility that the manufacturing process modifies its typical behavior, thus leading to the necessity of recalibrating not only the manufacturing process itself, but also the measurement system. This is particularly true if a signature-based sampling strategy is chosen, because a modification of the process usually leads to a modification of the signature, thus making the signature-based sampling strategy inefficient, and perhaps damaging, because it will maybe tend to sample points only in areas of the features which are no longer the ones that deviate most from the nominal geometry.

Finally, Hwang *et al.* (2002) proposed a “hybrid neuro-fuzzy” approach for planning the sample size. This approach consists in modeling the behavior of an expert operator when choosing the sample size, that is, several measurement tasks are assigned to different operators, and, on the basis of the sample sizes, they proposed a hybrid neuro-fuzzy model of the expert operator behavior is developed. The sampling pattern suggested is based on the Hammersley strategy. Even though this can effectively mimic actual industrial practice, the method is mainly subjective (like the human operators that are used to calibrate it) and therefore may not provide adequate results with consistency.

4.4.1 An Economic Criterion for the Choice of the Sample Size

In this section, a cost function is proposed which is aimed at supporting the decision procedure for finding an agreement between the manufacturer and the customer on the measurement uncertainty, which, in turn, depends on the sample size. The cost function provides a measure of the overall “inspection cost”, *i.e.*, the sum of the measurement cost and the inspection error cost.

The general form of the inspection cost C_I is

$$C_I = C_M + C_E , \tag{4.10}$$

where C_M is the measurement cost and C_E is the inspection error cost. The cost function will therefore separately define these contributions.

Assuming that the cost associated with the measurement process depends essentially on the measurement time, and assuming the measurement time is related to the sample size (which, in particular, applies to point-to-point measuring systems), we can express C_M as

$$C_M = c_M t \approx c_M t_p n = c_p n , \tag{4.11}$$

where c_M is the hourly cost of the measuring system, t is the time required to perform the measurement task, t_p is the time for sampling a single point, $c_p = t_p c_M$ is the cost of sampling a single point, and n is the number of sampling points.

In Equation 4.11 the cost related to each sampling point is supposed to be constant, meaning that each point requires the same time to be sampled. This is not always true, because point-to-point distances may differ, and the travel time varies depending on the location of the point. Therefore, this evaluation of the measurement cost is only approximate. Moreover, the time required to set up the machine, align the part, *etc.*, is not considered because it does not depend on the sampling strategy.

Evaluation of C_E, the inspection error cost, is quite subjective, and depends on the approach chosen by the manufacturer to deal with inspection errors. In fact, declaration of a part to be defective while it is actually conforming may lead to it being discarded, or reworked (if possible), or to other expensive and unnecessary actions. It is even harder to define the cost associated with declaring as conforming an actually defective part, being related to the possibility of making some finished product not working, or with decreasing customer satisfaction. A simple approach is proposed in the following.

Suppose that conformance to a tolerance has to be verified, and the ISO 14253-1 (1998) standard is followed. According to the standard, a part is assumed as conforming if and only if the measurement result y is smaller than the specification limit SL reduced by the expanded uncertainty U. Moreover, suppose that x, the real geometric error of the part, behaves according to some statistical distribution (*e.g.*, a Gaussian distribution): if the uncertainty increases, a higher number of parts will be rejected, even if they should be accepted. Under these assumptions, C_E can be evaluated as

$$C_E = c_w P\left(\mathrm{SL} - U < x < \mathrm{SL}\right), \tag{4.12}$$

where c_w is the value of a part, or of reworking it, or of any action to be performed on the part itself when it is declared as nonconforming, and $P(\mathrm{SL}\text{-}U < x < \mathrm{SL})$ is the probability that the real geometric error falls between SL-U and SL; therefore, this probability represents the expected fraction of rejected or reworked conforming parts (Figure 4.19). In Equation 4.12 it was assumed that only an upper bound exists for x, as usual for geometric tolerances; however, if both upper and lower bounds exist, like for dimensional tolerances, Equation 4.12 can be easily modified.

It could be pointed out that this formulation of the inspection error cost does not take into account the cost of declaring as conforming an actually nonconforming part, that is, the cost of a type II error. However, the ISO 14253-1 criterion for assessing conformance was designed in order to avoid this kind of error. In fact, it is supposed that the probability that a measurement result y < SL-U is obtained when measuring a nonconforming part, namely, characterized by x > SL, is very small, which is exactly the aim of the criterion proposed by the international standard.

Finally, an evaluation of the measurement uncertainty is required to apply this cost function. This subject has already been addressed in previous sections and

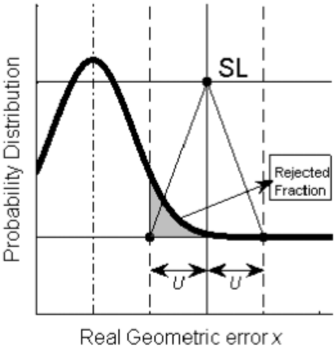


Figure 4.19 Rejected fraction of conforming parts

will not be treated here; any uncertainty evaluation technique may be fit for this method anyway.

Having defined a model for the inspection cost, and assuming that a sampling pattern (uniform, adaptive, process-based, *etc.*) has been chosen, so that the link between uncertainty and sample size is known, finding a trade-off between measurement cost and inspection error cost is straightforward. It is sufficient to identify the sample size that minimizes the inspection cost.

4.4.2 Case Studies: Roundness and Flatness

The case studies previously introduced for comparing different sampling patterns in terms of uncertainty may also be useful to compare different strategies in terms of inspection cost. The data required to perform this calculation are already available from preceding sections, with the exclusion of the statistical distribution of x, which may be derived from the parts measured to obtain the raw data, or from the manufacturing signature model.

In order to find the required trade-off between the measurement cost and the error cost, the optimization algorithm adopted in Sections 4.3.4.3 and "Example Strategies for Roundness" may be modified to choose the pattern, leaving the sample size free to change. This is computationally quite expensive, but leads to the correct trade-off.

Of course, if no process-based sampling strategy is of interest, a blind sampling strategy (*e.g.*, a uniform strategy) could be simply selected, and then the sample size may be left free to change: if the pattern is fixed, the link between sample size and uncertainty is univocally determined, and then the identification of the optimal

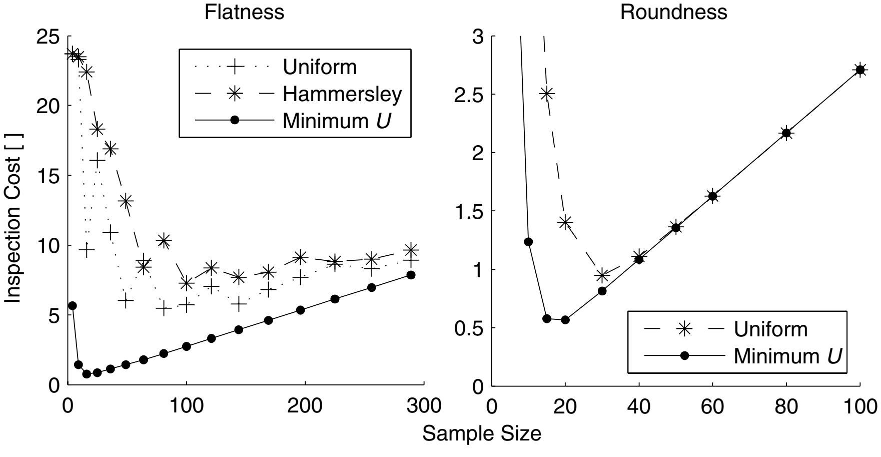


Figure 4.20 Inspection cost for flatness inspection, and roundness inspection

sample size is straightforward. Figure 4.20 clearly identifies the trade-off: if the sample size is small, then the inspection error cost is high, that is, several parts are wrongly rejected; if the sample size is large, inspection errors seldom occur, but the measurement cost is high because the sample size is large. Optimal sampling strategies may be identified between these conditions.

Different methods for defining the sampling pattern yield different inspection costs: in particular, process-based strategies show a significantly lower inspection cost, and a lower cost of the trade-off solution. However, it should be noted that as evident in the case of roundness, if the sample size is large enough there is no cost difference between blind and process-based strategies. This is because if the sample size is large enough, the blind sampling strategy sufficiently covers the feature to be inspected and provides an uncertainty that is not significantly different from the one granted by process-based strategy. Therefore, if there is no need to reduce the sample size because the inspection costs are considered not relevant anyway, a blind sampling strategy as dense as possible may be the correct solution, avoiding any effort to acquire data on the actual manufacturing process. However, if reducing the sample size is unavoidable (*e.g.*, because a very expensive measuring instrument is adopted or the measurement time available is short), the cost of acquiring raw data or a modeling signature, which are not accounted for in the inspection cost model, may be justified.

4.5 Conclusions

A correct sampling strategy definition is relevant for an effective CMM inspection. A wrong sampling strategy may lead to great measurement uncertainty, or to an unnecessarily high inspection cost.

Several methods may be proposed for planning the sampling strategy, but any method has to be able to define the sample size and sampling point pattern. Several coordinate measuring systems do not allow a completely free definition of the sampling strategy, so the operator is forced to choose some particular and specific sampling strategy.

The problem of the sampling pattern may be addressed with blind, adaptive, or process-based strategies. Blind strategies are easy to adopt, because they are based on simple and well-known patterns of points, and may be automatically generated by most CMM control systems. Adaptive sampling strategies allow for a reduction in the sample size, given the measurement uncertainty, which is important if the measuring system is slow and expensive, but their application depends on the control system of the measuring instrument, and current commercial control systems usually do not allow adaptive strategies. Process-based strategies are usually capable of achieving a low uncertainty with a small sample size, but require a research investment to acquire data on the manufacturing process; moreover, they are sensitive to process instability, which must be monitored by means of a suitable statistical control technique.

Finally, the problem of defining the correct sample size was depicted in its economic nature. The choice of the sample size, significantly influencing the measurement uncertainty, should be considered as an economic problem. A correct choice of the sample size should balance the inspection error cost, inversely proportional to the sample size, and the measurement cost, usually proportional to the sample size. A cost model was proposed as a support for finding this trade-off.

References

ISO 1101 (2004) Geometrical product specifications (GPS) – tolerances of form, orientation, location and run out, 2nd edn. International Organization for Standardization, Geneva

ISO 10360-1 (2000) Geometrical product specifications (GPS) – acceptance and reverification tests for coordinate measuring machines (CMM) – part 1: vocabulary, 1st edn. International Organization for Standardization, Geneva

ISO 10360-2 (2001) Geometrical product specifications (GPS) – acceptance and reverification tests for coordinate measuring machines (CMM) – part 2: CMMs used for measuring size, 2nd edn. International Organization for Standardization, Geneva

ISO 10360-3 (2000) Geometrical product specifications (GPS) – acceptance and reverification tests for coordinate measuring machines (CMM) – part 3: CMMs with the axis of a rotary table as the fourth axis, 1st edn. International Organization for Standardization, Geneva

ISO 10360-4 (2000) Geometrical product specifications (GPS) – acceptance and reverification tests for coordinate measuring machines (CMM) – part 4: CMMs used in scanning measuring mode, 1st edn. International Organization for Standardization, Geneva

ISO 10360-5 (2000) Geometrical product specifications (GPS) – acceptance and reverification tests for coordinate measuring machines (CMM) – part 5: CMMs using multiple-stylus probing systems, 1st edn. International Organization for Standardization, Geneva

ISO 10791-7 (1998) Test conditions for machining centres – part 7: accuracy of a finished test piece, 1st edn. International Organization for Standardization, Geneva

ISO 14253-1 (1998), Geometrical product specifications (GPS) – inspection by measurement of workpieces and measuring equipment – part 1: decision rules for proving conformance or nonconformance with specifications, 1st edn. International Organization for Standardization, Geneva

ISO 14660-1 (1999) Geometrical product specifications (GPS) – geometrical features – part 1: general terms and definitions, 1st edn. International Organization for Standardization, Geneva

ISO/DIS 10360-7 (2008) Geometrical product specifications (GPS) – acceptance and reverification tests for coordinate measuring machines (CMM) – part 7: CMMs equipped with imaging probing systems. International Organization for Standardization, Geneva

ISO/IEC Guide 98-3 (2008) Uncertainty of measurement – part 3: guide to the expression of uncertainty in measurement (GUM:1995), 1st edn. International Organization for Standardization, Geneva

ISO/IEC Guide 99:2007(E/F) (2007) International vocabulary of basic and general terms in metrology (VIM), 1st edn. International Organization for Standardization, Geneva

ISO/TS 12180-2 (2003) Geometrical product specifications (GPS) – cylindricity – part 2: specification operators, 1st edn. International Organization for Standardization, Geneva

ISO/TS 12181-2 (2003) Geometrical product specifications (GPS) – roundness – part 2: specification operators, 1st edn. International Organization for Standardization, Geneva

ISO/TS 12780-2 (2003) Geometrical product specifications (GPS) – straightness – part 2: specification operators, 1st edn. International Organization for Standardization, Geneva

ISO/TS 12781-2 (2003) Geometrical product specifications (GPS) – flatness – part 2: specification operators, 1st edn. International Organization for Standardization, Geneva

ISO/TS 14253-2 (1999) Geometrical product specifications (GPS) – inspection by measurement of workpieces and measuring equipment – part 2: guide to the estimation of uncertainty in GPS measurement, in calibration of measuring equipment and in product verification, 1st edn. International Organization for Standardization, Geneva

ISO/TS 14253-3 (2002) Geometrical product specifications (GPS) – inspection by measurement of workpieces and measuring equipment – part 3: guidelines for achieving agreements on measurement uncertainty statements, 1st edn. International Organization for Standardization, Geneva

ISO/TS 15530-3 (2004) Geometrical product specifications (GPS) – coordinate measuring machines (CMM): technique for determining the uncertainty of measurement – part 3: use of calibrated workpieces or standards, 1st edn. International Organization for Standardization, Geneva

ISO/TS 15530-4 (2008) Geometrical product specifications (GPS) – coordinate measuring machines (CMM): technique for determining the uncertainty of measurement – part 4: evaluating task-specific measurement uncertainty using simulation, 1st edn. International Organization for Standardization, Geneva

Anthony GT, Anthony HM, Bittner B, Butler BP, Cox MG, Drieschner R, Elligsen R, Forbes AB, Gross H, Hannaby SA, Harris PM, Kok J (1996) Reference software for finding Chebyshev best-fit geometric elements. Precis Eng 19(1):28–36

Badar MA, Raman S, Pulat PS (2003) Intelligent search-based selection of sample points for straightness and flatness estimation. J Manuf Sci Eng 125(2):263–217

Badar MA, Raman S, Pulat PS (2005) Experimental verification of manufacturing error pattern and its utilization in form tolerance sampling. Int J Mach Tools Manuf 45(1):63–73

Balsamo A, Di Ciommo M, Mugno R, Rebaglia BI, Ricci E, Grella R (1999) Evaluation of CMM uncertainty through Monte Carlo simulations. CIRP Ann 48(1):425–428

Barbato G, Barini EM, Pedone P, Romano D, Vicario G (2008) Sampling points sequential determination by kriging for tolerance verification with CMM. In: Proceedings of the 9th biennial ASME conference on engineering systems design and analysis, Haifa, Israel, CD-ROM

Bosch JA (ed) (1995) Coordinate measuring machines and system. Dekker, New York

Bracewell RN (2000) The Fourier transform and its applications. McGraw-Hill, New York

Buonadonna P, Concas F, Dionoro G, Pedone P, Romano D (2007) Model-based sampling plans for CMM inspection of form tolerances. In: Proceedings of the 8th AITEM conference, Montecatini Terme, Italy, CD-ROM

Burdick RK, Borror CM, Montgomery DC (2003) A review of methods for measurement systems capability analysis. J Qual Technol 35(4):342–354

Cadima JFCL, Jolliffe IT (2001) Variable selection and the interpretation of principal subspaces. J Agric Biol Environ Stat 6(1):62–79

Colosimo BM, Gutierrez Moya E, Moroni G, Petrò S (2008a) Statistical sampling strategies for geometric tolerance inspection by CMM. Econ Qual Control 23(1):109–121

Colosimo BM, Semeraro Q, Pacella M (2008b) Statistical process control for geometric specifications: on the monitoring of roundness profiles. J Qual Technol 40(1):1–18

Colosimo BM, Mammarella F, Petrò S (2010) Quality control of manufactured surfaces. In Lenz HJ and Wilrich PT (eds) Frontiers of Statistical Quality Control, vol 9. Springer, Wien

Cressie NAC (1993) Statistics for spatial data, 1st edn. Wiley-Interscience, New York

Deng JL (1982) Control problems of grey systems. Syst Control Lett 1(5):288–294

Dowling MM, Griffin PM, Tsui KL, Zhou C (1997) Statistical issues in geometric feature inspection using coordinate measuring machines. Technometrics 39(1):3–17

Draper NR, Smith H (1998) Applied regression analysis. 3rd edn. Wiley-Interscience, New York

Edgeworth R, Wilhelm RG (1999) Adaptive sampling for coordinate metrology. Precis Eng 23(3):144–154

Fiorentini F, Moroni G, Palezzato P, Semeraro Q (1992) Feature selection for an automatic inspection system. In: Proceedings of 24th CIRP international seminar of manufacturing systems, Copenhagen, Denmark, pp 199–208

Hammersley JM, Handscomb D (1964) Monte Carlo methods. Wiley, New York

Holland JH (1992) Adaptation in natural and artificial systems, 2nd edn. MIT Press, Cambridge

Hwang I, Lee H, Ha S (2002) Hybrid neuro-fuzzy approach to the generation of measuring points for knowledge-based inspection planning. Int J Prod Res 40(11):2507–2520. doi:10.1080/00207540210134506

Kim WS, Raman S (2000) On the selection of flatness measurement points in coordinate measuring machine inspection. Int J Mach Tools Manuf 40(3):427–443

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220(4598):671–680

Lee G, Mou J, Shen Y (1997) Sampling strategy design for dimensional measurement of geometric features using coordinate measuring machine. Int J Mach Tools Manuf 37(7):917–934

Lin ZC, Lin WS (2001a) Measurement point prediction of flatness geometric tolerance by using grey theory. Precis Eng 25(3):171–184

Lin ZC, Lin WS (2001b) The application of grey theory to the prediction of measurement points for circularity geometric tolerance. Int J Adv Manuf Technol 17(5):348–360

Montgomery DC (2004) Design and analysis of experiments, 6th edn. Wiley, New York

Moroni G, Pacella M (2008) An approach based on process signature modeling for roundness evaluation of manufactured items. J Comput Inf Sci Eng 8(2):021003. doi:10.1115/1.2904923, http://link.aip.org/link/?CIS/8/021003/1

Moroni G, Petrò S (2008) CMM measurement uncertainty reduction via sampling strategy optimization. In: Proceedings of the 9th biennial ASME conference on engineering systems design and analysis, Haifa, Israel, CD-ROM

Moroni G, Petrò S (2009) Virtual CMM based sampling strategy optimization. In: Proceedings of the 11th CIRP international conference on computer aided tolerancing, Annecy, France, CD-ROM

Moroni G, Polini W, Semeraro Q (1998) Knowledge based method for touch probe configuration in an automated inspection system. J Mater Process Technol 76(1):153–160

Moroni G, Polini W, Rasella M (2002) CMM trajectory generation. CIRP J Manuf Syst 31(6):469–475

Namboothiri VNN, Shunmugam MS (1999) On determination of sample size in form error evaluation using coordinate metrology. Int J Prod Res 37(4):793–804

Petrò S (2008) Geometric tolerances verification: strategy optimization for CMM measurement. PhD thesis, Politecnico di Milano, Milan

Raghunandan R, Rao PV (2007) Selection of an optimum sample size for flatness error estimation while using coordinate measuring machine. Int J Mach Tools Manuf 47(3–4):477–482

Rossi A (2001) A form of deviation-based method for coordinate measuring machine sampling optimization in an assessment of roundness. Proc IME B J Eng Manuf 215(11):1505–1518

Sartori S, Zhang GX (1995) Geometric error measurement and compensation of machines. CIRP Ann Manuf Technol 44(2):599–609

Savio E, De Chiffre L, Schmitt R (2007) Metrology of freeform shaped parts. CIRP Ann Manuf Technol 56(2):810–835. doi:10.1016/j.cirp.2007.10.008

Schwenke H, Siebert BRL, Wäldele F, Kunzmann H (2000) Assessment of uncertainties in dimensional metrology by Monte Carlo simulation: proposal of a modular and visual software. CIRP Ann 49(1):395–398

Schwenke H, Wäldele F, Weiskirch C, Kunzmann H (2001) Opto-tactile sensor for 2d and 3d measurement of small structures on coordinate measuring machines. CIRP Ann Manuf Technol 50(1):361–364

Schwenke H, Neuschaefer-Rube U, Pfeifer T, Kunzmann H (2002) Optical methods for dimensional metrology in production engineering. CIRP Ann Manuf Technol 51(2):685–699. doi:10.1016/S0007-8506(07)61707-7

Schwenke H, Knapp W, Haitjema H, Weckenmann A, Schmitt R, Delbressine F (2008) Geometric error measurement and compensation of machines – an update. CIRP Ann Manuf Technol 57(2):660–675. doi:10.1016/j.cirp.2008.09.008

Summerhays KD, Henke RP, Baldwin JM, Cassou RM, Brown CW (2002) Optimizing discrete point sample patterns and measurement data analysis on internal cylindrical surfaces with systematic form deviations. Precis Eng 26:105–121

van Dorp BW, Haitjema H, Delbressine F, Bergmans RH, Schellekens PHJ (2001) Virtual CMM using Monte Carlo methods based on frequency content of the error signal. Proc SPIE 4401:158–167. doi:10.1117/12.445616

Weckenmann A, Estler T, Peggs G, McMurtry D (2004) Probing systems in dimensional metrology. CIRP Ann Manuf Technol 53(2):657–684

Wilhelm RG, Hocken R, Schwenke H (2001) Task specific uncertainty in coordinate measurement. CIRP Ann 50(2):553–563

Yang TH, Jackman J (2000) Form error estimation using spatial statistics. J Manuf Sci Eng 122:262–272

Yang TH, Jackman J (2002) A Shannon sampling approach to form error estimation. Proc IME B J Eng Manuf 216(2):225–233

Yau HT, Menq CH (1995) Automated CMM path planning for dimensional inspection of dies and molds having complex surfaces. Int J Mach Tools Manuf 35(6):861–876

Zayed AI (1993) Advances in Shannon's sampling theory. CRC, Boca Raton

Standard under Development

ISO/CD TS 15530-2 Geometrical product specifications (GPS) – coordinate measuring machines (CMM): technique for determining the uncertainty of measurement – part 2: use of multiple measurements strategies in calibration artefacts. International Organization for Standardization, Geneva

Chapter 5
Identification of Microtopographic Surface Features and Form Error Assessment

Nicola Senin, Stefano Pini, and Roberto Groppetti

Abstract This work is concerned with quality inspection of microtopographic surface features, such as those that may be commonly found in semiconductor products, microelectromechanical systems, and other microcomponents. Surface microtopography data are assumed to be available as a height map, acquired through raster scanning over the region of interest, by means of a 3D profilometer or a 3D scanning microscope. An algorithmic procedure is proposed for form error assessment, which comprises several steps: first the feature of interest is localized and identified within the height map; then it is extracted and aligned with a reference (*i.e.*, nominal) geometry modeled by means of a CAD system; finally, form error is evaluated from the volume enclosed between the two aligned geometries. Feature identification is implemented through a modified version of the ring projection transform, adapted to operate on topography height maps; alignment comprises two steps (coarse alignment, consisting in an exhaustive search over discrete angular positions; and fine alignment, done with the iterative closest point technique). The final form error assessment procedure is applied to aligned geometries. The approach is illustrated and validated first through its application to an artificially generated case study, then to a real-life case of industrial relevance.

N. Senin
Dipartimento di Ingegneria Industriale, Università degli Studi di Perugia,
Via G. Duranti 67, 06125 Perugia, Italy,
e-mail: nsenin@unipg.it

S. Pini
Dipartimento di Ingegneria Industriale, Università degli Studi di Parma,
Parco Area delle Scienze 181/A, 43100 Parma, Italy,
e-mail: stefano.pini@nemo.unipr.it

R. Groppetti
Dipartimento di Ingegneria Industriale, Università degli Studi di Parma,
Parco Area delle Scienze 181/A, 43100 Parma, Italy,
e-mail: roberto.groppetti@unipr.it

5.1 Introduction

5.1.1 Scenario

This work is concerned with assessing *form error* of surface topography features manufactured at the micro and submicro scales, and whose geometry has been acquired by means of *3D profilometers* and/or *3D scanning microscopes.* This is a common scenario in quality inspection of microtopographic features manufactured on the surface of small parts such as semiconductor products, microelectromechanical systems (MEMS), and a wide array of other types of microcomponents. Additional valid examples include standard-sized parts on whose surfaces microtopographic patterns have been manufactured.

The peculiar aspects of the problem being considered, with respect to other form error assessment tasks, reside in the following considerations:

- Generally, the surface feature of interest is extremely small; under such premises, 3D profilometers and 3D scanning microscopes are usually selected as measurement instruments for quality inspection, since they are the only types of instruments which are capable of operating at such small dimensional scales with the required precision.
- Microtopography acquisition with 3D profilometers and 3D scanning microscopes is usually done through a *raster scanning process*, which consists in sampling *z* coordinates on points lying on a uniform *x*, *y* grid. Geometric data are available as a *height map*, essentially a discrete 2.5D geometry, which may also be considered as formally equivalent to a digital grayscale image (*z* coordinates being equivalent to pixel gray levels). Sampling is therefore often suboptimal with respect to the geometry and orientation of the feature being inspected.
- Given the demanding requirements in terms of precision, measurement instruments are often limited in terms of *range*. This leads to the difficulty of acquiring –within the same measurement – the entire topography of the feature being inspected together with the topography of additional datum surfaces which may be needed for localization of the feature itself, which may be located further apart; when this happens, feature localization errors are difficult to assess.
- When accurate localization is hard or impossible, common practice consists in acquiring a rectangular portion of the surface topography, covering a region larger than the feature, but within which the feature is known to be found; in order to assess form error, it is necessary to spatially register the measured geometry of the feature with its nominal counterpart.

Under these premises, current industrial approaches adopt *ad hoc* procedures depending on each specific application scenario, and a generalized solution is not available.

To favor repeatability and reproducibility, a generalized approach should include the adoption of algorithmic solutions for the issues being considered: computer-assisted data processing solutions are needed to separate the points belonging to the actual feature of interest from the remaining acquired scene points, for registering identified feature points to the nominal geometry, and for computing form error from the aligned geometries.

In this work, the nominal geometry of the feature of interest is assumed to be available as a CAD model, while the manufactured feature is supposed to be available as a part of a larger surface topography region, acquired by means of raster scanning with a 3D profilometer or a 3D microscope. Only 2.5D geometries are considered, as these are what most raster scanning devices are capable of acquiring. A novel algorithmic solution for identifying the subset of points actually belonging to the feature of interest, for aligning such points with the nominal geometry, and for performing a comparative assessment of the aligned geometries to evaluate the overall form error is presented, and is validated through the application to several case studies.

5.1.2 Main Terminology and Outline of the Proposed Approach

The nominal geometry of the feature of interest is referred to as the *template*. The template is supposed to be available as a *triangulated surface*, defined by a standard tessellated language (STL) model (STL is a common format for storing triangulated geometry), even though the proposed approach is generally applicable to templates defined by any mathematical surface representation.

The measured surface topography is referred to as the *scene*. Consistent with 3D profilometers for microtopography and 3D scanning microscopes, the scene is available as a simple set of height points organized over a regular grid (a height map). Raster scanning ensures that the topological arrangement of points is known (*i.e.*, they form a structured data set); thus, intermediate surface coordinate values can be obtained through interpolation on request.

The proposed approach comprises two main steps, briefly summarized in the following:

- *Feature identification and extraction*: A moving window, the size of the template, is used to scan the scene to search for positive matches with the template. Best-matching regions are selected as candidate regions (*i.e.*, scene regions containing the feature). It is assumed that in general more than one feature instance may be present in the scene; hence, more than one candidate may be found. Best-matching candidate regions are extracted and turned into standalone surfaces. At this point it is assumed that each stand-alone surface contains an instance of the feature.
- *Feature form error assessment*: Each extracted feature is aligned with the template through a sequence of coarse and fine alignment steps. The final alignment geometries can be used as the starting point for evaluating the form error.

5.2 Previous Work

From an overall point of view, the problem discussed here is fundamentally novel:

- The need to assess form error on micromanufactured surface features is recent and comes from the increased production of semiconductors, MEMS, and other microcomponents; current solutions are based on *ad hoc* approaches.
- The widespread usage of measurement solutions based on 3D profilometers and microscopes, combined with raster scanning, raises specific issues, unusual in other form error assessment problems, and related both to uniform sampling and to the 2.5D nature of measured geometry.

As briefly stated in the previous sections, the proposed approach identifies two main steps: *feature identification and extraction*, and *form error assessment*, the latter including *geometry alignment*. Each of these steps is related to specific literature domains, as illustrated in the following.

5.2.1 Previous Work on Feature Identification and Extraction

The analysis of the topography of manufactured surfaces at the micro and submicro scales, in particular as acquired by means of 3D profilometers, is currently based for the most part on approaches that rely on the computation of descriptors of mean surface texture properties (*e.g.*, roughness parameters) (Dong *et al.* 1994; Lonardo *et al.* 1996). Feature identification is still a relatively uncommon issue in the realm of micro- and nanotopography, and most of the previous work focused on the identification of basins and watersheds (Scott 1998; Barré and Lopez 2000) and on the application of filtering techniques such as wavelet filtering (Jiang *et al.* 2004; Zeng *et al.* 2005) for surface feature recognition.

Under the assumption of uniform sampling and the measured geometry being intrinsically 2.5D, the analogy with grayscale images holds, and surface topography analysis can obtain a great benefit from the adaptation of techniques originally developed within the realm of digital image analysis and processing. Feature recognition is no exception, as novel solutions can draw inspiration from the literature on *pattern recognition* for digital images. Overviews of pattern-matching-related issues for digital images can be found in Brown (1992) and Zitovà and Flusser (2003). The specific approach proposed in this work is based on preprocessing topography data before searching for matches. Preprocessing is done with a combination of *segmentation* and *edge detection* techniques: references on such techniques as applied to digital images can be found in Pal and Pal (1993) and Lucchese and Mitra (2001), while a solution specifically addressing surface topography data is proposed in previous work by the authors (Senin *et al.* 2007). The proposed approach for template matching is a novel solution resulting from the adaptation of a recent technique developed for images and based on the *ring pro-*

jection transform (Lin and Chen 2008). Transforms such as this one are *shape coding solutions*, *i.e.*, they turn shape information into a simpler numerical form, more suitable to be handled in quantitative shape comparison tasks, and generally designed to be invariant to shape localization, orientation, and sometimes scale as well (also called *shape descriptors*). Other notable approaches to shape coding include *moment-based encoding* (Flusser 2006), *ridgelets, fast Fourier transforms, and wavelet transforms* (see Chen *et al.* 2005 for a combined use of the aforementioned techniques).

5.2.2 Previous Work on Geometry Alignment and Form Error Assessment

Form error assessment implies aligning the geometry to be evaluated with its nominal counterpart. Since the goal is to evaluate form error from a complete 3D perspective, alignment must take place in three dimensions; therefore, alignment techniques developed for digital images – that operate on the more constrained image plane – cannot be used in this case. Instead, the literature related to alignment in 3D space must be considered.

One of the most established and well-known approaches for aligning geometries is the *iterative closest point* (ICP) technique (Besl and McKay 1992; Zhang 1994), with its variants (Audette *et al.* 2000; Rusinkiewicz and Levoy 2001). The ICP technique allows for alignment of geometries with great precision, but since the result is affected by the initial placement of the geometries, it is mainly used as a fine alignment technique. Thus, a coarse alignment solution must be obtained as a starting point for the ICP technique: among the most popular approaches available in the literature for coarse-alignment, many of which are summarized in Audette *et al.* (2000), it is worth mentioning those based on matching relevant axes, such as symmetry axes (Kazhdan *et al.* 2004), and principal component analysis axes, and those based on matching significant geometric features belonging to both geometries. In the latter case, features can be recognized by looking at local curvatures, *e.g.*, in the curvature scale space method (Mokhtarian *et al.* 2001), or by looking at other local shape descriptors defined to ease the recognition process, such as spin images (Johnson and Hebert 1999) and shape contexts (Kortgen *et al.* 2003).

Form error assessment is concerned with obtaining a quantitative measure of the difference between two geometries. The approach that has been followed in this work consists in aligning geometries first, and then computing the form error by combining local surface-to-surface Euclidean distances into a measure of the volume. Most 3D shape comparisons in the literature rely on *shape descriptors* instead, which, as mentioned earlier, allow for encoding shape information into simpler numerical forms. With descriptors, shape comparison becomes a matter of comparing descriptor results. Many of the techniques mentioned previously for feature identification and geometry alignment provide example of descriptors which are intrinsically capable of measuring shape similarity/difference. Additional survey work can be found in Loncaric (1998), Tangelder and Veltkamp (2004), and Funkhouser *et al.* (2005).

5.3 Outline of the Proposed Approach

In this section, the proposed approach is illustrated with the help of an artificially generated case study, designed to highlight the specific aspects of the problem and the steps of the procedure.

5.3.1 Simulated Case Study

To better describe the steps of the proposed approach, a simulated case study is introduced. The STL geometry shown in Figure 5.1a represents the *template* of an example surface feature to be recognized. In Figure 5.1b an artificially generated (*i.e.*, simulated) *scene* is shown, containing several instances of the template with different placement and orientation.

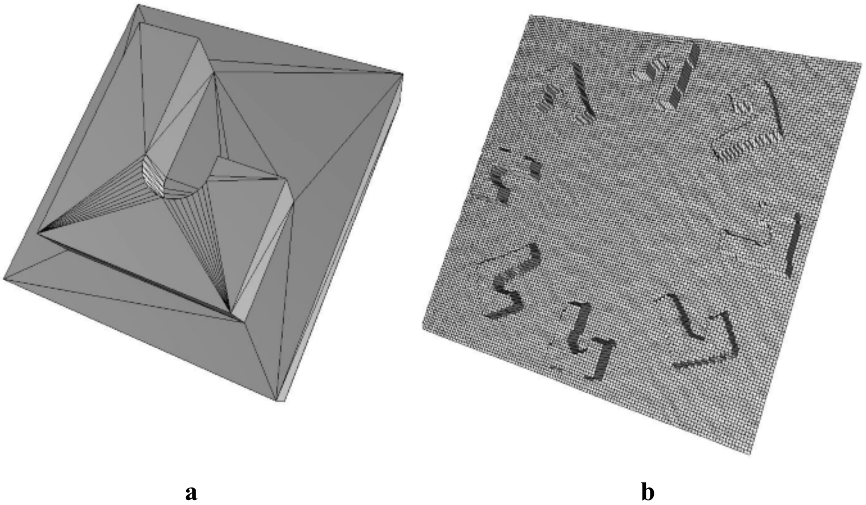


Figure 5.1 Geometries for the artificially generated case study: **a** template standard tessellated language (STL) geometry, and **b** simulated scene. The template is not visualized at the same scale as the scene

While the template geometry is supposed to represent the nominal conditions of the surface feature (and in this case also a portion of its surroundings), the scene model is meant to represent the result of a measurement process applied to a real manufactured surface, its geometry being affected by several sources of error. To obtain the simulated scene shown in Figure 5.1b, an STL model was generated first and populated with some feature instances with different placement and orientation (Figure 5.2a). Then it was altered by applying a sequence of transforms consisting in redoing the triangulation (Figure 5.2b) to achieve more uniform vertex spacing, and in the application of randomized algorithmic deformation effects such as local bends, twists, widespread waviness, and distributed Gaussian noise (also in Figure 5.2b). In this work simulated error had the sole purpose of introducing small modifications with respect to the nominal geometry, while no attention was given to reproducing actual manufacturing errors. Finally, the scene geometry was subjected to *virtual raster scanning*, *i.e.*, a simulated measurement process where a virtual 3D profilometer or 3D scanning microscope acquires points of a surface topography through raster scanning (Figure 5.2c). No simulated measurement error

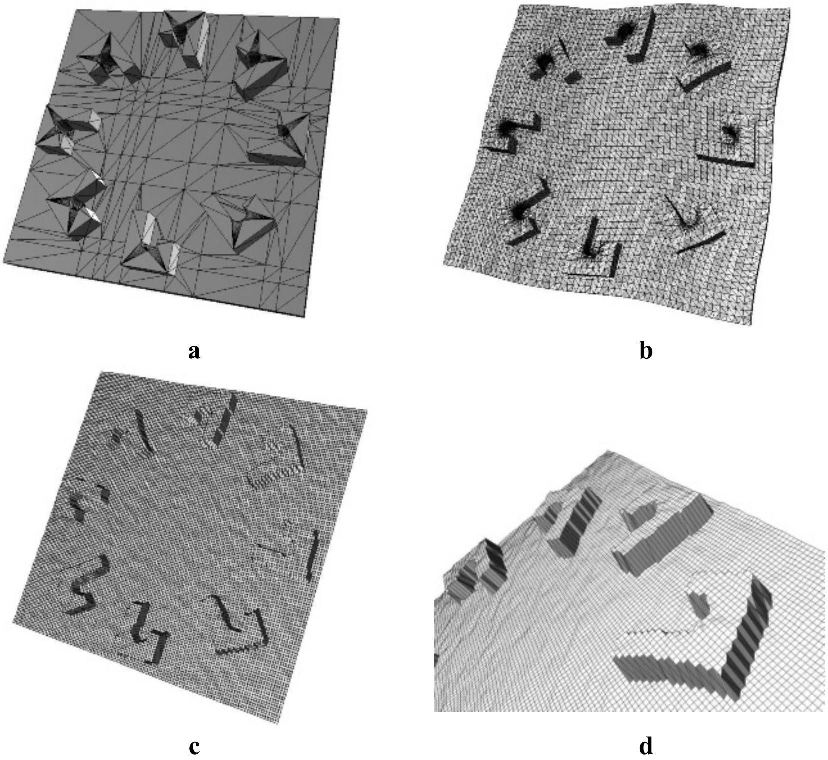


Figure 5.2 Creation of the simulated scene geometry: **a** original STL model, **b** new triangulation and application of algorithmic deformation effects, **c** virtual raster scanning, and **d** close-up view of two feature instances on the raster-scanned surface

was added to the simulation. As the typical conventions of 3D surface topography measurement are kept also in virtual raster scanning, the result of the simulated measurement process is a height map, *i.e.*, a set of points equally spaced in the x, y coordinates, each defined by a different z (height) value. Figure 5.2d shows a detail of the final surface. Figure 5.2c, d was obtained by triangulating the points obtained through virtual raster scanning to convey a clearer visual representation.

A visual comparison of the measured features shown in Figure 5.2d with the template in Figure 5.1a highlights the geometric differences to be expected in real-life operation, due to both the actual differences between the nominal and the manufactured geometry, and to the effect of raster scanning, regardless of considering measurement error (in the ongoing example, virtual raster scanning was performed with considerably coarse spacing to highlight its effect on measured geometry).

5.3.2 Overall Schema of the Proposed Approach

The complete structure of the proposed approach is summarized by the schema in Figure 5.3. The template geometry (STL model of the feature in its nominal state), and the scene geometry (the point set resulting from the actual measurement process, or the simulated one in the ongoing example) are supposed to be available as inputs to the process.

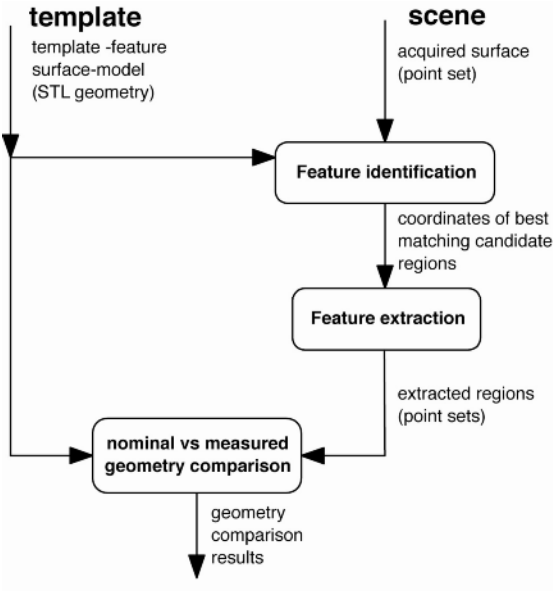


Figure 5.3 The overall process of feature recognition, extraction, and form error assessment for surface features against a template geometry

Through comparison with the template, best matches for the nominal feature are identified on the scene surface (*feature identification*) and are extracted as separate geometries (*feature extraction*). Each geometry extracted from the scene is then aligned to the template geometry to assess nominal-to-measured form error.

The single steps contained in the schema depicted in Figure 5.3 are discussed in detail in the next sections.

5.4 Feature Identification and Extraction

The detailed breakdown of the feature identification step is shown in Figure 5.4. Identification is performed through a main scanning loop, where a moving window that scans the original scene is used to select a portion of the surface to be subjected to pattern matching with the template. Both the template and the scene

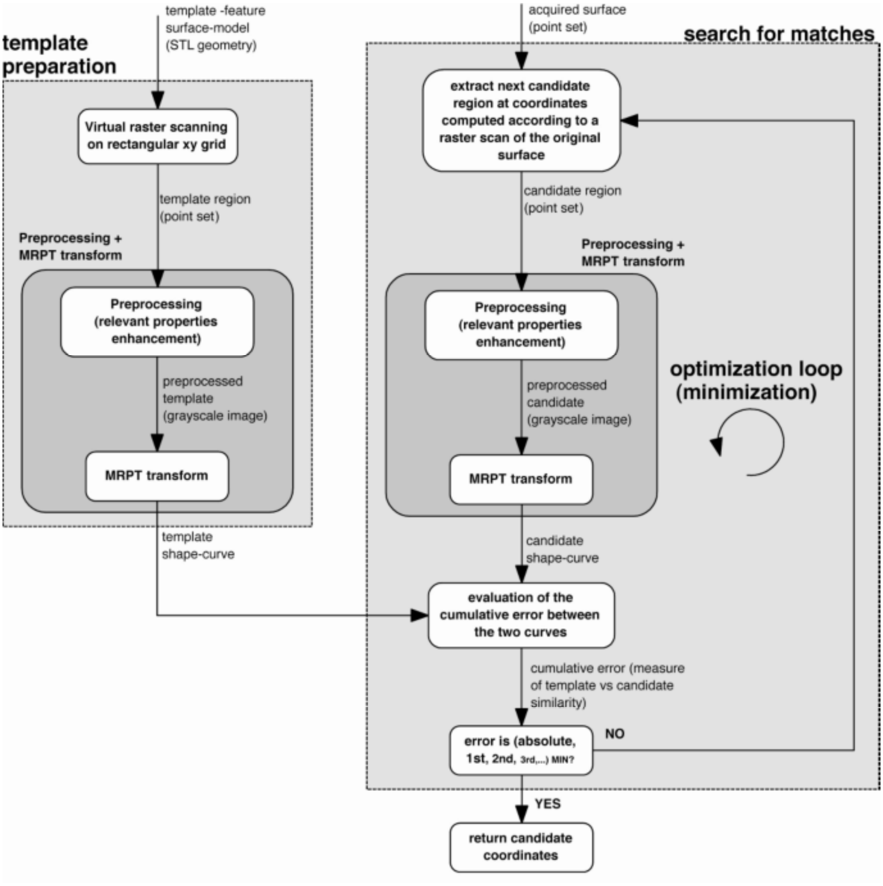

template preparation
template -feature surface-model (STL geometry)
Virtual raster scanning on rectangular xy grid
template region (point set)
Preprocessing + MRPT transform
Preprocessing (relevant properties enhancement)
preprocessed template (grayscale image)
MRPT transform
template shape-curve
acquired surface (point set)
search for matches
extract next candidate region at coordinates computed according to a raster scan of the original surface
candidate region (point set)
Preprocessing + MRPT transform
Preprocessing (relevant properties enhancement)
preprocessed candidate (grayscale image)
MRPT transform
optimization loop (minimization)
candidate shape-curve
evaluation of the cumulative error between the two curves
cumulative error (measure of template vs candidate similarity)
error is (absolute, 1st, 2nd, 3rd,...) MIN?
NO
YES
return candidate coordinates

Figure 5.4 The feature identification step. *MRPT* modified ring projection transform

geometry enclosed in the window are preprocessed first, to enhance salient attributes and to make the approach more robust to noise and other variations. The result of pattern matching is a matching score for each position of the moving window. At the end of the scanning process, window positions corresponding to the highest matching scores are selected as the best candidate results of feature identification, and the corresponding surface regions are extracted from the original topography as representative of positive identifications.

5.4.1 The Main Scanning Loop

The moving window is translated over the original scene surface by selecting its new center point at each step, according to a raster scanning strategy that covers all surface points. The window itself extends about the center point, and *zero padding* is used to make up for the missing points when the window is close to the scene boundaries. The region enclosed within the moving window is called the *candidate region*. For coarse investigation of a large scene, raster scanning can be modified to consider only a subset of scene points, by skipping a predefined number of positions while moving over the scene surface. The size of the moving window/candidate region is important, as it must match the spatial extensions of the template geometry.

5.4.2 Template Preparation

Before the actual comparison between the template and each candidate region can take place, the template, originally available as an STL model, must be turned into a form suitable for being compared with the candidate. For this purpose, the template is subjected to the same *virtual raster scanning process* the simulated scene was subjected to; if the scene is the result of an actual measurement process, virtual raster scanning of the template ensures that the measurement conditions applied when acquiring the surface topography of the scene are replicated for the template. This step must be done only once, and ensures that both the template and the candidate region topographies contain the same spatial frequencies, and thus are comparable.

5.4.3 Template and Candidate Region Preprocessing

Both the candidate region and the template undergo a preprocessing stage before being compared. Preprocessing is aimed at highlighting the most relevant attributes of the surface feature to be identified, and it makes identification less sensitive to nonrelevant forms of variation (disturbances). A correct choice of preprocessing operations may have a fundamental impact on the performance of the feature identification step. Since it acts as a filter for highlighting what is relevant, it must be designed depending on the specific feature one is after; on the other hand it should be considered that while the template geometry is preprocessed only once, the candidate region changes at each position of the moving window, and thus it must be preprocessed many times. Computational load may become an issue to consider before opting for overly complex preprocessing solutions.

Preprocessing may include *leveling* (by means of subtracting the *z* coordinates of the least-squares mean plane from the coordinates of the actual surface) if the relevant attributes of the feature are not deemed to be related to the overall orientation of the surface normal for the candidate region with respect to the overall surface. Some of the most common preprocessing approaches are illustrated in the following.

5.4.3.1 Height-based Binarization

The simplest preprocessing approach that can be adopted is *height-based binarization*, which consists in assigning 0-1 flags to surface points depending on whether the point height values are above or below a given threshold (*e.g.*, the mean height of the region). The resulting 0-1 map can be treated as a binary image. Binarization is ideal for describing surface features which may be summarized as 2D shapes either protruding or regressing with respect to the surrounding surface. In the case of the ongoing example, binarization has the effect of returning a sharp outline of the feature, as shown in Figure 5.5 for the template, and for a candidate region belonging to the scene in the ongoing example.

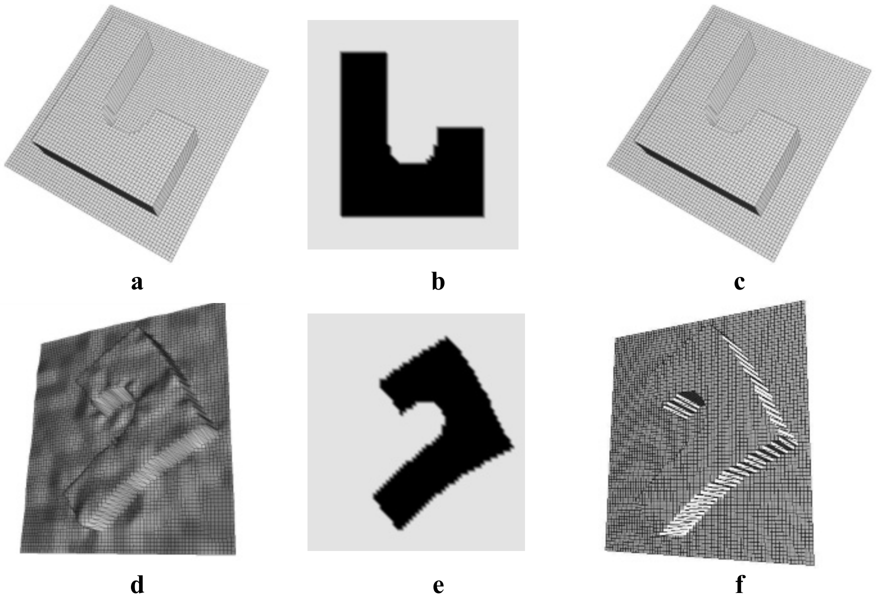


Figure 5.5 Preprocessing with height-based binarization: **a** original raster-scanned template, **b** binary map of the template, **c** binarized template topography, **d** candidate region from the scene, **e** binary map of the candidate, and **f** binarized candidate topography

With reference to Figure 5.5 one can see that in this case the effect of preprocessing is to turn the identification problem into a 2D binary pattern matching problem. Of course this is applicable when a 2D silhouette is deemed as sufficient for discrimination; other, more complex cases may require more complex preprocessing strategies.

5.4.3.2 Attribute-based *n*-segmentation

Attribute-based *n*-segmentation refers to a segmentation process, *i.e.*, a *partitioning* process, where a surface topography is divided into *n* regions, each featuring some uniform shape and/or texture properties. The result of a segmentation process is a map of scalar values, each value representing the identifier of a specific region. This greatly simplifies the scene, and allows for highlighting elements of interest (with a careful selection of the properties driving the segmentation process) while at the same time removing most of the disturbances that differentiate the nominal geometry from the measured one. The proposed approach implements surface topography segmentation by means of a solution developed previously by Senin *et al.* (2007), and is fundamentally based on computing point properties and then using *k-means clustering* to group points with uniform properties.

An example segmentation suitable for application in preprocessing is *slope-based* segmentation, which works well in describing features that can be seen as collections of roughly planar surfaces lying at different orientations (*e.g.*, the example surface depicted in Figure 5.6a). Slope-based segmentation can be used as the starting point for identifying feature *edges* (as the boundaries of the segmented regions), as illustrated in Figure 5.6b, c. For features defined by irregular surfaces, this approach is often more robust than direct detection of edges based on local curvature analysis.

The preprocessing approaches mentioned above are only some of the options available; the choice of a suitable preprocessing strategy strongly depends on the

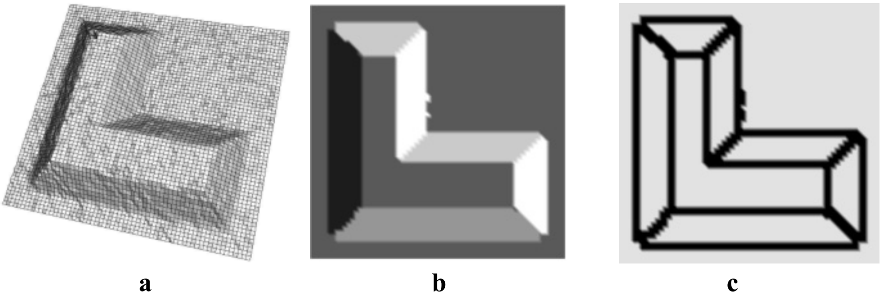


Figure 5.6 Preprocessing with slope-based segmentation to detect feature edges: **a** example surface topography, **b** segment map (each color identifies a region with uniform slope), and **c** edge map (edges are the borders of the uniform-slope regions)

peculiar aspects of the feature recognition problem. In any case, the result of the preprocessing step applied to a rectangular portion of surface topography (whether the template or a candidate region) consists of a map of scalar values, which in terms of data structures can be assimilated to a grayscale image and handled as such, thus paving the way to use pattern matching algorithms developed in the realm of digital image processing.

5.4.4 Template and Candidate Region Comparison Through Pattern Matching

Once the template and the candidate region geometries have been preprocessed and the corresponding maps of scalar values have been generated, a quantitative comparison of the template and the candidate must be done. From a general standpoint, a quantitative comparison between shapes relies on the availability of quantitative measures of shape, *i.e.*, *shape descriptors*, mapping shape properties to numerical values. Once such values are known, shape comparison can be mapped to a simple difference of values.

The choice of an appropriate shape descriptor depends on the nature of the data structures it must operate upon, and on the specific requirements of the feature identification process. In this case, the requirements are, on one hand, to cooperate with the preprocessing substep to ensure there is the required robustness (to noise and partial occlusion) and, on the other hand, that the descriptor must capture the salient traits of the shape without being influenced by shape *localization* and *orientation*. Robustness and invariance requirements are often counterbalanced by discrimination power, *i.e.*, with the capability of telling apart two different shapes.

The proposed approach draws inspiration from the *ring projection transform*, recently applied in the literature to pattern matching problems for digital images (Lin and Chen 2008). This work adopts a modification of the transform, which will be referred to as the modified ring projection transform (MRPT), illustrated by Equation 5.1 and with the support of Figure 5.7.

$$C\left(r_j\right) = \sum_{i=1}^{n_\theta} r_j d\theta \cdot z\left(x_c + r_j \cos\theta_i, y_c + r_j \sin\theta_i\right). \tag{5.1}$$

An integral (a summation, in discrete form) is computed on the *np* values of the map points lying on the circumference. This process is repeated at each radius from the center, and the results are collected together into a curve, which is shown in Figure 5.7b.

The biggest difference with respect to the original formulation consists in assigning more weight to the features that are far from the center, which is like saying that shape elements that experience a larger displacement when the feature is rotated count more in the overall determination of the descriptor.

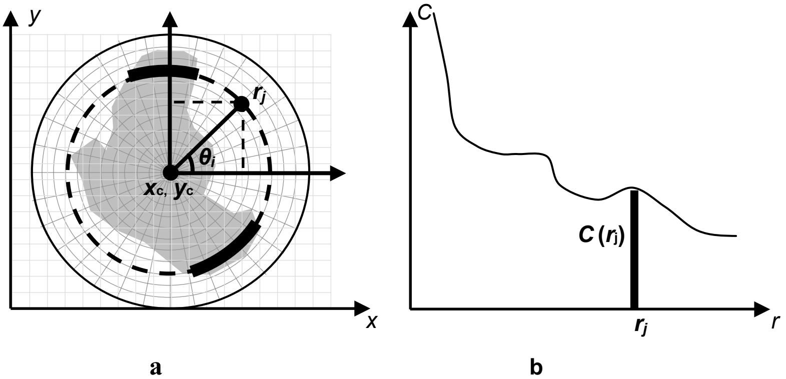


Figure 5.7 Conceptual schema of the MRPT: **a** geometric construction of the circumference each integral is computed upon, and **b** resulting MRPT curve

To compute the similarity (or difference) of two shapes, one can build the two MRPT curves, and then compute their difference in several ways. In this work three main approaches were investigated:

1. the cross-correlation coefficient computed on the two curves;
2. the maximum value of the cross-correlation function computed on the two curves; and
3. the integral of the point distances of the two curves.

The second solution captures curve-shape differences but is less sensitive to the relative positioning of the two curves. The first and third solutions are sensitive to relative positions of the curves. The third approach was selected as the most promising, and was implemented according to the expression illustrated by Equation 5.2:

$$\mathrm{Err}_{tot} = \frac{1}{r_{\max}} \sum_{j=1}^{n_r} \left| C_{tpl}\left(r_j\right) - C_{scn}\left(r_j\right) \right|. \tag{5.2}$$

The cumulative error Err_{tot} is computed as the summation of the errors at each r_j value, C_{tpl} being the MRPT curve of the template, and C_{scn} the MRPT curve of the candidate region (portion of the scene). To make the cumulative error independent of the feature size, the cumulative error is divided by the maximum radius of the MRPT. Finally, remember also that this approach returns an error, while cross-correlation-based approaches return a similarity measure. To treat this as a similarity problem, the error function must be inverted.

Finally, notice that all the approaches to MRPT curve comparison, Equation 5.2 being no exception, capture cumulative differences between the two curves, but not local ones, which is like saying that more sophisticated ways to compare the MRPT curves may lead to an increase in the discrimination power of the similarity metric selected.

5.4.5 Some Considerations on the Sensitivity and Robustness of the Preprocessed-shape Comparison Substep

Given that the MRPT integral is computed over a circumference, the angular orientation of the shape does not change its value. However, the position of the center each circumference is referred to does make a significant difference; hence, the approach is independent of orientation but not of localization of the shape. Secondly, it should be noted that any shape containing the same amount of "mass" (*i.e.*, values) at the same radial distance from the center will be encoded with the same $C(r)$ value; this is a limitation in terms of discriminative power of the approach. In other words, the "averaging" effect of integration, on one hand, increases the robustness to small shape variation, but, on the other hand, decreases the discriminative power as different shapes may be encoded similarly owing to what was illustrated above. Remember, however, that the descriptor is sensitive to localization of the given feature. This is actually not a problem, as the center of the transform is actually the center of the moving window translating on the original surface topography (see Section 5.4.4). The best match between MRPT curves will be obtained corresponding to the match of the template center and the feature center (see the examples in Figure 5.8).

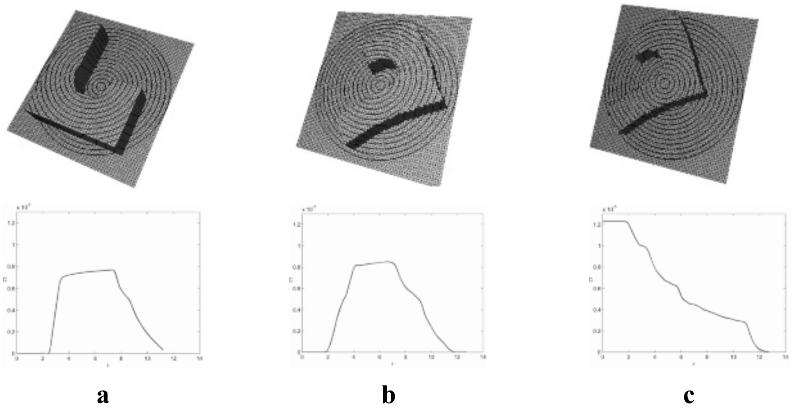


Figure 5.8 MRPT construction on binarized topography, and corresponding MRPT curves: **a** template, **b** candidate feature, center-aligned with template, and **c** off-centered candidate feature

5.4.6 Final Identification of the Features

At each step of the raster scanning process, a candidate–template match score is generated by the procedure illustrated earlier. At the end of the scanning process, each scene point ends up being associated with a match score, all the scores being

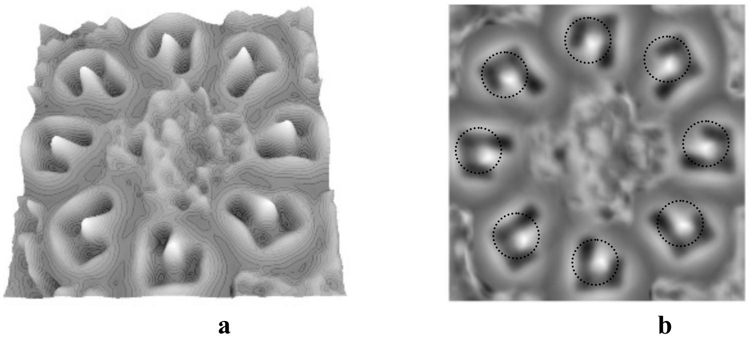


Figure 5.9 Match response surface for the ongoing example; a z-const contour plot has been overlaid to better visualize the location of the local maxima: **a** perspective view, and **b** top view (local maxima are *circled*)

viewable as a surface, whose z coordinates are the actual scores themselves. This surface is referred to as the *match response surface*; the match response surface obtained for the ongoing example is shown in Figure 5.9. In a match response surface, peaks (local maxima) are the points where the best matches between the candidate region and the template were found; *i.e.*, they correspond to the best guesses in terms of identifying the feature of interest on the scene surface. In the result obtained for the ongoing example, shown in Figure 5.9, the first eight highest maxima correspond exactly to the centers of the features in the scene surface, meaning that all eight instances of the feature were positively identified, regardless of their orientation. In the example, the highest matching scores are for the scene features that have undergone the least deformation with respect to the template, while the most deformed ones achieve lower scores.

It is expected that as the differences between the scene and template features increase, some positive identifications may be missed; at the current state of research it is difficult to determine when this may happen, as it also depends on the intrinsic properties of the feature being analyzed, on measurement process parameters, and on the preprocessing step discussed previously.

5.4.7 Feature Extraction

The local maxima of the match response surface correspond to the centers of the scene regions to be extracted as the best candidates for the feature identification problem. The process of extracting such regions from the original surface topography is called *feature extraction*: since it basically involves just an extraction of a finite set of points belonging to the original scene set, the process is straightforward and does not need to be discussed further. Consistently with the raster scanning process, if the center point is located close to one of the scene borders, the extraction will use zero padding to determine the values of the missing points.

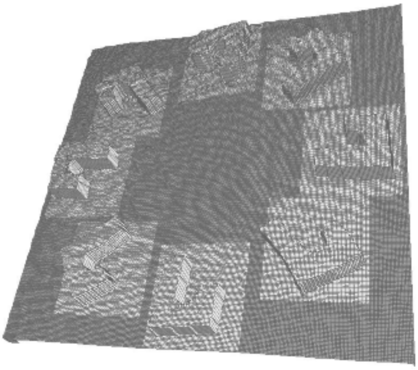


Figure 5.10 Regions to be extracted as the best-matching candidates in the original scene topography (their centers correspond to the first eight local maxima in the match response surface shown in Figure 5.9)

In Figure 5.10 the extracted regions (centers corresponding to local maxima in the match response surface) are highlighted on the original scene topography); it can be seen that the procedure results in all positive matches for the scene feature instances.

5.5 Nominal Versus Measured Feature Comparison

The last part of the proposed approach is related to performing an accurate comparison of the template and each measured feature (the regions extracted from the scene as the best-matching candidates) in order to assess the *form error*. The overall process is described in Figure 5.11.

Form error assessment implies aligning the geometries of the template and the recognized feature first, and then computing the differences between the two aligned geometries. However, the entire process is completely dependent on the type of form error that must be assessed. Several types of form error may require *datum surfaces* to be identified, if this is the case, datum surfaces should be defined on the template, their counterparts located on the candidate region, and both used for alignment. In this work it is assumed that no datum is available; therefore, the only alignment option consists in *globally aligning the entire geometries*, minimizing the sum of squared distances computed between each template point and its closest counterpart on the candidate region.

The global alignment problem discussed here is special since at the end of the feature identification and extraction steps, the template (nominal feature) and the candidate region (measured feature) are already partially aligned. In fact, feature identification itself, by locating the points corresponding to the best matches, provides a hint on the relative placement of feature centers; however, no information is given concerning rotational alignment, the MRPT being invariant to rotation.

Therefore, the alignment problem becomes for the most part a rotational alignment problem (about the *z*-axis), even though some rotational and translational alignment about the other axes may still be necessary to compensate for discretization and preprocessing-induced errors in the matching stages.

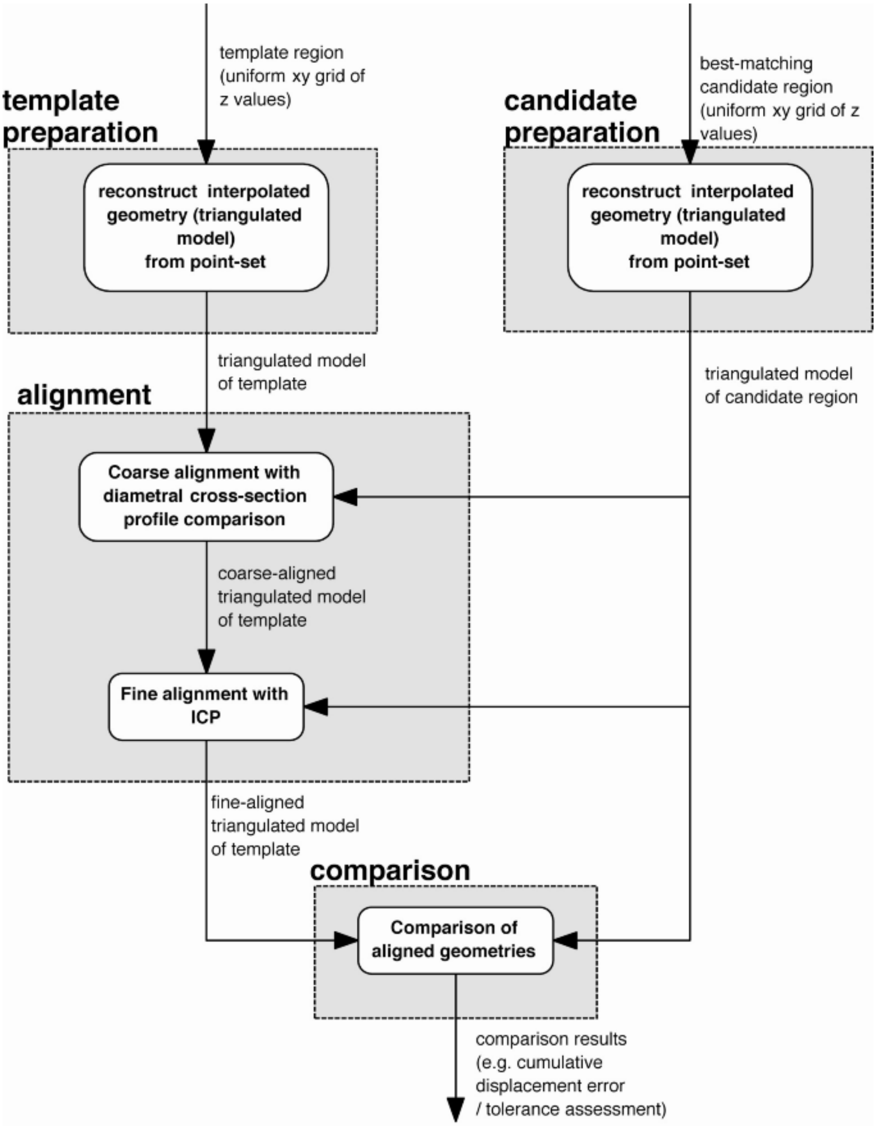

template preparation
template region (uniform xy grid of z values)
reconstruct interpolated geometry (triangulated model) from point-set
candidate preparation
best-matching candidate region (uniform xy grid of z values)
reconstruct interpolated geometry (triangulated model) from point-set
triangulated model of template
triangulated model of candidate region
alignment
Coarse alignment with diametral cross-section profile comparison
coarse-aligned triangulated model of template
Fine alignment with ICP
fine-aligned triangulated model of template
comparison
Comparison of aligned geometries
comparison results (e.g. cumulative displacement error / tolerance assessment)

Figure 5.11 Template versus measured feature comparison step. *ICP* iterative closest point

5.5.1 Coarse and Fine Alignment

The most widespread technique available in the literature for aligning two geometries is known as *iterative closest point* (ICP) (Besl and McKay 1992; Zhang 1994). ICP is suitable for application to the problem discussed here. However, it is well known that ICP tends to converge to local minima; therefore, geometries need to start from an initial position which is quite close to the optimal result. While in this case this is true for translational alignment, it definitely does not hold for rotational alignment, since the matched feature and the template may be in completely different angular orientations with respect to the z-axis. In order to deal with this problem, in this work ICP was used as a fine alignment solution, while a preliminary coarse alignment step was accomplished by means of another technique based on rotational alignment about the z-axis only.

5.5.2 Template and Candidate Geometry Preprocessing for Alignment Purposes

Both the template and the candidate region need to be preprocessed for better alignment. Recall that the candidate region originates as a rectangular portion of the original scene, defined by a regular grid of equally spaced points. The candidate region is turned into a triangulated geometry so that a continuous (C^0) surface representation is available in case interpolated points are needed (and in fact they will be needed by both the coarse and the fine alignment procedures). The template geometry is retriangulated as well, even though it was already provided as a triangulated geometry from the beginning (recall that the template was originally defined from STL geometry). The goal of triangulation is to produce a mesh finer than the original to facilitate optimal alignment with algorithms that operate on points.

5.5.3 Coarse Rotational Alignment with Diametral Cross-section Profile Comparison

To accomplish the coarse alignment step (rotational and about the z-axis), the following procedure is proposed. First, the template and candidate geometries are trimmed into circular shapes (as seen from a z-aligned viewpoint), the diameters being selected so that the feature is completely enclosed within the circumference. This is done to eliminate the peripheral points lying in the rectangular corners of the topographies, which would influence too much the overall outcome of angular alignment.

The template and candidate geometric centers are registered within a common Cartesian reference frame, then both geometries are sliced with diametral planes passing through the z-axis and through the geometric centers, each creating a diametral cross-section profile (a silhouette) of the topography. A predefined number of cross-section profiles is computed, equally spaced at discrete angular values about the *z*-axis (see Figure 5.12a, b).

After cross-section profiles have been generated, the entire range of relative, discrete angular displacements between the template and the candidate is tested. At each rotational displacement, the silhouette profiles of the template and the candidate are paired according to the current displacement angle, and the overall cumulative error between the two profile sets is computed with the same formula used to compare MRPT curves (see Equation 5.2). The cumulative error can be plotted as a function of the relative angular displacement, yielding the *match score profile* (see Figure 5.12c). The smaller the error, the better the match between the two profile sets, and hence the better the alignment between the two geometries. Therefore, the absolute minimum of the match score profile is taken as the optimal *coarse alignment angle*, to be used for rotating the template into its coarsely rotationally aligned position with respect to the candidate (see Figure 5.12d).

The coarseness of the alignment is mainly controlled by angular discretization, which in turn depends ultimately on the complexity of the feature to be aligned. This choice is currently *ad hoc*: for the ongoing example, 256 angular values were considered for coarse alignment (although only 32 radial profiles are rendered in Figure 5.12a, b, d).

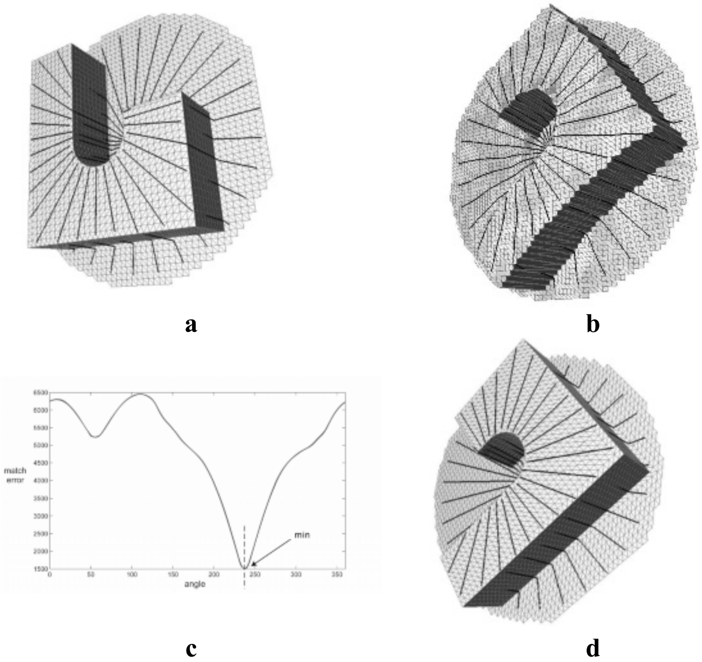


Figure 5.12 Coarse alignment procedure applied to the ongoing example: **a** template with radial profiles, **b** unaligned candidate and radial profiles, **c** match score profile (minimum corresponding to optimal alignment angle), and **d** rotationally aligned candidate

5.5.4 Fine Alignment with ICP

In the ICP procedure implemented in this work, the disk-shaped coarse-aligned template plays the role of the *model* (according to ICP terminology; Besl and McKay 1992), while the disk-shaped candidate region is the ICP *scene*. The ICP procedure consists in an iterative process where the model is brought closer to the scene to improve alignment. At each step of the iteration, each model point is paired with its closest counterpart in the scene. A rigid transform is then found that minimizes the squared sum of distances between the points of each pairing; this is known as *the absolute orientation* problem, for which a closed-form solution is available in the literature (Horn 1987). The resulting rigid transformation matrix can be used to move the model points (the template geometry) closer to the scene points. Given the new position of the model, new pairings can be generated with scene points, and the minimization process can be repeated. The iterations continue until a termination criterion is met, usually related to the rate of change in the sum of squared errors computed at each absolute orientation resolution step.

Figure 5.13 shows the ICP procedure applied to the ongoing example. Recall that the original geometries were previously subjected to coarse alignment. Their initial position is shown in Figure 5.13a for the ongoing example: the darker mesh

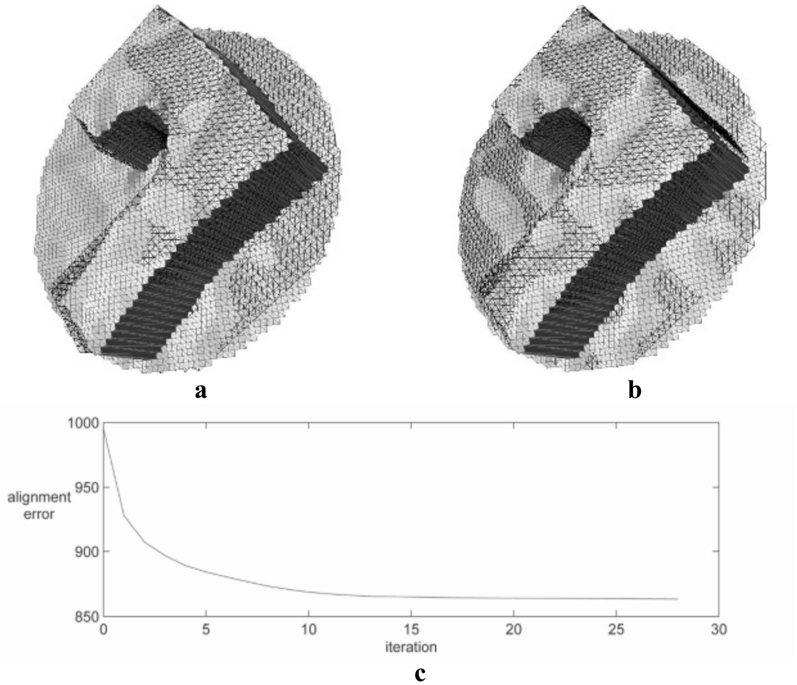


Figure 5.13 ICP fine alignment applied to the ongoing example: **a** model and scene in their original position, **b** model and scene in their final aligned position, and **c** displacement error plot

belongs to the template, while the light-gray mesh belongs to the candidate geometry; the final alignment, reached after about 30 iterations, is shown in Figure 5.13b. Figure 5.13c shows the cumulative displacement error as a function of the ICP iteration; in this implementation convergence to a final alignment is assessed by setting a minimum threshold value on the variation of the displacement error.

5.5.5 Comparison of Aligned Geometries

Given the underlying assumption of the absence of reference datums (hence the global alignment approach), form error can be treated as a quantitative measure related to overall differences between the template and candidate geometries. However, the candidate geometry cannot be directly compared with the template geometry yet. The template in fact represents a geometry which, even when ideally replicated on the manufactured surface, would still appear different at measurement, owing to the very same nature of the acquisition process: raster scanning introduces a discretization which – at these scales – usually cannot be neglected. Form error assessment becomes the problem of comparing the candidate geometry with an ideal geometry, *as it would appear after measurement*.

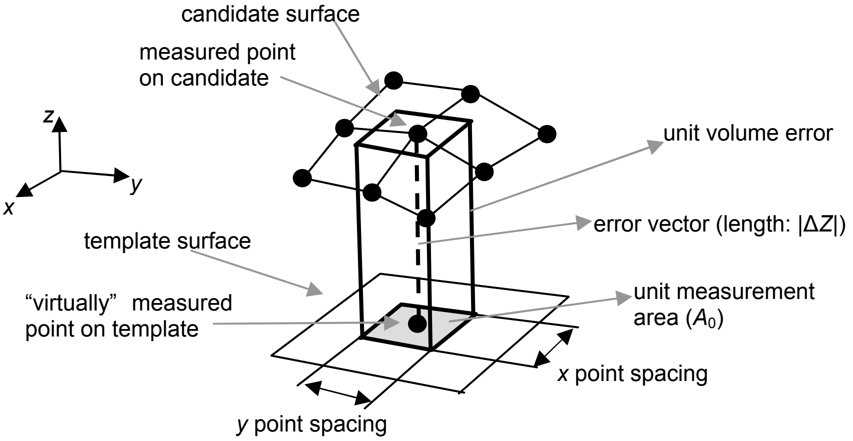

candidate surface
measured point on candidate
z
y
x
unit volume error
error vector (length: |ΔZ|)
template surface
"virtually" measured point on template
unit measurement area (A_0)
x point spacing
y point spacing

Figure 5.14 Geometric construction for the error vector and the unit volume error

To accomplish this, we propose a form error metric that works as follows. From each measured point belonging to the candidate, a vector is drawn in the *z* direction (measurement direction) to intersect the template surface (see Figure 5.14): the length of such a vector represents the error between the measured point on the candidate and a point that would be measured on the template, also the *x,y* coordinates (which is equivalent to considering the template geometry after

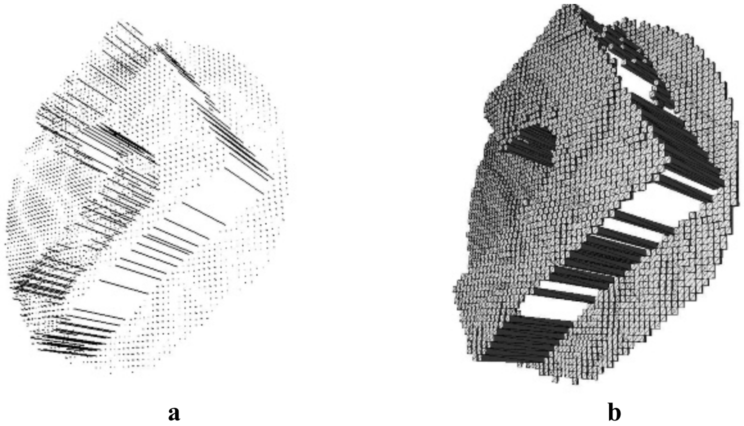


Figure 5.15 Error vectors in the ongoing example: **a** error vector map, and **b** unit volume errors

virtual raster scanning). The *error vector* can be used to compute a *unit volume error* by multiplying it by the unit measurement area, which in raster scanning is *x point spacing* times *y point spacing* (see Figure 5.14).

The sum of all unit volume errors computed for all candidate points that have a corresponding counterpart on the template is the *total volume error*. The approach is also summarized by Equation 5.3:

$$E_{tot} = A_0 \cdot \sum_{i=1}^{n_c} \left| \Delta z_i \right| , \tag{5.3}$$

where *n* is the number of candidate points that have a corresponding counterpart on the template, the index *i* refers to the *i*th point, and the other variables are documented in Figure 5.14.

In Figure 5.15a the *error vector map* (the set of all error vectors) is shown for the ongoing example; a close-up view of error vectors defined between the candidate and template geometries is shown in Figure 5.15b.

The ongoing example is useful for highlighting a peculiar property of this form error metric, which is more evident with steplike features: *x*,*y* misalignments between the candidate and the template are responsible for most of the overall volumetric error (see Figure 5.14).

5.6 Validation of the Proposed Approach

To validate the proposed approach, the application to a real case study is now illustrated. The subject of the analysis is the microtopography of the surface of a steel embossing roller for texturing a thin polymeric film. The embossing pattern, whose functional role requires a certain degree of geometric uniformity, has been

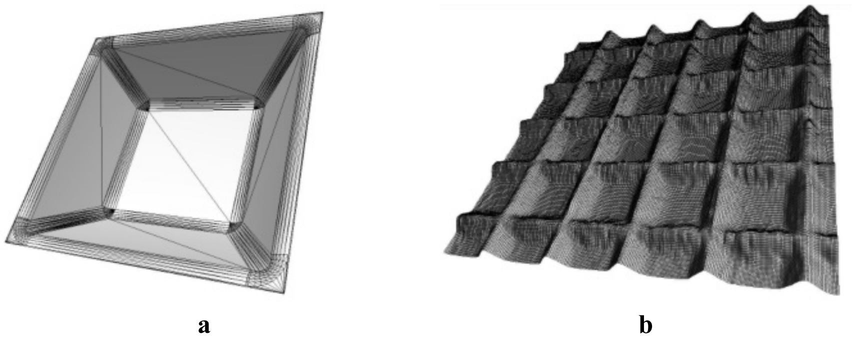


Figure 5.16 Surface topography of embossing roller: **a** template geometry (STL), and **b** measured surface (approximate projected area 1,473 μm × 1,531 μm). The template is not visualized at the same scale as the scene

engraved on the roller. The measured surface topography (a portion of the roller surface) is shown in Figure 5.16b; it was obtained through raster scanning with a 3D profilometer equipped with a noncontact laser sensor (conoscopic holography; Optimet Conoscan 1000), consists of 280 × 290 scanned points with 5.28-μm spacing in *x*, *y* and *z* resolution less than 0.1 μm, and covers an approximate area of 1,473 μm × 1,531 μm. To apply the proposed approach, the nominal geometry shown in Figure 5.16a was defined to represent the characteristic shape of a pattern element, and was modeled as an STL triangulated geometry by means of a CAD system.

5.6.1 Feature Identification and Extraction

Before the MRPT was applied, the template was slightly extended at the sides with flat surfaces so that the circular patterns of the MRPT would cover the entire feature definition area, as shown in Figure 5.17a.

The combination of slope-based segmentation and region edge detection illustrated earlier was then selected as the preprocessing step for the virtual raster scanned template and the candidate regions. Figure 5.17b shows the preprocessed template; slight asymmetries in the final edge map are due to discretization errors in raster scanning, segmentation, and edge detection. Figure 5.17c shows the MRPT curve computed on the preprocessed template. Finally, Figure 5.17d shows the preprocessed scene; notice how the irregularities on the measured surface result in a slightly more irregular edge map.

At the end of the main scanning loop, a match score surface is obtained, as shown in Figure 5.18a. Local maxima represent the locations of the highest-scoring matches of the feature-identification process. The match score surface is complex, as expected given the many self-similar regions generated by the preprocessing transform.

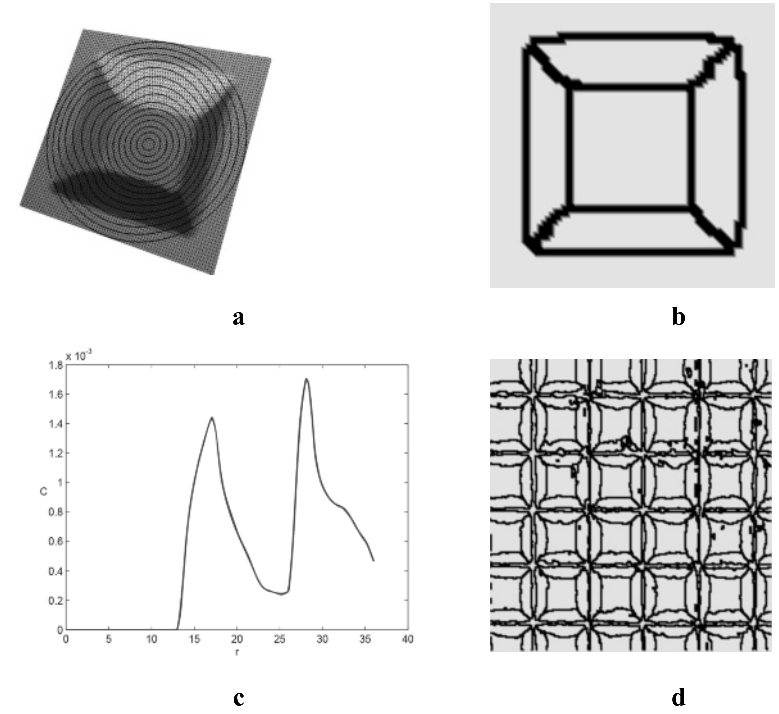


Figure 5.17 Preprocessing steps: **a** template surface extended for the MRPT and subjected to virtual raster scanning, with overlaid circular paths, **b** template binary map after preprocessing (slope-based segmentation and edge detection on the segment map), **c** MRPT curve for the preprocessed template, and **d** scene binary map after preprocessing (slope-based segmentation and edge detection on the segment map)

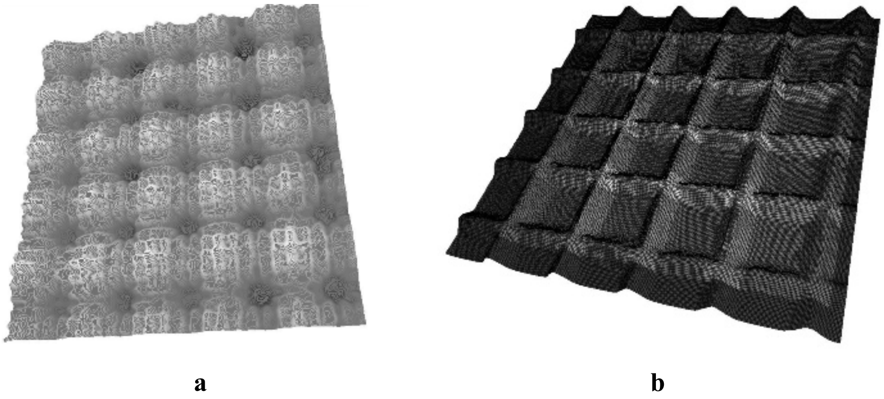


Figure 5.18 Feature identification results: **a** match score surface, and **b** corresponding locations of the regions defined by the 16 local maxima identified

While some maxima are strong where we expected them to be, and thus properly locate some of the features, others are weaker and are as high as other peaks referring to features that are partially located outside the acquired area. To recognize the 16 features properly, rules were added to the extraction of local maxima so that they could not be too close to each other and/or too close to the border of the acquired region. This is basically imposing a constraint on feature localization of the feature recognition process. Given such constraints, the 16 highest matching scores correspond to the 16 regions highlighted in Figure 5.18b.

5.6.2 Feature Alignment and Form Error Assessment

As for the previous example, no datums can be identified to do the alignment between the template and the measured feature; hence, global alignment is pursued. The original template geometry (Figure 5.16a) is used as the additional surface of the extended template (Figure 5.17a) would negatively affect the alignment results.

Moreover, differently from the previous example, no coarse alignment needs to be done as the measured features are not too-differently oriented with respect to the template. Fine alignment for one of the extracted candidate regions produces the results shown in Figure 5.19b; the unit volume errors are shown in Figure 5.19c. Once again, the meaningfulness of the computed form error depends on the alignment procedure, which in turn depends on the form error assessment goals to be pursued. In this case the result can be used to assess global deviations from the nominal shape; if form error is computed individually for all the 16 recognized candidates, it can be used also to assess pattern regularity.

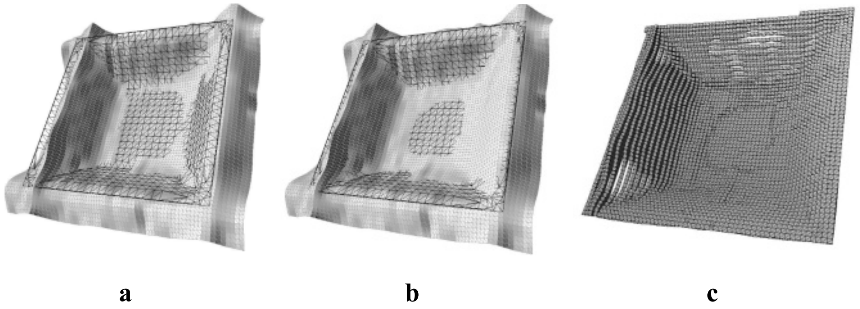


Figure 5.19 Fine alignment results with one of the extracted candidates: **a** template and candidate before fine alignment, **b** fine-aligned geometries, and **c** unit volume errors

5.7 Conclusions

In this work a novel algorithmic approach was proposed for form error assessment for microtopography features such as those that are commonly found on semiconductor products, MEMS and microcomponents in general, as well as for microtopographic patterns that can be micromanufactured on the surfaces of standard-sized parts. The surface topography is assumed to be available as a height map, *i.e.*, 2.5D geometry, as is typically acquired through raster scanning by means of a 3D profilometer or a 3D scanning microscope. The approach comprises several steps. First, the feature of interest is localized and identified within the measured region by means of a modified version of the ring projection transform, adapted to operate on topography data. Then it is extracted and aligned with a CAD-modeled reference (representing the nominal geometry): alignment comprises two steps (coarse alignment, through an exhaustive search over discrete angular positions, and fine alignment, with ICP). Finally, the form error is computed as the volume enclosed between the two aligned geometries: computation also takes into account discretization induced by raster scanning. The approach was illustrated and validated first through its application to an artificially generated case study, then to a real-life case of industrial relevance. The results show that the approach is indeed effective in identifying the features of interests, aligning them to their nominal references (templates), and obtaining a measure of the form error. However, some open issues still remain.

5.7.1 Issues Related to Feature Identification

The performance of feature identification needs to be further investigated in the presence of occlusion and different types/amounts of disturbances. Also, the discriminating power of the identification approach with respect to different geometric elements must be investigated in more detail.

The recognition of false positives in feature identification (*i.e.*, candidate regions erroneously recognized as good matches) must be investigated. This is needed as the outcome of the subsequent form error assessment depends on it. False positives may arise since identification is based on retrieving local maxima in the match score surface. In the industrial case example this was accomplished by establishing rules on the placement of the maxima (see Figure 5.18); more general approaches may be searched for.

Feature recognition is based on similarity assessment (through the MRPT); the form error is again a measure of similarity. In the presence of large form errors, feature identification may not succeed in the first place, thus limiting the application of the approach to those cases where the form error is not so large that identification is compromised. Further investigation is required to identify this boundary condition.

5.7.2 Issues in Feature Alignment and Form Error Assessment

The performance of the proposed approach for coarse alignment depends on the selection of the angular discretization parameter, which must be carefully selected depending on the geometric properties of the feature to be aligned. This dependency is not that important for features with strong asymmetry traits, but becomes relevant for quasi-symmetric features, where a too-coarse discretization may lead to unsatisfactory alignment. Algorithmic determination of suitable angular discretization should be pursued.

Types of form error assessment requiring reference datums to be registered first require changes to the alignment procedure, as a global alignment solution would not be adequate. In such situations, alignment should take identify the datums on the candidate (the datum should be explicitly defined on the template) and then overlay them; then, the remaining alignment transforms should be constrained to move the features without breaking datum alignment.

Final mention should be reserved for the problem of *measurement uncertainty*, which was not dealt with in this chapter, but is, however, of fundamental importance for any quality inspection procedure. Owing to their rapid evolution, most profilometers and microscopes dedicated to the analysis of 3D surface microtopography lack proper measurement traceability, and their accuracy (trueness and precision) cannot be easily determined. International standards are beginning to be available for specific classes of instruments (ISO 25178-601:2010, ISO 25178-602:2010 and ISO 25178-701:2010) and others are being developed (ISO/DIS 25178-603, ISO/DIS 25178-604, ISO/DIS 25178-605); however, some more time is necessary before such efforts reach maturity and gain widespread acceptance.

References

ISO 25178-601:2010 Geometrical product specifications (GPS) – Surface texture: Areal – Part 601: Nominal characteristics of contact (stylus) instruments, International Organization for Standardization, Geneva

ISO 25178-602:2010 Geometrical product specifications (GPS) – Surface texture: Areal – Part 602: Nominal characteristics of non-contact (confocal chromatic probe) instruments, International Organization for Standardization, Geneva

ISO 25178-701:2010 Geometrical product specifications (GPS) – Surface texture: Areal – Part 701: Calibration and measurement standards for contact (stylus) instruments, International Organization for Standardization, Geneva

Audette MA, Ferrie FP, Peters TM (2000) An algorithmic overview of surface registration techniques for medical imaging. Med Image Anal 4(3):201–217

Barré F, Lopez J (2000) Watershed lines and catchment basins: a new 3D-motif method. Int J Mach Tools Manuf 40: 1171–1184

Besl PJ, McKay ND (1992) A method for registration of 3-D shapes IEEE Trans Pattern Anal Mach Intell 14(2):239–256

Brown LG (1992) A survey of image registration techniques. Comput Surv 24(4):325–376

Chen GY, Bui TD, Krzyzak A (2005) Rotation invariant pattern recognition using ridgelets, wavelet cycle-spinning and Fourier features. Pattern Recognit 38(12):2314–2322

Dong WP, Sullivan PJ, Stout KJ (1994) Comprehensive study of parameters for characterising 3D surface topography III: parameters for characterising amplitude and some functional properties, & IV: parameters for characterising spatial and hybrid properties. Wear 178:29–60

Flusser J (2006) Moment invariants in image analysis. Proc World Acad Sci Eng Technol 11:196–201

Funkhouser T, Kazhdan M, Min P et al (2005) Shape-based retrieval and analysis of 3D models. Commun ACM 48(6):58–64

Horn BKP (1987) Closed-form solution of absolute orientation using unit quaternions. J Opt Soc Am A 4:629–642

Jiang XQ, Blunt L, Stout KJ (2004) Third generation wavelet for the extraction of morphological features from micro and nano scalar surfaces. Wear 257:1235–1240

Johnson AE, Hebert M (1999) Using spin images for efficient object recognition in cluttered 3D scenes. IEEE Trans Pattern Anal Mach Intell 21(5):433–449

Kazhdan M, Chazelle B, Dobkin D et al (2004) A reflective symmetry descriptor for 3D models. Algorithmica 38(1):201–225

Kortgen M, Park G-J, Novotni M et al (2003) 3D shape matching with 3D shape contexts. In: Proceedings of the 7th central European seminar on computer graphics, Budmerice, Slovakia

Lin YH, Chen CH (2008) Template matching using the parametric template vector with translation, rotation and scale invariance. Pattern Recognit 41:2413–2421

Lonardo PM, Trumpold H, De Chiffre L (1996) Progress in 3D surface microtopography characterization. Ann CIRP 42(2):589–598

Loncaric S (1998) A survey of shape analysis techniques. Pattern Recognit 31(8):983–1001

Lucchese L, Mitra SK (2001) Color image segmentation: a state-of-the-art survey. Image processing, vision, and pattern recognition, Indian National Science Academy, New Delhi

Mokhtarian F, Khalili N, Yuen P (2001) Curvature computation on free-form 3-D meshes at multiple scales. Comput Vis Image Understand 83(2):118–139

Pal NR, Pal SK (1993) A review on image segmentation techniques. Pattern Recognit 26(9):1277–1294

Rusinkiewicz S, Levoy M (2001) Efficient variants of the ICP algorithm. In: Third international conference on 3-D digital imaging and modeling (3DIM '01), p 145

Scott PJ (1998) Foundation of topological characterisation of surface texture. Int J Mach Tools Manuf 38(5–6): 556–559

Senin N, Ziliotti M, Groppetti R (2007) Three-dimensional surface topography segmentation through clustering. Wear 262(3–4):395–410

Tangelder JWH, Veltkamp RC (2004) A survey of content based 3D shape retrieval methods. In: Proceedings shape modeling international, pp 145–156

Zeng W, Jiang X, Scott P (2005) Complex ridgelets for the extraction of morphological features on engineering surfaces. J Phys Conf Ser 13:246–249

Zhang Z (1994) Iterative point matching for registration of free-form curves and surfaces. Int J Comput Vis 13:119–152

Zitovà B, Flusser J (2003) Image registration methods: a survey. Image Vis Comput 21:97721000

Standards under Development

ISO/CD 25178-604 Geometrical product specifications (GPS) – surface texture: areal – part 604: Nominal characteristics of non-contact (coherence scanning interferometry) instruments. International Organization for Standardization, Geneva

ISO/CD 25178-605 Geometrical product specifications (GPS) – surface texture: areal – part 605: nominal characteristics of non-contact (point autofocusing) instruments. International Organization for Standardization, Geneva

ISO/DIS 25178-603 Geometrical product specifications (GPS) – surface texture: areal – part 603: nominal characteristics of non-contact (phase-shifting interferometric microscopy) instruments. International Organization for Standardization

Chapter 6
Geometric Tolerance Evaluation Using Combined Vision – Contact Techniques and Other Data Fusion Approaches

Gianni Campatelli

Abstract The development of ever-pressing requirements for geometric tolerances has produced two main measuring needs: to obtain the geometric values of industrial products with higher precision and to obtain these values in a reduced time span. In order to accomplish these objectives, one of the most investigated and applied approaches is the use of multiple sensors on a traditional coordinate measuring machine (CMM). The resulting machine is usually referred to as a hybrid CMM and it is able to combine the data from optical and contact sensors in order to produce the measurement of a specific object with higher precision and in less time with respect to the traditional CMM approach. This chapter briefly explains the hybrid CMM characteristics and the working principles of the most used sensors. Then the method for the treatment of the data acquired by the multiple sensors is presented, starting from the basic problem of data registration to the algorithm to integrate and fuse the optical and touch probe data.

6.1 Introduction to Hybrid Coordinate Measuring Machine Systems

Hybrid coordinate measuring machines (CMMs) use more than one sensor in order to obtain information faster regarding the test piece geometry. The main difference with respect to traditional CMMs is the use of optical sensors to acquire geometric data. The optical acquisition source can be mounted on the arm of the

G. Campatelli
Dipartimento di Meccanica e Tecnologie Industriali, Università di Firenze,
Via di S. Marta 3, 50139 Florence, Italy,
e-mail: gianni.campatelli@unifi.it

CMM (*i.e.*, camera) or can be external to the CMM structure depending on the technology and acquisition strategy chosen. The use of optical sensors does not exclude the successive use of a contact probe in order to collect a high-resolution dataset in a time-consuming manner.

The classic use of a hybrid CMM is for medium-sized to small products; so the traditional implementation of such systems is a medium-sized CMM with a workspace volume ranging from about 0.5 to 1 m^3. Products with larger dimensions are usually measured by using a pure optical system or, at least, with a system where the trigger probe position is measured using optical sensors (*i.e.*, laser triangulation or optical orientation of the touch probe, a solution often used for naval and car body measurements); however the aim of this chapter is the presentation of a hybrid CMM where optical and touch probe approaches are both used in order to obtain a measurement, so these large-scale applications will not be treated in this chapter.

The optical systems with which a hybrid CMM is equipped are usually based on techniques that allow a high rate of point acquisition, such as laser triangulation, structured light projection, and interferometric systems. Other optical sensor techniques characterized by higher precisions but lower acquisition rates, such as laser interferometry and conoscopy, are not used, at least in an industrial field of application, in a hybrid CMM.

The hybrid CMM idea is spreading (Chen and Lin 1997) because the actual optical probes can provide good precision in measuring well-defined geometries, theoretically as good as the precision obtained with contact probes in certain specific and controlled conditions (Chen and Lin 1991). Within the category of hybrid CMMs there are two main groups:

1. The imaging probe CMM group, which encompasses, as defined by the ISO/DIN 10360-7/2008 standard, all the CMMs equipped with a vision probe system. A subcategory of these is the optical CMM (OCMM), where the sensor is a camera.
2. The group of CMMs with standstill optical systems, which include all the approaches that have the optical sensors not rigidly connected to the CMM (*i.e.*, mounted on the arm) but used externally to provide information for the successive use of a touch probe. An example of an application is the use of an independent structured light system supporting the CMM.

It is, however, necessary to point out that the use of optical probes introduces some specific error sources, such as the type of illumination, the optical lens image distortion, the distance from the part, and the optical characteristics of the measured part. For these reasons the real resolution of a CMM equipped with an optical sensor is typically much lower than that of a CMM with a touch probe. By "resolution", we refer to the smallest spatial distance that a measuring device can distinguish.

A CMM could have a resolution as fine as 0.5 µm, while an optical system mounted on a CMM typically has a resolution of the order of 10 µm (Shen *et al.* 2000). Laser scanners have a very high data acquisition rate (up to 2,500 data points per second), and good resolution, of the order of 10 µm (Chan *et al.* 1997). Vision systems have lower resolutions (about 20–30 µm); however, they can acquire thousands of data points simultaneously over a large spatial range (El-Hakim and Pizzi 1993) without moving the optical head.

Optical methods, such as interferometry, stereo vision, structured light projection, and shape from focus/defocus, have long received extensive attention. Unlike classic touch probe CMMs, measurement systems developed from optical principles are noncontact in nature. This characteristic is responsible for the absence of wear and deformation during the measuring process that sometimes can happen when contact systems are used with low hardness and low stiffness measurands.

A typical application of an OCMM is in the field of reverse modeling, where a large number of data points must be acquired and the touch probe approach would be too slow. However, if high precision is required for reverse modeling, the use of a touch probe is mandatory as it is if there is a reduced number of points. For this reason, an OCMM system usually integrates both contact and noncontact acquisition systems.

Nashman *et al.* (Nashman 1993; Nashman *et al.* 1996) were among the first to develop integrated vision–touch probe systems, which emphasize the integration of a CMM, a 3D analog touch probe, a video camera, and a laser triangulation probe, all managed by a unique system controller. Their work shows that it is possible to create multisensor cooperative interaction of a vision and a touch probe system which provides sensory feedback to a CMM for a measuring task.

Examples of the application of OCMMs can be found in many companies. In Figure 6.1 a CMM with an optical sensor mounted and a detailed picture of an optical sensor (laser line scanner) are illustrated.

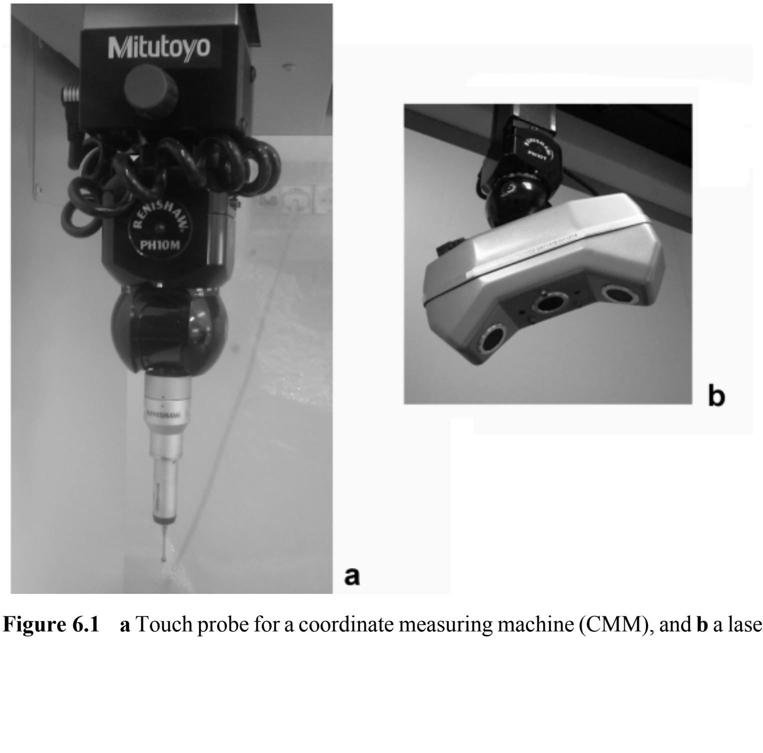


Figure 6.1 **a** Touch probe for a coordinate measuring machine (CMM), and **b** a laser line scanner

6.1.1 Brief Description of Optical Measurement Systems

Currently, the standard systems for surface digitization are the digital light processing (DLP) projector and laser line scanners.

The DLP projector is one of the most used solutions for surface digitization thanks to the flexibility and high precision that can be obtained with this system, especially if compared with the older technology of LCD projectors. The DLP projector is based on the technology of a digital micromirror device (DMD) (Hornbeck 1998). The DMD is an array of micromirrors fabricated onto a memory chip and directly controlled by it. Each micromirror is able to turn on or off a single pixel of light controlled by the input digital signal. The low inertia of the system allows the mirrors to be switched at very high speed. The switching creates an image that is reflected from the surface of the DMD. This allows the DLP projector to create easily, quickly, and without interpolation error a large variety of digital patterns with high resolution and precision. The most used structured light patterns range from a single light beam to a grid light, a single light stripe, and multiple light stripes. The DLP projector is associated with a CCD camera that grabs images of the object. Through the sampling of the 2D image coordinates of the surface points illuminated by the light patterns, the 3D coordinates of these surface points can be calculated using a triangulation algorithm. The principle of projection light reconstruction is that projecting a narrow band of light onto a three-dimensionally shaped surface produces a line of illumination that appears distorted from perspectives other than the perspective of the projector, and can be used for an exact geometric reconstruction of the surface shape. For this reason the camera is rotated by a certain angle with respect to the axis of the projector. The modern systems also use multiple cameras to increase the precision of acquisition by averaging the coordinates of the point acquired. Usually patterns with many stripes at once, or with

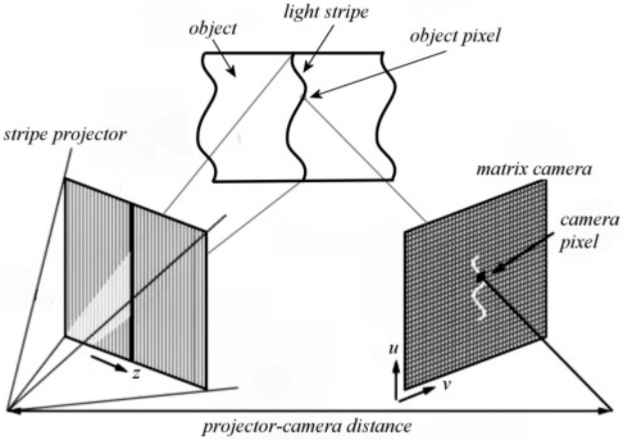


Figure 6.2 Working scheme for a LCD structured light approach

arbitrary fringes, are used in order to have a fast acquisition of the whole surface. This initial phase, not considering calibration where the reference system of the projector and the CMM are to be roughly aligned, is very fast. A simple scheme of the process is shown in Figure 6.2.

Another approach used with DLP projectors is to use a light stripe scanning method, which sequentially projects a single light stripe, or more frequently a pattern of stripes in order to save time, onto the object surface with the aim of exploring the whole product. The advantage of this approach is that there are no moving parts in the acquisition system and all the scanning is controlled by software. Moving mechanisms often introduce errors and usually complicate the whole system (Yang *et al.* 1995).

The other most used type of sensor is the laser scanning head. This system is based on the projection of a laser line (sometimes also only a dot) onto a surface and the acquisition of the resulting image using a CCD camera. The camera has to be positioned in a known position with respect to the laser source in order to calculate the position of the surface points. Depending on how far away the laser light collided with the surface, the laser line (or dot) appears at different places in the camera's field of view. This technique is called triangulation because the laser line, the camera, and the laser source form a triangle. The triangulation evaluates the position of each point of the laser line because the distance and the angle between the camera and the laser source are known. Given the position of the pixels on the CCD camera that see the projected line, it is possible to reconstruct the position of the whole curve where the line is projected. As for the DLP approach, this method is very fast but it is surface-dependent. For reflective surfaces the projected light may not be seen, and for translucent surfaces the signal could be very noisy; only for Lambertian surfaces the method provides very good results (Forest and Salvi 2002; Zussman *et al.* 1994). Often the products to be measured with this approach are painted with a matt white paint. An example of the behavior of laser light on different surfaces is shown in Figure 6.3.

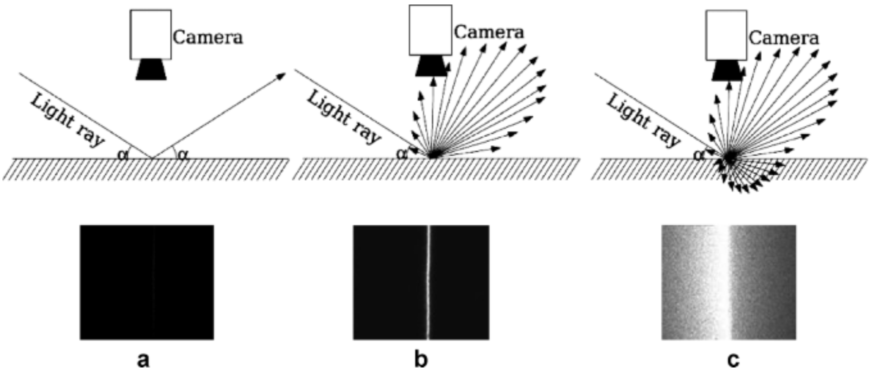


Figure 6.3 Effect of the surface type on the laser light reflection: **a** specular surface, **b** Lambertian surface, and **c** translucent surface

Both of these systems use CCD cameras for the image acquisition. When this solution is adopted, it is necessary to evaluate and consider lens distortion. The lens distortion is a characteristic feature of all real optical systems. The lens distortion can be evaluated using different models (Weng *et al.* 1992); the basic distortion components are defined as radial distortion, decentering distortion, and thin prism distortion, but more complex models are also available. For the lenses that are suitable for machine vision, the radial distortion is usually much more significant than the other two types of distortion. Therefore, usually only this term is included in the lens distortion model. This model can be evaluated and the vision of the CCD corrected by acquiring and processing a known image, usually a chessboard. Similar corrective actions are developed for the lens and DLP system of the projector (Chen *et al.* 2009; Zexiao *et al.* 2007).

6.2 Starting Problem: Precise Data Registration

The starting problem for the integration of the data from multiple sources is given by the need for data registration in order to obtain a spatial "closeness" among the same points acquired by two, or more, measuring systems. The techniques used to fuse together close points acquired also at different resolutions will be treated in the following section, but each method applied needs a first preprocessing step of registration in order to provide reliable results.

Each measuring instrument provides a cloud of points, sometimes obtained using a regular grid but often with an irregular pattern, that have to be related to, and later integrated with, the output of the other acquisition instruments. The characteristics of the optical spatial data, which are usually unordered, with fixed density and independent of the surface curvature, make this integration nontrivial.

The main problem is that the different acquiring instruments nearly always use different spatial references (origin and orientation of the axes). Even if the part is measured on a single measuring platform carrying both a touch and a laser probe, independent calibration is usually needed for each probe, resulting in a misalignment.

The misalignment among the datasets is usually small because, during the calibration of the tools, it is normal practice to refer the acquiring devices to the same spatial coordinate system. However, also using this precaution, a small misalignment is always found among the many datasets. The misalignment is very important because the numerical algorithms to achieve an alignment are different if one starts from two or more clouds of points that are heavily misaligned or if one starts from clouds that are nearly aligned. The algorithms to be used in the first case are based on pattern recognition or on the evaluation of the moments of inertia of the cloud; these approaches are very fast but provide only a rough alignment of the clouds. In the second case, the one that is usually applied in the case of multisensor measuring machines such as the OCMM, the approaches are mainly based on the popular iterative closest point (ICP) algorithm.

The ICP algorithm proposed by Besl and McKay (1992) is a standard solution to the alignment problem. The ICP algorithm has three basic steps, as reported below considering $\mathcal{P}$ and $\mathcal{M}$ as two set of points to be aligned:

1. Pair each point of $\mathcal{P}$ to the closest point in $\mathcal{M}$.
2. Compute the motion that minimizes the mean square error (MSE) between the paired points.
3. Apply the optimal motion to $\mathcal{P}$ and update the MSE. Some iterations are needed to obtain a satisfactory solution.

The ICP algorithm is a very popular algorithm, so many variants have been proposed over the years. These variants can be classified according to how the algorithms:

- select subsets of $\mathcal{P}$ and $\mathcal{M}$;
- pair points;
- weight the pairs;
- reject some pairs;
- assign the error metric; and
- minimize the error metric.

Many modifications of the ICP algorithm have been proposed in order to improve the robustness of the algorithm. In the case of noisy clouds of points, the ICP algorithm sometimes provides faulty results. The source of the problem is that the original algorithm assumes outlier-free data and $\mathcal{P}$ being a subset of $\mathcal{M}$; this implies that each point of the first dataset has a corresponding point in the second dataset. Therefore, this approach requires $\mathcal{M}$ has a larger number of points than $\mathcal{P}$, as usually happens using optical and touch measuring machines.

Many attempts have been conducted in order to make the ICP algorithm more robust by rejecting wrong pairs. An example of this strategy can be found in the works by Pulli (1997) and Rousseeuw and Van Zomerman (1990) that introduced two algorithms called least median of square (LMedS) and least trimmed square (LTS). The general idea is to make the linear regression used to evaluate the least mean square insensitive to outliers.

Consider the expression

$$y_i = \beta_0 + \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_j x_{ij} + \sum_{i=1}^{n} e_i = \beta_0 + \beta_1 x_{11} + \ldots + \beta_m x_{nm} + e_1 + \ldots + e_n \qquad (6.1)$$

for $i = 1, \ldots, n$. Here y_{i} are the response variables, x_{ii} to x_{im} the explanatory variables, and β_j for $j = 0, \ldots, m$ the coefficients (the model parameters). In the classical regression theory the errors e_i are assumed to have a Gaussian distribution with zero mean. The standard least (mean) squares method computes the parameters β_j such that the sum of the squares of the residuals $r_i = f(\beta_0, \ldots, \beta_m)$ is minimal.

A simple but useful measure of robustness for the proposed variant to the ICP algorithm is the breakdown point, that is, the fraction of the dataset that can be contaminated with outliers before the algorithm crashes (wrong results). Both the

LTS and the LMedS have a breakdown value of 50%. On the basis of the study of the breakdown values of the algorithms, research into a new solution has started to consider with more care the statistical properties of the method adopted.

Among the many solutions developed over the years, it is worth mentioning the algorithm proposed by Chetverikov *et al.* (2005), which in the application to real registration problems has proved its robustness and usefulness. This is a good model to explain the "trimmed" approach of the ICP algorithm used by many other authors. The algorithm proposed is called the trimmed ICP (TrICP) algorithm. The advantages of the proposed TrICP algorithm are its robustness to noise and to poor initial estimation.

The variant with respect to the original ICP algorithm is mainly due to the introduction of an overlap function (ξ), the fraction of the points of the smaller set ($\mathcal{P}$) that have to be paired, which can be computed automatically by the algorithm with some iteration. The steps of the TrICP algorithm are as follows:

1. For each point of $\mathcal{P}$; find the closest point in $\mathcal{M}$ and compute the individual distances of each point to the paired one.
2. Sort the individual distances in increasing order, then select the ξ lowest values and calculate their sum (STS).
3. If the value of the MSE or number of iterations has reached the stopping value, stop the algorithm, otherwise update the value of STS and continue.
4. Compute for the selected pairs the optimal Euclidean motion that minimizes STS.
5. Transform $\mathcal{P}$ according to the motion and return to step 1.

6.3 Introduction to Serial Data Integration, Data Fusion, and the Hybrid Model

In the field of multisensor measurements two different approaches may be used in order to fully use the capabilities of an OCMM: serial data integration and data fusion. The first solution uses the optical sensors to acquire a great amount of data and to organize the data acquisition that follows with higher precision devices (usually a touch probe). The final measurement data used for the inspection of the geometric tolerances or acquisition of the geometry of the product for reverse modeling are only from the second, high-resolution, device. This is a serial approach where the data from the optical source are used only during the acquisition process and cannot be found in the final dataset. The optical device is used only to increase the speed and efficiency of the touch probe acquisition. In the case of geometric tolerance verification, the first system is used to localize the product to be tested and to execute subsequently the touch probe tolerance inspection. In the case of reverse modeling of a specimen, the first step has the objective to evaluate the general dimensions, the boundaries, and, if possible, the most relevant features of the specimen in order to optimize the following touch probe measurements.

On the other hand, data fusion has the aim of creating a complementary acquisition of surface data by multiple sensors. The difference from the serial data integration is that the resultant dataset is a fusion of the data acquired both by the low-resolution and by the high-resolution sources. Actually there are only a few applications of data fusion in the metrology field, while this approach is extensively used in other fields of research such as the nondestructive testing, geospatial applications, weather forecasting, and airplane control.

Data fusion is generally defined as the use of techniques that combine data from multiple sources and gather that information in order to make inferences, which will be more efficient and potentially more accurate than if they were made by means of a single source. Data fusion is a low-level fusion process. Fusion processes are often categorized as low, intermediate, or high, depending on the processing stage at which fusion takes place. Low-level fusion combines several sources of raw data to produce new raw data. The expectation is that fused data are more informative and synthetic than the original inputs.

From this definition it is possible to evaluate the differences among serial data integration and data fusion. Serial data integration has the main advantages of simplicity of application, the deterministic choice of the precision of the result (given by the second measuring tool to be used), and the possibility of creating general-purpose systems to analyze a large variety of products. On the other hand, the use of serial data integration has a lower efficiency than the use of data fusion because some of the acquired data are discarded during the process and the geometric information has to be reacquired with a slower high-precision system. Data fusion, however, has some major problems: the main one is the inability to determine *a priori* the precision of the information for a specific measurement. The limit of the precision that can be reached, for classical application of data fusion that does not acquire replicas of the measurement, is always a fraction of the precision of the better performing device used. Moreover, the data fusion cannot be used for general-purpose measurement but, at the current state of the art, the strategy and the following characteristic parameters have to be evaluated for each specific case. The most used techniques for data fusion are the neural network, the fuzzy logic, and, specifically for the measurement field, the Bayesian approaches. These limitations, and the greater simplicity of the serial data acquisition approaches in terms of practical applications, have undermined the development of "pure" data fusion approaches in the field of industrial measurement. Nowadays, many OCMMs work with a serial data integration approach, while data fusion is the preferred strategy in other fields, such as geographical and environmental measurements. However, hybrid models have been proposed in order to implement some of the advantages of both approaches. These hybrid models provide a resultant dataset that takes the information mainly from the touch probes but partly also from the optical sensors. These approaches are based on the implementation of advanced feature recognition strategies based on the output of the optical sensors. Currently they are also providing promising results, albeit limited only to applications with a small number of specific features. In the following sections more detailed descriptions of some exemplary implementations of the serial and hybrid data integration and data fusion approaches are reported.

6.4 Serial Data Integration Approaches

Serial data integration is used both for reverse modeling of free-form surfaces and for inspection of geometric tolerances. In the first case the optical sensors have the aim of acquiring the general dimensions and boundaries of the sample in order to automatically generate and optimize the path of the touch probe. In the second case the aim of the optical sensors is to locate a sample of known geometry inside the measuring range and execute a preset trigger probe tolerance inspection. In both cases the main advantage of a serial approach is the reduction of the time needed for the geometry/tolerance acquisition of the sample. In the traditional touch-probe-only approach there is a relevant manual activity that can be automated using a serial approach. The results in terms of performances are very interesting; some researchers (Chen and Lin 1997) reported a reduction in the time needed for the reverse modeling, maintaining the same precision of traditional approaches, of about 85–90%.

The general scheme of a serial data integration approach is always based on four simple steps:

1. acquisition of low-resolution data, usually using optical sensors;
2. processing of the data in order to obtain a low-resolution surface;
3. automatic definition of the strategy for acquisition of the high-resolution data; and
4. creation of a reconstructed surface based on the high-resolution dataset.

In order to explain in detail the serial data integration, two different approaches will be presented. The first approach was proposed by Chen and Lin (1997) and the second approach was proposed by Carbone *et al.* (2001). The two approaches are just two examples of the vivid creativity that one can find in the field of serial data integration. These two approaches were chosen as representative because they use very different solutions, for the acquisition of the low-resolution data (one uses stereo vision, while the other uses a structured light device), for the processing (one is model-free and uses a triangulation of the form, while the other uses Bézier curve representation), and for the touch acquisition tool path generation (planar in one case, 3D in the other). Most of the later proposed methods are a sort of permutation of the different strategies for the various steps presented in these two cases.

6.4.1 Serial Data Integration: Vision-aided Reverse Engineering Approach

The vision-aided reverse engineering approach (VAREA) was initially proposed by Chan *et al.* and later modified by many other authors (Chan *et al.* 1997). The system presented is constituted by a simple CCD camera mounted on the CMM arm and a trigger touch probe. The limitations of this approach are that the reconstructed surface must satisfy two conditions:

1. The measuring free-form surface has continuous derivatives at least up to the second order.
2. The measuring surface is such that all its normals are within the probe's local accessibility cone (Spitz *et al.* 1999).

The first condition is essential for the edge detection algorithm to work properly, the second is necessary for the following digitization by a touch probe. The approach is constituted by two main steps:

1. the initial vision-driven surface triangulation process;
2. the adaptive model-based digitizing process.

A detailed scheme of the approach as presented by Chan *et al.* is shown in Figure 6.4.

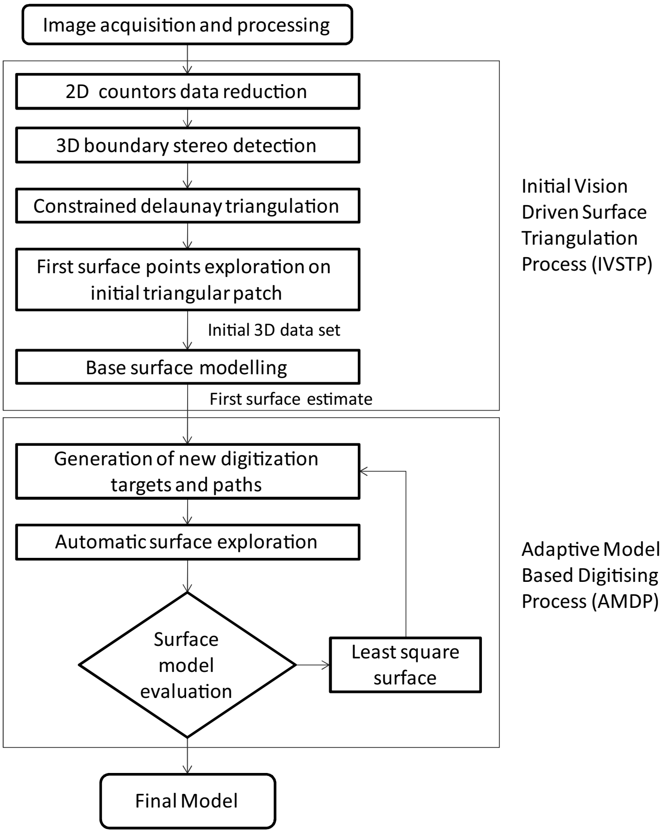


Image acquisition and processing
2D countors data reduction
3D boundary stereo detection
Constrained delaunay triangulation
First surface points exploration on initial triangular patch
Initial 3D data set
Base surface modelling
First surface estimate
Initial Vision Driven Surface Triangulation Process (IVSTP)
Generation of new digitization targets and paths
Automatic surface exploration
Surface model evaluation
Least square surface
Adaptive Model Based Digitising Process (AMDP)
Final Model

Figure 6.4 The vision-aided reverse engineering approach

In the first step, a vision system is used to detect 3D surface boundaries by using a 3D stereo detection method based on multiple image analysis (stereo vision). Obviously, before image processing the images are filtered in order to reduce noise to an acceptable level. Then a classical Laplacian edge detection algorithm is used to extract the surface boundaries effectively. Later the boundaries are reduced to a number of characteristic points using a data reduction strategy in order to create a first dataset that can be processed with a constrained Delaunay triangulation algorithm.

The geometric features extracted from the edge detection algorithm applied to the acquired image normally contain many redundant points which are not required for the initial surface triangulation. To achieve a reduction in terms of the number of points, an algorithm is used that computes the relative angle between two consecutive points that have been obtained. Only when the angle between the approximating curve of the sequence of points until the selected point and the segment that links this point to the following point is greater than a specific threshold, usually 20° is used, the point is considered in the dataset. This algorithm was developed by Fujimoto and Kadya (1993). For each point chosen the vertical coordinate is evaluated by using stereo-vision techniques. This approach provides only approximate values (tolerance of 0.1–0.3 mm), but it is enough for the rough model of the product. The next step is the development of a Delaunay triangulation based on the boundary points chosen. Other points internal to the boundary are chosen automatically in order to reduce the deformation of the triangles using the Delwall algorithm. This is a very fast step because this first model is created on the basis of only a few points. For all the vertices of the internal points of this first triangulation the vertical coordinate is acquired, either using again

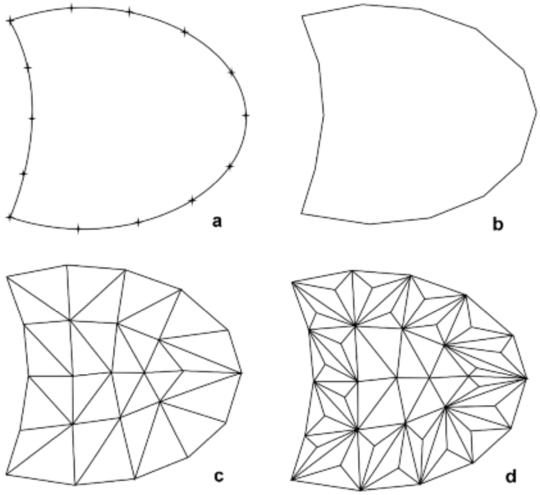


Figure 6.5 Digitizing of a ceramic plate: **a** boundary points, **b** boundary schematization, **c** first triangulation, and **d** refined triangular model

stereo-vision techniques or as a mean of the vertical coordinates of the connected boundary points.

Thanks to the triangulation of the rough surface, it is possible to create a collision-free path for the touch probe with a high safety margin. The first points to be accurately digitized are the boundary points and the chosen internal points. The main idea of the approach is to digitize sequentially the model using the triangulation as a basis for the choice of the points to be acquired by the touch probe. The surface triangulation refinement process explores the model-based midpoint estimates on the existing triangular patch and continues to refine the triangular patches, creating new midpoints for the triangular patches whose deviation of acquired values from the model values exceeds the specified digitizing accuracy. The adaptive model-based digitizing process continues with the automatic data exploration process until the digitizing accuracy reaches the user's specified deviation tolerance. With this approach it is possible to obtain automatically a triangular model of the product that satisfies the user's specified digitizing accuracy for every shape.

An example of the first step in the digitizing of a ceramic plate is shown in Figure 6.5.

6.4.2 Serial Data Integration: Serial Bandwidth

The second approach is intrinsically different from the VAREA because in this case the data acquired with both vision and a touch probe are organized in order to create a CAD model based on Bézier curves. This is a more complex representation with respect to the triangulation approach and it allows one to reconstruct also surfaces that are not always visible from a single direction like in the first case. This approach is also a serial one, where the points obtained with the optical probe are finally discarded and the object model is based only on the higher-accuracy touch probe data; however, the information from the optical device is considered to create a stopping condition and to define if the model has to be refined during the acquisition steps.

Briefly, the method can be summarized in the following steps. The first step is represented by the optical digitization of the object using many clouds of points. The clouds are merged together using the "bandwidth approach", which allows for an easy reconstruction of the surface based on Bézier curves. Basically the cloud of points rough model is somehow sliced by many bands and each band constitutes the basis to create a curve. This first rough model is used to define the tool path for the touch probe depending on the surface curvature (Yau and Menq 1996). Then the process is iterated until the error between the actual curve and the one computed in the previous step is under a user-defined threshold. Usually one or two iterations are enough for most surfaces, while a higher number of iterations could be needed for complex shapes. The scheme of the approach, as presented by Carbone *et al.* (2001), is shown in Figure 6.6.

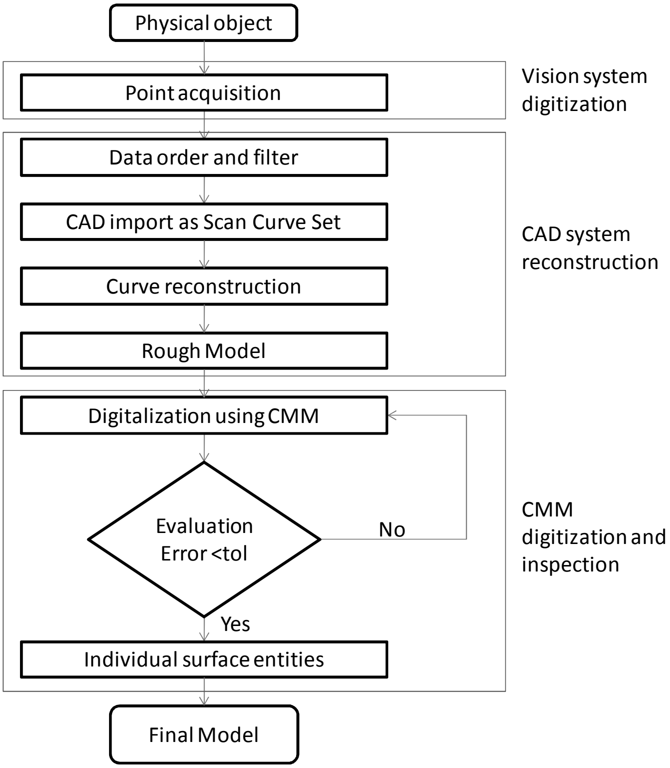


Figure 6.6 The serial bandwidth approach

The peculiar characteristic of the serial bandwidth approach is to identify, during the optical data acquisition, subsets of data points that are later simplified into CAD curves (typically, Bézier curves). Such subsets are actually regions of contiguous points, referred to as "bands", hence the name of the approach. Pairs of contiguous Bézier curves are then transformed into surface patches, and are used as a geometric reference to be refined by the subsequent high-resolution data acquisition.

In summary, while in the previous approach the final CAD model was built starting from the smaller dataset acquired by the touch probe, in this approach the CAD geometry is reconstructed by first creating a continuous geometric representation based on the points obtained by the optical probe (Bézier curves and surface patches), and then refining it by means of the points from the touch probe.

The starting point is the acquisition of a number of clouds of points using the 3D vision system. The system used for the first acquisition is a DLP projector of structured light and a CCD camera. The projector is oriented along multiple directions in order to "see" globally all the surfaces of the product. In general, for simple geometries one or two acquisitions from different directions are enough, while for more complex products three to five acquisitions could be needed.

The data obtained from the structured light processing projector are then filtered thanks to the choice of arbitrary bands in which to group the points. The bands are chosen starting from the definition of the reference plane, the ordering direction of the cloud of points, and the bandwidth between the arrays of points. The enveloping direction of the bands is perpendicular to the defined ordering direction and relates to the reference plane. The data are divided into bands and the points in the same band are ordered along the ordering direction. After that, the direction of adjacent curves is alternately changed.

An associated style curve is computed for each band using the Bézier representation. The style curve is obtained by filtering the huge number of points of the band with a threshold filtering angle similar to the one already used by the VAREA. From the filtered points the Bézier spline is automatically calculated, taking care to segment the curve where a discontinuity arises. The rough model is constituted by nearly parallel Bézier curves (the bands are parallel, while the 3D curves do not have this constraint, but they only have to be inside the associated band) that can be processed in order to create a rough surface schematization of the object. This aggregation into Bézier surfaces is usually carried out automatically by many commercial CAD software packages. The final accuracy of this model is usually of the order of 0.5 mm, enough to generate a collision-free touch probe tool path.

The touch probe acquisition process, which includes the definition of the measurement sequence, the number of measurement points, the number of probes, and their configuration, is planned using an algorithm that considers the curvature of the surface and defines the density of points to be acquired accordingly. Surfaces are then digitized using the touch probe, and the measurements are carried out until the stopping condition is found. This condition is given by the achievement of a specific user-defined value of the deviation between nominal points (those on the reconstructed Bézier surface) and actual points (those acquired by the CMM). If this value is greater than the threshold, a new set of points is acquired and the surfaces are iteratively reconstructed. In this case the first step of the iteration uses the surfaces generated by the structured light acquisition as references, while the following steps use the surface generated by the points acquired by the touch probe.

6.5 Geometric Data Integration Approaches

The geometric approaches are hybrid approaches that consider the information from the optical and the touch probes for the final digitization of the surfaces but do not apply a real fusion process. The hybrid approaches are used mainly for reverse modeling and scarcely for tolerance verification. This is because for the tolerance verification the information regarding the orientation of the sample, which a simpler serial approach can also provide, is deemed sufficient. In the hybrid approaches the information from the two sources is merged together in order to create a final representation of the product that is constituted mainly by

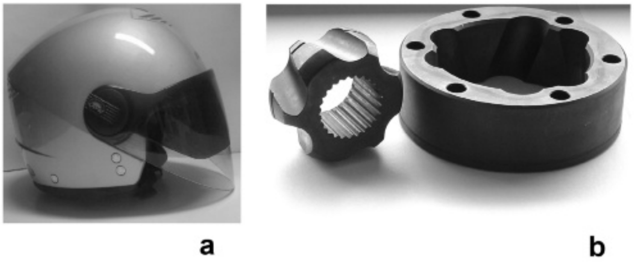


Figure 6.7 **a** Example of sculptured surface, and **b** a mechanical product

the points acquired by the touch probe but where the geometric features are evaluated thanks to the vision system. In this case higher-level information is extracted from the vision system with respect to the other approaches that also includes also geometric information. These approaches provide an advantage in terms of the time needed for the analysis of the sample with respect to the serial approaches when the surface can be simplified as an assembly of basic features, such as in the case of mechanical products; there is no advantage when the surface is a sculptured free form. In Figure 6.7 two examples of a sculptured surface and a mechanical product are shown, where the geometric approach has higher efficiency.

6.5.1 Geometric Approach: Geometric Reasoning

Many multisensor approaches have been developed on the basis of the geometric feature evaluation; the one that it is explained in detail in this chapter was developed at Columbus University by Shen *et al.* (2000). It was chosen because of the high degree of integration of the solution proposed and it could be considered, in its general definition, as a representative scheme for all similar approaches.

The multiple-sensor integrated system developed consists of three main subsystems:

1. a CMM with an active 3D vision system and a touch probe;
2. an information aggregation module for data fusion based on feature recognition; and
3. an inspection planning algorithm for the final touch probe surface digitization.

The vision system chosen is constituted by a DLP projector and a CCD camera. The DLP projector is one of the most used solutions for surface digitization thanks to the flexibility and high precision that this system can provide, especially when compared with the older technology of LCD projectors. The information aggregation module responsible for the data aggregation analyzes the acquired coordinate data from the 3D vision system to extract the surface geometric information of the object. This module performs three main tasks:

1. data segmentation and grouping;
2. surface identification and visualization; and
3. geometric reasoning and evaluation.

The resulting geometric abstractions form a preliminary description of the surface geometry and feature topology of the object that will be used for the part localization, sensor selection, and touch probe path selection.

Owing to surface features or occlusion that does not allow the acquisition process of the CCD camera, the cloud of points obtained is usually constituted by several patches of points, called view patches. Each view patch can be characterized using its shape as a criterion: planar, cylindrical, spherical, multifeature, or sculpture shape. For each patch the outer boundary and the inner boundaries can be defined using many algorithms (Daniels *et al.* 2007). The scheme of the approach, as proposed by Shen *et al.*, for the feature extraction is shown in Figure 6.8.

The first two steps have the objective of creating the view patches based on the local information of the acquired points. The neighbor algorithm creates a structure that evaluates which points are nearest to each point of the cloud. This information is then processed by the patch grouping step, which creates the view patches by grouping together all the points that have at least one point in the neighbor structure of another point of the patch. The view patches are defined as a continuous group of neighbor points. The surface features are then extracted from the view patches using surface fitting algorithms (Cernuschi-Frias 1984). If a view patch cannot be fitted with a single feature surface, the patch has to be segmented into smaller feature surfaces, or fitted with free-form parametric surfaces. A similar approach is used for the external and internal boundaries of the view patches in order to extract the curve feature from the cloud of points. The patch boundary extraction is carried out using a growing process (Huang and Menq 1999), while for the boundary segmentation the algorithm proposed by Lee and Menq (1995) is used.

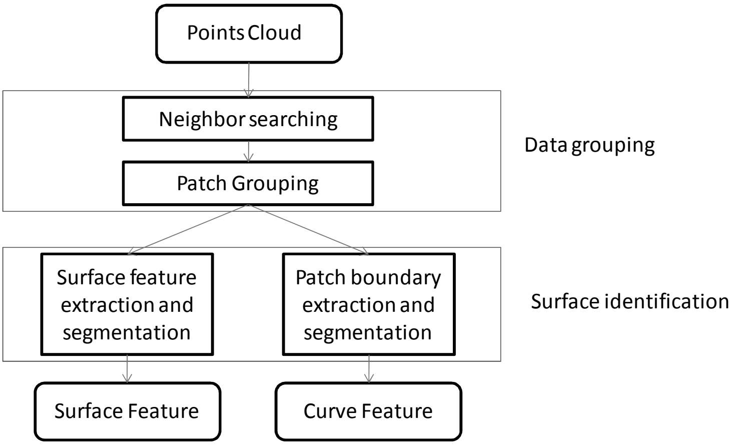


Figure 6.8 The feature extraction algorithm

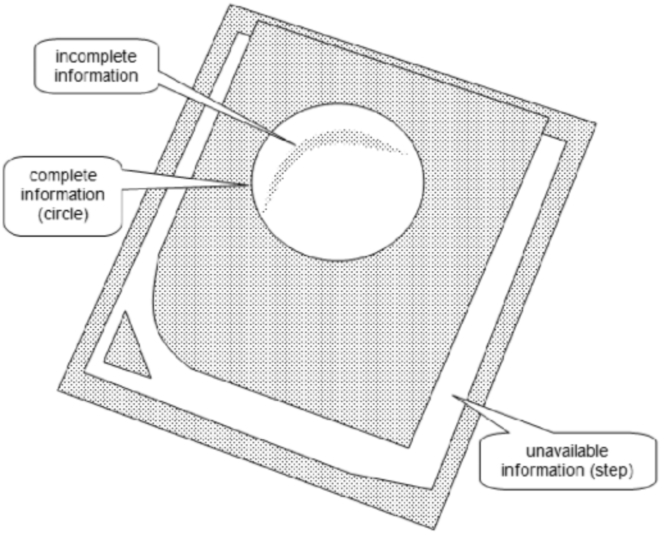


Figure 6.9 Example of complete and incomplete feature extraction

The extracted features are then grouped on the basis of the content of the information:

1. complete feature information;
2. incomplete feature information; and
3. unavailable feature information.

For complete feature information the CMM can be instructed to acquire the points needed to digitize the surface features with a touch probe that permits higher accuracy with respect to the optical system. If the feature information is incomplete or unavailable for the surface reconstruction, other optical acquisitions are needed in order to create a low-level geometric model useful for planning the touch probe acquisition. An example of complete and incomplete data is shown for a mechanical product in Figure 6.9.

The touch probe path is defined on the basis of the geometric feature: planar, circle, free form, *etc.* For each type of feature an optimal strategy can be planned depending on the precision that the measurement or the reverse modeling has the aim of achieving.

6.5.2 Geometric Approach: Self-organizing Map Feature Recognition

Similar to the geometric reasoning for the sequence of phases, but very different in terms of the algorithms used, is the self-organizing map (SOM) feature recognition approach proposed by Chan *et al.* (2001). The relevant aspect of the proposed approach is the use of an advanced neural network, a Kohonen SOM, for the surface segmentation and feature extraction. This approach is useful because its further implementation could be the basis for a real multisensor data fusion (Section 6.6) at least for simple features such as holes. The use of a neural network for feature recognition is widely spread in medical science, especially for the study of tumors, lesions, and other abnormalities in medical images. The general idea is to use only a CCD camera to acquire an image of the object and to extract the feature using the SOM. The patches are created by the study of the image color and intensity gradient.

A competitive learning network, the SOM, consists of *n* layers of 2D arrays of neurons; usually a 5 × 5 strategy is considered a good compromise between data processing time and precision. Each neuron is connected to its immediate neighbor on its own layer and to (*n* – 1) neurons on the layers immediately below and above its location. The input neurons are on the first layer and receive the input from the CCD image. For each iteration of the SOM, the strength of the neurons on each layers is measured and for each input dataset it is possible to evaluate which layer has the greatest strength; this is the winning layer and the pixel associated with the 5 × 5 dataset is associated with this layer. Each layer represents a possible patch, so the SOM has to be designed with a number of layers compatible with the characteristic of the geometry that has to be acquired; the number of layers has to be greater than the number of possible patches that the SOM has to recognize.

Given the patch of the object, it is possible to search for simple feature such as holes. The routine proposed is based on the evaluation of the external and internal boundaries of each patch. The internal boundaries are then processed in order to evaluate if the boundary is associated with a hole or another type of feature. On the basis of this analysis, a specific CMM acquisition process is defined automatically. The scheme for the hole recognition, as presented by Chan *et al.* (2001), is shown in Figure 6.10.

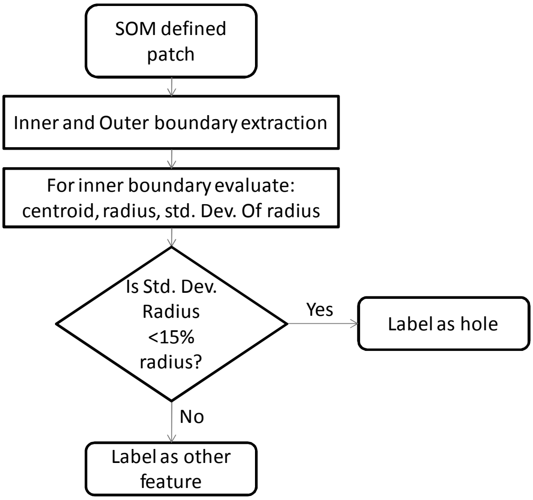

SOM defined patch
Inner and Outer boundary extraction
For inner boundary evaluate: centroid, radius, std. Dev. Of radius
Is Std. Dev. Radius <15% radius?
Yes
Label as hole
No
Label as other feature

Figure 6.10 The hole recognition algorithm. *SOM* self-organizing map

6.6 Data Fusion Approach

The integration approaches presented in the previous sections use the data from the lower-resolution system, usually the optical system, in order to optimize and automate the touch probe acquisition. The higher-resolution data of the touch probe simply overwrite the data already acquired with the other systems. Data fusion has the aim of integrating together the measurements from the single acquisition systems in order to obtain a measurement that has higher precision than each single dataset.

The data fusion approach applied to a hybrid CMM is still a new field of research. While multisensor data fusion is widely used in many fields such as weather forecasting, medical science, acoustics, and nondestructive testing, very few attempts have been made to integrate together the data from different sources in the mechanical measurement field. Huang *et al.* (2007) stated that "Multi sensor data fusion (MDF) is an emerging technology to fuse data from multiple sensors in order to make a more accurate estimation of the environment through measurement and detection" and "Although the concept is not new, MDF technology is still in its infancy".

The difficulty in applying the data fusion approaches already used in other fields is related to the differences among the variables involved. The use of data fusion in weather forecasting, for example, takes into consideration factors such as air humidity, soil and air temperature, and wind velocity in order to predict the probability of rain. The fusion of data from many sensors adds a variable to a problem that is usually underdefined, and with an output that is different from the input. In this case the data fusion has the aim of creating a model to link together complementary information on the same process. The methods used for this data fusion have to create a model to link multiple parameters to a different output. The data fusion has to find a model to approximate an unknown model. For this reason the approaches used are usually without a predetermined model such as fuzzy logic (Chen and Huang 2000) or neural network (Huang *et al.* 2007).

The case of mechanical measurement is completely different. In this case the inputs are all measurements of the product and so is the result. Moreover, in

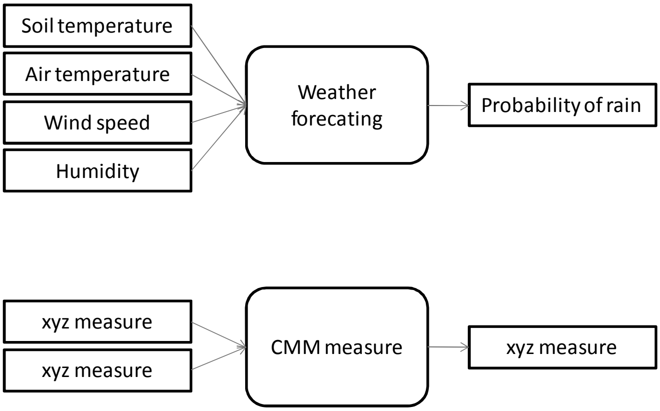


Figure 6.11 The data fusion model for weather forecasting and a mechanical measurement

order to improve the precision of the data, both inputs have to be referred to the same spatial position. The strategy used to solve the data fusion process in the mechanical measurement field is to create a linkage model between the high-resolution data and the low-resolution data. Such a model would allow the prediction of the high-resolution data starting from the low-resolution data. From such a model it would be possible to obtain measurements with the same spatial coordinate from the low-resolution and high-resolution data that could be fused together. A proposal for such model was made by Xia *et al.* (2008a), based on a Bayesian hierarchical model, and it is explained briefly in the following paragraphs.

The difference between input and output for weather forecasting and measurement using a hybrid CMM is reported in Figure 6.11.

Recent studies on the problem of synthesizing spatial data collected at different scales and resolutions by Ferreira *et al.* (2005) and on the problem of calibrating computer simulation models of different accuracies by Kennedy and O'Hagan (2001) and Qian and Wu (2006) have proven that multiresolution data can be fused together, once a fit linkage model has been chosen. These studies however do not consider the misalignment issue, which is a fundamental aspect for multiresolution data in the field of mechanical measurements.

The approach proposed by Xia *et al.* (2008a) was called a Bayesian hierarchical model by the authors. This starts from the modeling of the low-resolution data as a Gaussian process – this was already proven to be a valid approach for manufactured parts by Xia *et al.* (2008b) – and then the low-resolution and high-resolution data are aligned with a variant of the TrICP algorithm (Section 6.2) in order to create a linkage model represented as a K kernel function. The linkage model is then used to fuse together the low-resolution and high-resolution data thanks to a Bayesian prediction. The general scheme of the approach, as presented by Xia *et al.* (2008b), is shown in Figure 6.12.

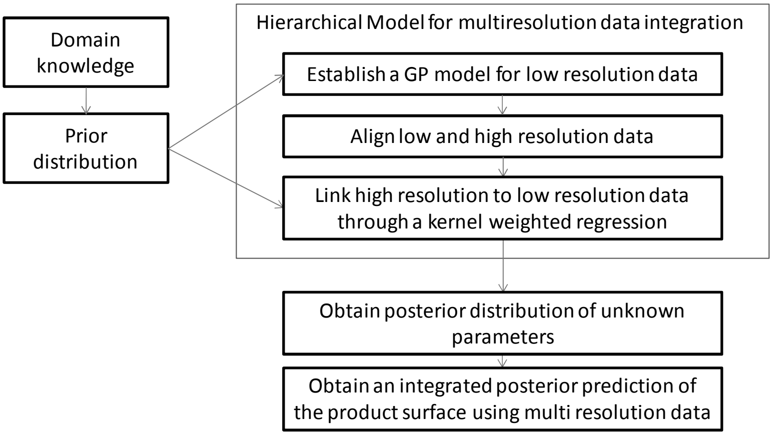

Domain knowledge
Prior distribution
Hierarchical Model for multiresolution data integration
Establish a GP model for low resolution data
Align low and high resolution data
Link high resolution to low resolution data through a kernel weighted regression
Obtain posterior distribution of unknown parameters
Obtain an integrated posterior prediction of the product surface using multi resolution data

Figure 6.12 The Bayesian hierarchical model. *GP* Gaussian process

Modeling the low-resolution data as a Gaussian process means that the value of each measurement acquired for coordinate x_i is composed of a mean value and an error:

$$y_l(x_i) = \eta_l(x_i) + \varepsilon_l(x_i), \tag{6.2}$$

where $y_l(x_i)$ is the acquired low-resolution data, $\eta_l(x_i)$ is the actual value, and $\varepsilon_l(x_i)$ is the error. As in the Gaussian process, the mean value is made up of two components, the mean component and the covariance function:

$$\mathrm{cov}\left[\eta_l(x_i), \eta_l(x_j)\right] = k_l^2 R(x_i, x_j), \tag{6.3}$$

where k is the variance and R the correlation function, which can be expressed as

$$R(x_i, x_j) = \prod_{k=1}^{d} \exp\left[-v_k\left(x_{ki} - x_{kj}\right)^2\right], \tag{6.4}$$

where d is the dimension of the input variable and v_k are the scale parameters that control the "damping" of the correlation with respect to the distance from x_i. The modeling of the low-resolution data with a Gaussian process is useful for the definition of the kernel function, which is necessary for the creation of a linkage model between the high-resolution and the low-resolution data.

After the modeling, the two datasets are aligned in order to create a linkage model. For the precise alignment the algorithm proposed is a small variation of the TrICP algorithm presented in Section 6.2. The general idea is to perform the alignment using only a subset of the data. The points considered are the whole high-resolution dataset and an iteratively chosen group of low-resolution data. The iterations of the algorithm that choose the new dataset at each iteration are carried out until there is no improvement in the objective function, defined as in the TrICP algorithm.

The neighborhood linkage model has a double objective: to create a linking model between low-resolution and high-resolution data and to enable a fine alignment with greater precision with respect to the previous one. The first alignment carried out with the TrICP algorithm is necessary to give a first approximate alignment for the creation of a reliable linking model. The linking model is constituted by an adaptive kernel function:

$$y_h(x_j) = \alpha_1 \sum_{i=1}^{l} K(x_j, x_i)\eta_l(x_i) + \alpha_0 + e(x_j), \tag{6.5}$$

where $y_h(x_j)$ is the high-resolution data, α_0 and α_1 are the scale and location coefficients of the alignment, e is the residual, and $K(x_j, x_i)$ is the kernel function defined as

$$K(x_i, x_j) = D\left[\sum_{k=1}^{d}\left(\frac{x_{ki} - x_{kj}}{\lambda_k}\right)\right], \tag{6.6}$$

with $D(t) = (1 - t^{3/2})^3$ if $t < 1$ or 0 if $t > 1$ and λ the size of the neighborhood automatically assigned and controlled by the kernel.

In order to fuse together the data, a prior distribution $p(\theta)$ is defined as the product of the prior distributions of the individual parameters. From the prior it is possible to define a Bayesian predictor that integrates the multiresolution data in order to provide the value of $y_h(x_0)$ at the input location x_0 considering the neighbor data from the low-resolution and high-resolution datasets. The predictor can be defined as

$$p\left[y_h(x_0)\middle|y_l, y_h\right] = \int_{\theta_l} p\left[y_h(x_0)\middle|\eta_l, y_h\right] \cdot p\left(\eta_l\middle|\theta_l, y_l\right) \cdot p\left(\theta_l\right) \cdot d\theta_l. \tag{6.7}$$

This equation needs to be solved numerically using a Markov chain Monte Carlo method to approximate some terms in the integral.

The proposed approach has been tested and the results are very encouraging. The improvement in precision that could be attained using multiresolution data, instead of high-resolution data only, for some preliminary tests conducted by Xia *et al.* (2008a) is of about 1 order of magnitude.

The multiresolution fusion process developed is a novel approach for the use of multisensor CMMs and, although actually mathematically complex, could be a first step in the design of new CMM control software and acquisition processes. The automation of such an approach could allow for a consistent reduction in the time needed to obtain a reverse modeling or a geometric tolerance measurement of a product.

6.7 Concluding Remarks

The introduction of OCMMs in the field of product inspection and reverse modeling has brought a great advance in measurement technologies. The advantages are mainly related to the possibility to automate the entire process and to obtain a result in ever-decreasing processing time. Most of the approaches developed in the last decade are based on a sequential acquisition of data: a fast acquisition of low-resolution data is done first, followed by a more time-consuming acquisition of higher accuracy data. This second acquisition can be optimized thanks to the analysis of the low-resolution dataset; many approaches that have been developed are also able to create the CMM operation code for the high-resolution data acquisition with almost entirely automated procedures. This field of research is extremely active. Apart from continuous work to improve the alignment algorithms, the new frontier of the research is to shift from a sequential approach to a real data fusion in order to save even more time and to improve the precision of the final dataset. In the case of data fusion, a general model capable of handling various applications is very difficult to achieve. For a specific area of research and application, a specific data fusion model must be built. Recent activities in this field have shown encouraging results that, we hope, will motivate more and more researchers to develop new approaches and seek the automation of the data fusion process in the field of CMMs.

References

Besl P, McKay N (1992) A method for registration of 3-D shapes. IEEE Trans Pattern Anal Mach Intell 14:239–256

Carbone V, Carocci M, Savio E, Sansoni G, De Chiffre L (2001) Combination of a vision system and a coordinate measuring machine for the reverse engineering of freeform surfaces. Int J of Adv Manuf Technol 17:263–271

Cernuschi-Frias B (1984) Orientation and location parameter estimation of quadric surfaces in 3-D space from a sequence of images. PhD dissertation, Brown University of Providence

Chan V, Bradley C, Vickers W (1997) Automating laser scanning of 3-D surfaces for reverse engineering. SPIE 3204:156–164

Chan VH, Bradley C, Vickers GW (2001) A multi-sensor approach to automating co-ordinate measuring machine-based reverse engineering. Comput Ind 44:105–115

Chen LC, Lin GCI (1991) A vision-aided reverse engineering approach to reconstructing freeform surfaces. Robot Comput Integr Manuf 13(4):323–336

Chen LC, Lin GCI (1997) Reverse engineering of physical models employing a sensor integration between 3D stereo detection and contact digitization. Proc SPIE 3204:146–155

Chen YM, Huang HC (2000) Fuzzy logic approach to multisensor data association. Math Comput Simul 52:399–412

Chen X, Xia J, Jin Y, Sun J (2009) Accurate calibration for a camera–projector measurement system based on structured light projection. Opt Lasers Eng 47(3–4):310–319

Chetverikov D, Stepanov D, Krsek P (2005) Robust Euclidean alignment of 3D point sets: the trimmed iterative closest point algorithm. Image Vision Comp 23:299–309

Daniels J, Ha HK, Ochotta T, Silva CT (2007) Robust smooth feature extraction from point clouds. In: Proceedings of IEEE international conference on shape modeling and applications

El-Hakim SF, Pizzi N (1993) Multicamera vision-based approach to flexible feature measurement for inspection and reverse engineering. Opt Eng 32:2201–2215

Ferreira M, Higdon D, Lee H (2005) Multi-scale random field models. ISDS discussion paper 05-02. http://ftp.stat.duke.edu/WorkingPapers/05-02.html

Forest J, Salvi J (2002) A review of laser scanning three dimensional digitizers. Intell Robot Syst 73–78

Fujimoto M, Kadya K (1993) An improved method for digitised data reduction using an angle parameter. Measurement 12:113–122

Hornbeck LJ (1998) From cathode rays to digital micromirrors: a history of electronic projection display technology. Texas Instruments Technol J 15(3):7–40

Huang J, Menq CH (1999) Feature extraction from incomplete surface data. Coordinate Metrology Measurement Laboratory technical report CMML-99-03

Huang YB, Lan YB, Hoffmann WC, Lacey RE (2007) Multisensor data fusion for high quality data analysis and processing in measurement and instrumentation. J Bionic Eng 4:53–62

Kennedy M, O'Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc Ser B 63:425–464

Lee N L, Menq CH (1995) Automated recognition of geometric forms from B-rep models. In: Proceedings of ASME international computers in engineering conference, pp 805–816

Pulli K (1997) Surface reconstruction and display from range and color data. PhD thesis, University of Washington

Qian Z, Wu CFJ (2006) Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. Technometrics 50(2):192–204

Rousseeuw P, Van Zomeren B (1990) Unmasking multivariate outliers and leverage points. J Am Stat Assoc 85:633–651

Nashman M (1993) The use of vision and touch sensors for dimensional inspection tasks. Manuf Rev 6(2):155–162

Nashman M, Hong TH, Rippey WG, Herman M (1996) An integrated vision touch-probe system for dimensional inspection tasks. In: Proceedings of SME Applied Machine Vision '96 conference, Cincinnati, pp 243–255

Shen TS, Huang J, Menq CS (2000) Multiple-sensor integration for rapid and high-precision coordinate metrology. IEEE/ASME Trans Mechatronics 5(2):110–122

Spitz SN, Spyridi AJ, Requicha AAG (1999) Accessibility analysis for planning of dimensional inspection with coordinate measuring machines. IEEE Trans Robot Automat 15(4):714–727

Weng J, Cohen P, Herniou M (1992) Camera calibration with distortion models and accuracy evaluation. IEEE Trans Pattern Anal Machine Intell 14:965–980

Xia H, Ding Y, Mallicky BK (2008a) Bayesian hierarchical model for integrating multi-resolution metrology data. In: Proceedings of European Network for Business and Industrial Statistics conference

Xia H, Ding Y, Wang J (2008b) Gaussian process method for form error assessment using coordinate measurements. IIE Trans 40(10):931–946

Yang J, Lee NL, Menq CH (1995) Application of computer vision in reverse engineering for 3D coordinate acquisition. In: Proceedings of the symposium on concurrent product and process engineering, ASME international mechanical engineering congress and exposition, pp 143–156

Yau HT, Menq CH (1996) A unified least-squares approach to the evaluation of geometric errors using discrete measurement data. Int J Mach Tools Manuf 36(11):1269–1290

Zexiao X, Jianguo W, Ming J (2007) Study on a full field of view laser scanning system. Int J Mach Tools Manuf 47:33–43

Zussman E, Schuler H, Seliger G (1994) Analysis of the geometrical feature detectability constraints for laser-scanner sensor planning. Int J Adv Manuf Technol 9:56–64

Chapter 7
Statistical Shape Analysis of Manufacturing Data

Enrique del Castillo

Abstract We show how statistical shape analysis, a set of techniques used to model the shapes of biological and other kinds of objects in the natural sciences, can also be used to model the geometric shape of a manufactured part. We review Procrustes-based methods, and emphasize possible solutions to the basic problem of having corresponding, or matching, labels in the measured "landmarks", the locations of the measured points on each part acquired with a coordinate measuring machine or similar instrument.

7.1 Introduction

In statistical shape analysis (SSA) the *shape* of an object is defined as all the information of the object that is invariant with respect to similarity transformations on the Euclidean space (rotations, translations, and dilations or changes of scale). The goal of SSA is to analyze the shapes of objects in the presence of random error.

Analysis of shapes in manufacturing is critical because geometric tolerances (specifications) of roundness, flatness, cylindricity, *etc.*, need to be inspected, controlled, or optimized based on a cloud of two- or three-dimensional measurements taken on the machined surfaces of the part. These tasks are even more complex if the part geometry has a "free form", *i.e.*, there is no standard geometric construction that can represent the shape, a situation common in advanced manufacturing applications such as in the aerospace sector.

E. del Castillo
Department of Industrial and Manufacturing Engineering, The Pennsylvania State University,
University Park, PA 16802, USA,
e-mail: exd13@psu.edu

Over the last 20 years, SSA techniques have been developed and applied in many areas of the natural sciences where there is interest in characterizing differences between and variability within shapes, *e.g.*, biology, paleontology, and geology. A considerable intersection of ideas also exists with image and pattern recognition in computer science. In particular, SSA is known as *geometric morphometrics* in biology, and the type of techniques developed in the past two decades has been called the "morphometrics revolution" by some authors (Adams *et al.* 2004) given the success SSA had over previous techniques used to analyze shapes. For more on the history and foundations of SSA with applications in biological sciences we refer readers to the books by Dryden and Mardia (1998) and Zelditch *et al.* (2004).

Our interest in shape analysis stems in part from recent interest on "profile analysis" in the field of statistical process control (SPC) (Kang and Albin 2000; Colosimo *et al.* 2008; Woodall *et al.* 2004) (although we do *not* discuss SPC based on shape analysis in this chapter, this is certainly another potential area of research where SSA ideas can be used). In profile-based SPC, a parametric model is sought that describes the form that the response follows with respect to some variable of interest (in essence, one performs functional data analysis). The parameters of this model are fitted on the basis of process data and then multivariate SPC methods are applied to the estimated parameters.

By working with the shape directly, SSA techniques avoid the parametric model definition step, allow complicated shapes to be studied, and simplify the presentation of results. In SSA, one works with the whole shape of the object, so the geometry is not "thrown away" (Dryden and Mardia 1998).

The remainder of this chapter is organized as follows. Section 7.2 describes methods to solve the landmark matching problem, which occurs when two objects have point labels that do not correspond to each other. In Section 7.3, we review the main ideas of SSA based on the so-called Procrustes method. In this section, the notions of shape space, the generalized Procrustes algorithm (GPA), and tangent space coordinates are discussed. The chapter concludes with a discussion of other shape analysis techniques, including areas for further work.

7.2 The Landmark Matching Problem

In most of SSA, the main goal is statistical inference with shapes, in particular, to test if two or more objects have an equal shape or not, or to determine the directions in which most of the variability of a shape occurs. Some other authors' main interest (*e.g.*, in biology) is to describe how shapes of objects (*e.g.*, species of animals) change with time. In our case, the main goal is to study the shape of manufactured parts.

The techniques considered herein are based on shape data obtained by measuring the parts at specific *landmarks*, points of special interest or unique characteristics. In order to be amenable to data analysis, landmarks should refer to homologous points (points of correspondence) that match between objects. A landmark is given by the two- or three-dimensional Cartesian coordinates of a point on the object surface and a given label for the point, usually a sequential number $1,2,\ldots,k$ which corresponds from object to object. Assignment of landmarks to objects is in itself an important problem; in some areas, such as in archeology and biology, specific points of the objects are of scientific interest and this assignment is done manually. In manufacturing, considerable amounts of data can be acquired with a coordinate measuring machine (CMM) or through digital images of the objects. There is no guarantee in practice, however, that the measurements acquired will correspond to each other between parts. Homologous landmarks have the same label, hence we call the case of complete homologous landmarks the *labeled* case.

All the SSA methods considered in later sections of this chapter require labeled landmark data. Similar parts measured with a CMM do not always contain corresponding or labeled landmarks. This can be due to the difficulty in orienting the part when mounting it on the CMM. If the orientation is different between parts, the CMM measurements will not correspond to each other, since they will have different labels. Therefore, one first important problem that needs to be addressed is how to "match" the landmarks between two or more shapes so that we obtain corresponding shape data. This problem has received attention in the pattern recognition literature in recent years, where it is called the point matching or shape matching problem. The work by Ranjaragan *et al.* (Gold *et al.* 1998; Chui and Rangarajan 2000) is based on solving a highly nonlinear optimization problem where the objective is to minimize the sum of the Euclidean distance between points $\{i\}$ in shape 1 and the transformed points $\{j\}$ in shape 2. The rationale for this approach is that matching would be easier if the objects were oriented similarly, and have a similar location and scale (similarity transformations). Jointly determining the matching correspondences and the transformation necessary for "registering" object 2 to object 1 results in a hard optimization problem.

A completely different approach is that of Belongie *et al.* (2002), who proposed an efficient method for matching two two-dimensional shapes, although they left undefined some implementation details, as we will see below. Their method separates the landmark matching problem from the problem of registering the objects, that is, their matching method is in principle invariant with respect to location, scaling, and orientation of the two parts. The main idea is to measure the amount of data in the neighborhood of each point of each shape (given by the frequency of points in its neighborhood) and use these measures as costs to be minimized in a classic weighted matching problem, solvable via linear programming. For point *i* in a shape, Belongie *et al.* proposed computing a two-dimensional histogram where the number of points nearby are counted. If r is the Euclidean distance between two points of the shape, the two-dimensional histogram extends along $\log r$ and θ, measuring the distance and direction where the nearby points are located. The histogram bins are selected such that they are of constant width in $(\log r,\theta)$, giving more importance in this way to closer points. Let $h_i(l,s)$ be the observed frequency of nearby points in cell (l,s) of the histogram, where $l=1,\ldots,L$, $s=1,\ldots,S$. The two-dimensional histogram formed by the frequencies $h_i(\cdot,\cdot)$ is called the “context” of point i by these authors. The idea then is to match those points between two different shapes that have the most similar “contexts”. For this purpose, define the cost of matching point i in part 1 and point j in part 2 to be

$$C_{ij} = \sum_{l=1}^{L}\sum_{s=1}^{S}\frac{\left[h_i(l,s)-h_j(l,s)\right]^2}{h_i(l,s)+h_j(l,s)},\ i=1,2,\ldots,k, j=1,2,\ldots,k, \tag{7.1}$$

which is the classic χ^2 statistic (with $L\cdot S-1$ degrees of freedom) used to test for the difference between two distributions. Note that $C_{ij}\neq C_{ji}$. Let $B=(U,V,E)$ be a graph with two disjoint sets of points (U and V), *i.e.*, a bipartite graph, and a set of edges (E, to be decided) joining a point in U with a point in V (the “matching” set). Define the decision variables $X_{ij}=1$ if the edge joining points $v_i\in V$ and $u_j\in U$ is included in the matching, and $X_{ij}=0$ otherwise (Papadimitrou and Steiglitz 1982). Belongie *et al.* (2002) proposed solving the landmark matching or labeling problem by solving the following weighted matching problem (in our notation):

$$\begin{aligned}
&\min \textstyle\sum_{i=1}^{k}\sum_{j=1}^{k} C_{ij}X_{ij}\\
&\text{subject to}\\
&\textstyle\sum_{j=1}^{k} X_{ij}=1,\ i=1,2,\ldots,k,\\
&\textstyle\sum_{i=1}^{k} X_{ij}=1,\ j=1,2,\ldots,k,\\
&X_{ij}\geq 0,\ i=1,\ldots,k; j=1,\ldots,k.
\end{aligned}\tag{7.2}$$

The problem is then one of *linear* programming, for which, as is well known, there exist efficient algorithms. Note that the formulation does not include the constraints $X_{ij}\in\{0,1\}$, which turn out to be redundant (the linear programming solution is always binary) so the problem is *not* an integer programming problem, which would imply a considerably harder optimization problem. The matrix formulation of the problem is based on defining the $k^2\times 1$ vector of decision variables

$$\mathbf{x}' = (X_{11}, X_{12},\ldots,X_{1k},X_{21},X_{22},\ldots,X_{2k},\ldots,X_{k1},X_{k2},\ldots,X_{kk})$$

and defining the $2k\times k^2$ matrix of constraint coefficients:

$$\mathbf{A} = \left[\begin{array}{cccccccccc}
1 & 1 & \ldots & 1 & & & & & & \\
 & & & & 1 & 1 & \ldots & 1 & & \\
 & & & & & & & & \ddots & \\
 & & & & & & & & & 1 \quad 1 \quad \ldots \quad 1\\
1 & & & & 1 & & & & \ldots & 1\\
 & 1 & & & & 1 & & & \ldots & \quad 1\\
 & & \ddots & & & & \ddots & & \ldots & \quad\quad \ddots\\
 & & & 1 & & & & 1 & \ldots & \quad\quad\quad 1
\end{array}\right],$$

where all empty spaces are zeroes. If matrix $\mathbf{C} = \{C_{ij}\}$ is put into vector form as follows

$$\mathbf{c}' = (C_{11}, C_{12}, \ldots, C_{1k}, C_{21}, C_{22}, \ldots, C_{2k}, \ldots, C_{k1}, C_{k2}, \ldots, C_{kk}),$$

then the formulation is simply

$$\min \mathbf{c}'\mathbf{x}$$

$$\text{subject to}$$

$$\mathbf{A}\mathbf{x} = \mathbf{b},$$

$$\mathbf{x} \geq \mathbf{0},$$

where $\mathbf{b}$ is a $2k \times 1$ vector of ones. The property that ensures a $\{0,1\}$ solution is that matrix $\mathbf{A}$ is a totally unimodal matrix (Papadimitrou and Steiglitz 1982, Theorem 13.3). A matrix is a totally unimodal matrix if (1) it has zeroes except at two locations per column, where it has ones, and (2) the rows can be grouped in two sets such that the ones in each column belong to different sets. These two properties hold for matrix $\mathbf{A}$.

An important implementation detail is how to scale the distances. We suggest defining $r_{ij} = d_{ij} / \max(d_{ij})$, where d_{ij} is the Euclidean distance between points i and j in the object and the maximum is measured over all distances between any two points (landmarks). Therefore, $\max(r_{ij}) = 1$.

Example. Landmark matching. Suppose we have the two shapes shown in Figure 7.1. These are two handwritten digit 3's, each with 13 landmarks. Suppose the landmarks are labeled as shown in the Figure and in the first four columns of Table 7.1. We will keep the labels of shape 1 constant and will try to match the labels of the second shape to those of the first. The cost matrix $\mathbf{C}$ is shown in Table 7.2. This was obtained using a two-dimensional histogram at each point of each shape where $L = 10$ bins were used for $\log r$ (logarithm of Euclidean dis-

Table 7.1 Input matrices for the "digit 3's" problem (first four columns). The last two columns are the output, sorted matrix

14	41	21	25	9	39
21	42	22	19	15	39
29	42	25	22	21	40
35	37	9	39	25	36
32	33	21	27	23	31
26	30	21	40	21	27
16	26	15	39	19	25
25	26	8	17	21	25
29	24	25	36	23	24
33	20	15	17	25	22
30	16	19	25	22	19
23	11	23	31	15	17
16	12	23	24	8	17

Table 7.2 The cost matrix **C** for the two digit 3's problem. This is a $k \times k = 13 \times 13$ nonsymmetric matrix. *Bold numbers* correspond to costs for the optimal matching solution

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	7.0	8.3	7.8	**4.7**	9.3	6.8	4.4	11.5	10.2	9.5	10.1	10.2	7.3
2	7.7	7.0	9.5	6.8	9.2	5.5	**2.0**	10.4	8.0	6.7	10.1	9.2	8.0
3	6.2	3.5	6.3	5.7	8.2	**1.6**	5.9	6.0	6.7	2.7	10.7	7.0	6.8
4	9.6	7.7	8.8	9.0	6.3	7.5	8.2	7.0	**2.6**	5.3	9.0	3.8	8.5
5	6.3	6.1	8.2	10.7	7.0	7.2	7.7	8.8	5.6	5.8	8.8	**3.5**	8.7
6	8.5	9.3	12.5	10.7	**6.5**	7.8	4.8	10.1	7.2	6.9	11.0	9.2	9.7
7	7.8	8.0	10.0	9.0	7.8	6.5	3.2	6.3	7.3	5.3	**9.0**	7.5	8.5
8	**5.0**	8.5	9.2	11.7	5.7	6.6	7.7	11.7	7.5	6.8	10.2	7.2	9.4
9	3.0	7.1	6.7	9.8	6.2	4.7	7.7	11.0	6.8	6.7	9.5	7.0	**5.2**
10	5.2	6.3	**4.5**	7.5	8.3	4.8	5.8	10.6	7.6	7.7	9.7	8.2	4.3
11	5.8	**1.3**	4.5	5.8	6.3	3.2	4.2	8.9	7.0	5.4	11.0	5.5	5.7
12	5.3	5.3	8.1	6.1	8.6	2.8	4.7	6.7	4.6	**2.0**	12.5	9.5	6.1
13	6.3	6.9	7.7	7.7	10.2	3.1	8.9	**4.5**	5.1	3.3	12.0	9.4	6.2

tances) and $S = 9$ bins were used for θ. Specifically, the bin edges where set at $[0, \exp(-4), \exp(-3.5), \exp(-3), \exp(-2.5), \exp(-2), \exp(-1.5), \exp(-1), \exp(-0.5), 1]$ and $[-\pi, -3*\pi/4, -\pi/2, -\pi/4, 0, \pi/4, \pi/2, 3*\pi/4, \pi]$. (This is a higher-resolution histogram than that used by Belongie *et al.* 2002; we found the results vary considerably with the resolution of the histogram, given by the number of bins.

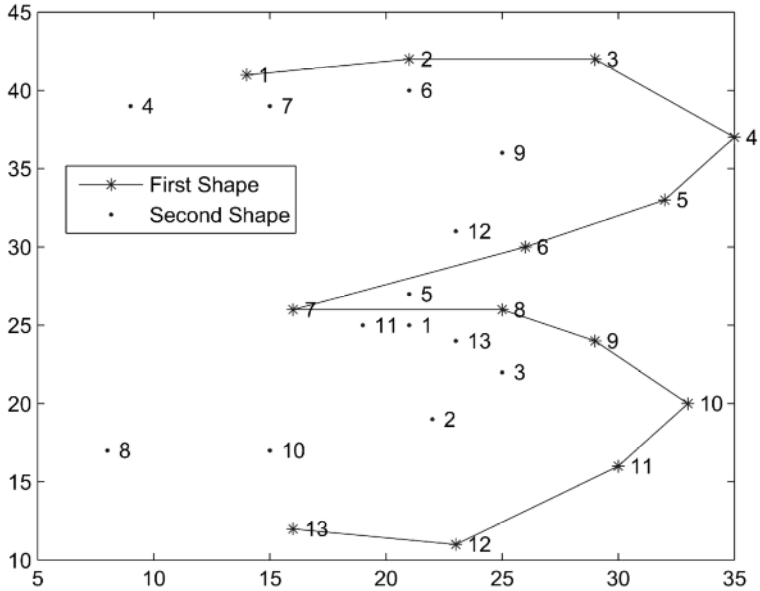


Figure 7.1 Two handwritten digit "3"s, each with k = 13 landmarks. The labels of the landmarks of the second digit were shuffled and do not correspond to those of the first digit 3

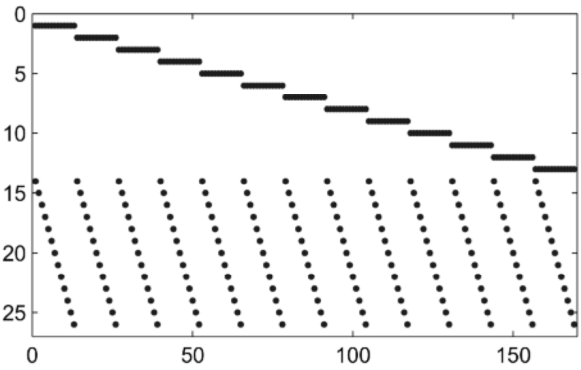


Figure 7.2 The **A** matrix in the weighted matching linear programming formulation applied to the two digit 3's matching problem (where $k = 13$). *Dots* indicate 1's, *empty spaces* indicate 0's. This is a $2k = k^2 = 26 \times 169$ matrix

Intuitively, the number of bins should be an increasing function of k, the number of landmark points.) The costs C_{ij} were computed by excluding those cells in the histograms that would lead to a zero denominator. The structure of the **A** matrix, a 26×169 matrix, is shown diagrammatically in Figure 7.2. Solving the resulting linear programming problem (we used the `linprog` routine in MATLAB), the optimal solution leads to the correspondences shown in Figure 7.3. The reordered landmark matrix for the second digit 3 is shown at the right in Table 7.1.

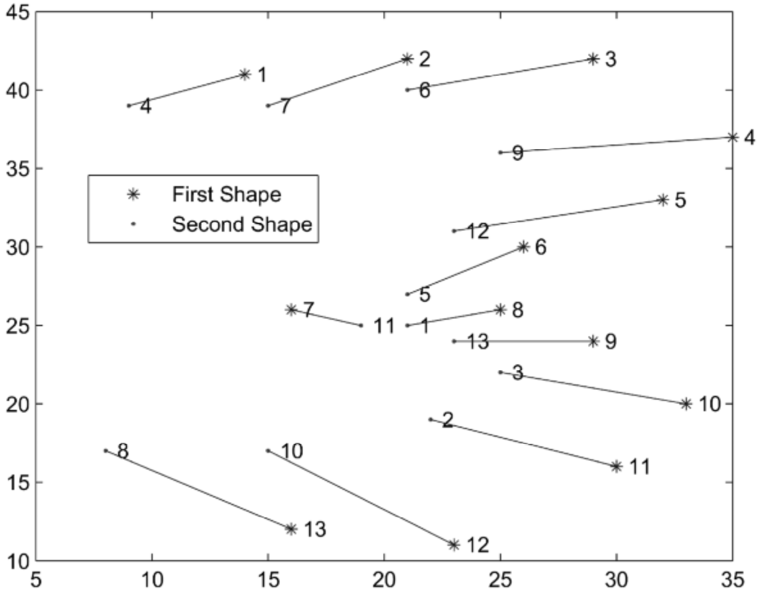


Figure 7.3 Optimal solution to the linear programming matching problem, digit 3 problem. Compare with Figure 7.1

We point out how the two configurations in the example did not have the same scale. We would like a matching algorithm to work even if the configurations are not equally oriented (*i.e.*, to be rotation-invariant; the method is already location- and scale-invariant). For configurations with different orientation, a simple solution, which we used in the previous example, is to compute the angles θ in the histogram with respect to the line defined by the two closest points to the point in question (this is a procedure similar to that suggested by Belongie *et al.* 2002, who suggested using the "tangent" line to each point as the axis of reference).

An apparently open problem in the literature is how to solve similar matching problems when there are n shapes, not only two. A first attempt to solve such a problem may involve matching shapes (1, 2), giving 2' (the relabeled object 2), then matching (2', 3), (3', 4), ... , (n–1', n) and then repeating matching (n', 1), (1', 2'), *etc.*, until there is convergence. It is unknown how effective such an approach is, and whether or not convergence is guaranteed.

7.3 A Review of Some SSA Concepts and Techniques

There is a very large body of literature on SSA techniques. Only the main precepts and techniques are presented here. For a more thorough presentation of SSA, we refer the reader to Dryden and Mardia (1998) and Goodall (1991) for developments up to 1998 and for more recent developments we refer the reader to Adams *et al.* (2004), Kent and Mardia (2001), Klingenberg and Monteiro (2005), and Green and Mardia (2006).

The mainstream of SSA that followed the "revolution" in morphometrics is based on two main steps. First, the objects under consideration are registered or superimposed with respect to each other in order to filter out rotation, translation, and isometric scaling (dilation) effects. This is done because the objects may have different orientations on the Euclidean space or have different locations or sizes, and therefore their shapes cannot be initially compared. The main technique for this task is the GPA.

An underlying assumption of the GPA is that landmarks refer to homologous or corresponding points in each object. Since this is not always the case in CMM data, the landmark matching problem discussed in the previous section must be solved first before attempting the registration. Matching before registering seems to be a simpler and better strategy than trying to jointly match the landmarks and register the objects, as attempted in Gold *et al.* (1998) and Chui and Rangarajan (2000).

Once objects have been registered, multivariate statistical methods of inference can be performed on the projections of the shapes on the space tangent to the mean shape. These two steps are explained below. We first give some geometric notions necessary to understand the algorithms.

7.3.1 Preshape and Shape Space

Let $\mathbf{X}$ be a $k \times m$ matrix containing the k landmarks (coordinate pairs or triples) of an object in m (two or three) dimensions. $\mathbf{X}$ is sometimes called a *configuration matrix*, which we could also refer to as a "profile matrix", following manufacturing practice for the case of two-diomensional closed contours (ASME Y14.5M 1994). With this notation, the shape of a configuration $\mathbf{X}$ is obtained, first, by removing location and scale effects by computing the so-called *preshape* $\mathbf{Z}$:

$$\mathbf{Z} = \frac{\mathbf{HX}}{\|\mathbf{HX}\|}, \tag{7.3}$$

where $\mathbf{H}$ is a $(k-1)\times k$ Helmert submatrix (Dryden and Mardia 1998) and $\|\cdot\|$ denotes the Frobenius norm of a matrix (*i.e.*, $\|\mathbf{A}\| = \sqrt{\sum_i \sum_j |a_{ij}|^2}$). If we define $h_j = -[j(j+1)]^{-1/2}$, then $\mathbf{H}$ is a matrix whose jth row is $(\underbrace{h_j, h_j, \ldots, h_j}_{j \text{ times}}, -jh_j, \underbrace{0, \ldots, 0}_{k-j-1 \text{ times}})$ for $j = 1, \ldots, k-1$. Note that $\mathbf{HH}' = \mathbf{I}_{k-1}$ and that the rows of $\mathbf{H}$ are contrasts. Alternatively, one could start with the *centered preshapes*, defined by $\mathbf{Z}_c = \mathbf{H}'\mathbf{Z}$ (these are $k \times m$ matrices).

The transformation in Equation 7.3 removes location effects via the numerator, and rescales the configurations to unit length via the denominator. Since we have not removed rotations from $\mathbf{Z}$, it is not yet the shape of $\mathbf{X}$, hence the name preshape. The centered preshapes are equivalent to centering each coordinate of each configuration by its centroid and dividing each by its norm.

The *shape* of configuration $\mathbf{X}$, denoted $[\mathbf{X}]$, is defined as the geometric information that is invariant to similarity transformations. Once location and scale effects have been filtered as above, the shape is then defined as

$$[\mathbf{X}] = \{\mathbf{Z\Gamma} : \mathbf{\Gamma} \in SO(m)\}, \tag{7.4}$$

where $\mathbf{Z}$ is the preshape of $\mathbf{X}$ and $\mathbf{\Gamma}$ is a rotation matrix [*i.e.*, a matrix such that $\mathbf{\Gamma}'\mathbf{\Gamma} = \mathbf{\Gamma\Gamma}' = \mathbf{I}_m$ with $\det(\mathbf{\Gamma}) = +1$] and $SO(m)$ is the space of all $m \times m$ rotation matrices, the special orthogonal group. Multiplication by a suitable matrix $\mathbf{\Gamma}$ reorients (rotates) the object. Note that a shape is therefore defined as a set.

The following geometric interpretation of these transformations is due to Kendall (1984, 1989). Given that preshapes are scaled and centered objects, they can be represented by vectors on a sphere of dimension $(k-1)m$, because the numerator in Equation 7.3 removes m degrees of freedom for location parameters and the denominator removes one additional degree of freedom for the change of scale. The preshapes, having unit length, are therefore on the surface of this (hyperspherical) space, which has $(k-1)m-1$ dimensions by virtue of being on the surface of a unit sphere. As one rotates a preshape $\mathbf{Z}$ via Equation 7.4, the vectors $\mathbf{Z\Gamma}$ describe an *orbit*, in effect, a geodesic, on the preshape space. All these vectors correspond to the same shape, since by definition the shape of an object is invariant to rotations. Thus, the orbits (also called *fibers*) of the preshape space are mapped one to one into single points in the *shape space*, the space where shapes reside. Two objects have the same shape if and only if their preshapes lie on the same fiber. Fibers do not overlap. The shape space, the space of all possible shapes, has dimension $M = (k-1)m - 1 - m(m-1)/2$ since in addition to losing location and dilation degrees of freedom we also lose $m(m-1)/2$ degrees of freedom in the specification of the (symmetric) $m \times m$ rotation matrix $\mathbf{\Gamma}$.

Example. Preshape space and shape space. In order to explain these ideas, consider one of the simplest possible cases, where we have two lines in R^2. Thus, we have that $m = 2$ and $k = 2$, where the obvious landmarks are the endpoints of the lines. After centering and scaling the two lines using Equation 7.3, one obtains the preshapes with matrices $\mathbf{Z}_1$ and $\mathbf{Z}_2$. Since the original objects evidently have the same shape (that of a line in Euclidean space), these two preshapes lie on the same fiber or orbit, generated as the preshapes are rotated using Equation 7.4. The shape space is of dimension $(k-1)m-1=1$, namely, the circumference of a circle. As the preshapes rotate (they can rotate clockwise or counterclockwise) they will eventually coincide, which corresponds to the centered and scaled lines coinciding. Finally, since there is a single shape, the shape space is the simplest possible, namely, a single point [the dimension is $M = (k-1)m - 1 - m(m-1)/2 = 0$].

In general, the shape space will also be a spherical, nonlinear space, of reduced dimension compared with the preshape space.

7.3.2 Generalized Procrustes Algorithm

Two preshapes $\mathbf{Z}_1$ and $\mathbf{Z}_2$ lying on different fibers correspond to two objects with different shapes. A measure of the similarity between two shapes is the shortest distance between the fibers, the *Procrustes distance* $\rho(\mathbf{X}_1, \mathbf{X}_2)$. This corresponds to the distance along the surface of the preshape space and is therefore a distance along a geodesic. Alternatively, two measures of distance over a linear space are the "partial Procrustes distance", given by

$$d_p(\mathbf{X}_1, \mathbf{X}_2) = \min_{\mathbf{\Gamma} \in SO(m)} \left\| \mathbf{Z}_2 - \mathbf{Z}_1 \mathbf{\Gamma} \right\|, \tag{7.5}$$

and the "full Procrustes distance", where the minimization is also done over a scale parameter:

$$d_F(\mathbf{X}_1, \mathbf{X}_2) = \min_{\mathbf{\Gamma} \in SO(m), \beta \in \mathbb{R}} \left\| \mathbf{Z}_2 - \beta \mathbf{Z}_1 \mathbf{\Gamma} \right\|. \tag{7.6}$$

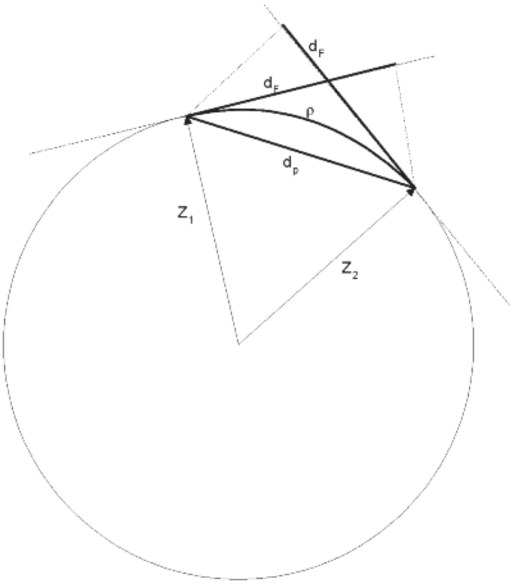


Figure 7.4 Distances between two shapes in preshape space. ρ is the Procrustes distance (along a geodesic), d_F is the full Procrustes distance (along a tangent), and d_p is the partial Procrustes distance (along the secant). The preshapes have $||Z_i|| = 1$

Geometrically, $d_p(\mathbf{X}_1, \mathbf{X}_2)$ is the secant between $\mathbf{Z}_1$ and $\mathbf{Z}_2$ in preshape space, and $d_F(\mathbf{X}_1, \mathbf{X}_2)$ is the distance along the tangent at either of the preshapes (see Figure 7.4). As can be seen, for objects with similar shapes, $\rho \approx d_F \approx d_p$.

For a collection of n registered configurations or profiles, the GPA registers or superimposes all the n objects by finding scaling factors $\beta_i \in \mathbb{R}$, rotation matrices $\mathbf{\Gamma}_i \in SO(m)$, and m-dimensional translation vectors $\boldsymbol{\gamma}_i$, $i = 1, \ldots, n$, such that they minimize the sum of squared full Procrustes distances between all objects:

$$G(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n) = \min_{\beta_i, \mathbf{\Gamma}_i, \gamma_i} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} || \beta_i \mathbf{X}_i \mathbf{\Gamma}_i + \mathbf{1}_k \boldsymbol{\gamma}_i' - (\beta_j \mathbf{X}_j \mathbf{\Gamma}_j + \mathbf{1}_k \boldsymbol{\gamma}_j') ||^2, \quad (7.7)$$

where $\mathbf{1}_k$ is a vector of k 1's. The resulting registered configurations are called the *full Procrustes fits*, defined as

$$\mathbf{X}_i^p = \hat{\beta}_i \, \mathbf{X}_i \hat{\mathbf{\Gamma}}_i + \mathbf{1}_k \hat{\boldsymbol{\gamma}}'_i, \; i = 1, \ldots, n. \quad (7.8)$$

The mean shape of the n objects is simply the average of the n configurations, namely, $\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^p$.

The minimization (Equation 7.7) needs to be subjected to a constraint that limits the scaling done, otherwise the optimal value of G will be zero. One such restriction is to use a constraint on the size of the mean shape, $S(\hat{\boldsymbol{\mu}}) = 1$, where the size of any configuration $\mathbf{X}$ is defined as $S(\mathbf{X}) = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{m}(X_{ij} - \bar{X}_j)^2} = \|\mathbf{CX}\|$, where $\mathbf{C} = \mathbf{I}_k - k^{-1}1_k1_k'$, $\bar{X}_j = \frac{1}{n}\sum_{i=1}^{k} X_{ij}$ and X_{ij} is the *j*th coordinate of the *i*th point in the configuration. Another common constraint, used in what follows, is to make the average of the squared sizes of the registered configurations $\mathbf{X}_i^p$ given by Equation 7.8 equal to the average of the squared sizes of the original objects:

$$\frac{1}{n}\sum_{i=1}^{n} S^2(\mathbf{X}_i^p) = \frac{1}{n}\sum_{i=1}^{n} S^2(\mathbf{X}_i)\,. \tag{7.9}$$

The GPA, as developed by Gower (1975) and Ten Berge (1977), proceeds as follows to solve Equation 7.7 subject to Equation 7.9:

1. Center (but do not scale) the configurations $\mathbf{X}_1, \ldots, \mathbf{X}_n$ by initially defining

   $$\mathbf{X}_i^p = \mathbf{H}\mathbf{X}_i,\ i = 1, \ldots, n$$

   [alternatively, we can define $\mathbf{H}'\mathbf{H}\mathbf{X}_i = \mathbf{C}\mathbf{X}_i = \mathbf{X}_i^p$ and the resulting matrices will be $k \times m$; note that $\mathbf{X}_i^p$ as defined above is instead $(k-1) \times m$]
2. Let $\overline{\mathbf{X}}_{(i)} = \frac{1}{n-1}\sum_{j \neq i} \mathbf{X}_j^p$, $i = 1, \ldots, n$. These are the "jackknifed" average shapes excluding object i.
3. Do a Procrustes fit (rotation only) of the current $\mathbf{X}_i^p$'s onto $\overline{\mathbf{X}}_{(i)}$. This yields rotation matrices $\hat{\boldsymbol{\Gamma}}_i$, from which we let

   $$\mathbf{X}_i^p \leftarrow \hat{\boldsymbol{\Gamma}}_i \mathbf{X}_i^p,\ i = 1, \ldots, n\,.$$

   We repeat steps 2 and 3 for all i.
4. Compute the $n \times n$ correlation matrix $\boldsymbol{\Phi} = corr(\mathbf{X}_v)$, where

   $$\mathbf{X}_v = [\mathrm{vec}(\mathbf{X}_1^p)\mathrm{vec}(\mathbf{X}_2^p)\ldots\mathrm{vec}(\mathbf{X}_n^p)]\,,$$

   where vec($\mathbf{X}$) returns a vector in which we stack the columns of matrix $\mathbf{X}$ on top of each other. Note we stack all the m dimensions together.
5. Let $\phi = (\phi_1, \ldots, \phi_n)'$ be the eigenvector of $\boldsymbol{\Phi}$ corresponding to its largest eigenvalue. Then set

   $$\hat{\beta}_i = \sqrt{\frac{\sum_{j=1}^{n}\left\|\mathbf{X}_j^p\right\|^2}{\left\|\mathbf{X}_i^p\right\|^2}}\phi_i,\ i = 1, \ldots, n$$

   and let $\mathbf{X}_i^p \leftarrow \hat{\beta}_i \mathbf{X}_i^p$. The algorithm repeats steps 2–5 until there is convergence.

The algorithm is guaranteed to converge (in the sense that the fitted $\mathbf{X}_i^p$ cease to vary as i increases), usually in just a few iterations (Ten Berge 1977). The exact solution to the Procrustes registration problem between two objects $\mathbf{X}_1$ and $\mathbf{X}_2$ required in step 3, implies finding $\mathbf{\Gamma} \in SO(m)$ that minimizes $d_p(\mathbf{X}_1, \mathbf{X}_2)$ (see Equation 7.5) for $\mathbf{X}_1 = \mathbf{X}_i^p$ and $\mathbf{X}_2 = \overline{\mathbf{X}}_{(i)}$, $i = 1, \ldots, n$. The exact solution to this problem is well known in both the statistics (Jackson 2003) and the computer vision (Horn *et al.* 1988) fields and is given by $\hat{\mathbf{\Gamma}} = \mathbf{UV}'$, where $\mathbf{U}$ and $\mathbf{V}$ are obtained from the singular value decomposition $\mathbf{Z}_2'\mathbf{Z}_1 = \mathbf{V\Lambda U}$. An important implementation detail of singular value decomposition for shape analysis is that to ensure we have $\det(\hat{\mathbf{\Gamma}}) = +1$ and hence a rotation matrix (as opposed to –1 and a reflection matrix), we can make instead $\hat{\mathbf{\Gamma}} = \mathbf{URV}'$, where $\mathbf{R}$ is the identity matrix except for the last diagonal element, for which we use $\det(\mathbf{UV}')$.

The GPA as described assumes the statistical model

$$\mathbf{X}_i = \beta_i(\boldsymbol{\mu} + \mathbf{E}_i)\mathbf{\Gamma}_i + \mathbf{1}_k \boldsymbol{\gamma}_i',\ i = 1, \ldots, n, \qquad (7.10)$$

where $\boldsymbol{\mu}$ is the mean shape of the objects and the $k \times m$ matrix of errors $\mathbf{E}_i$ is such that $\mathrm{vec}(\mathbf{E}_i) \sim (\mathbf{0}, \sigma^2 \mathbf{I}_{km \times km})$, where $\mathbf{0}$ is a vector of *km* zeroes and $\mathbf{I}_{km \times km}$ is the $km \times km$ identity.

The model then assumes *isotropic variance*, *i.e.*, the variance is the same at each landmark and at each of the m coordinates. Modification of the GPA for the case of a general covariance matrix of the errors $\mathbf{\Sigma}$ requires a straightforward modification of the definition of the d_F distances minimized in Equation 7.7 that accounts for $\mathbf{\Sigma}$. However, given that in general $\mathbf{\Sigma}$ is unknown and needs to be estimated, there is no known registration algorithm which guarantees convergence in the nonisotropic case. Common practice is to initially set $\mathbf{\Sigma} = \mathbf{I}$, run the GPA, then estimate $\mathbf{\Sigma}$ with

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{vec}(\mathbf{X}_i^p - \hat{\boldsymbol{\mu}})\mathrm{vec}(\mathbf{X}_i^p - \hat{\boldsymbol{\mu}})',$$

run the GPA again with the squared full Procrustes distances in Chui and Rangarajan (2000) replaced by the Mahalanobis squared distance $\mathrm{vec}(\mathbf{X}_i)'\hat{\mathbf{\Sigma}}^{-1}\mathrm{vec}(\mathbf{X}_j)$, and iterate this process (but convergence is not guaranteed).

Equation 7.10 implies that each object results from the rotation, scaling, and translation of the mean shape in the presence of random noise, *i.e.*, similarity transformations of the mean shape observed with noise generate the observed profiles of the objects.

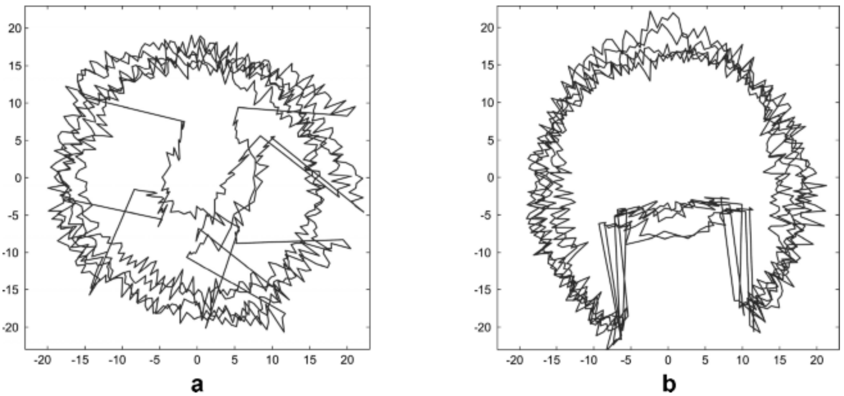


Figure 7.5 Example of generalized Procrustes algorithm (GPA) registration applied to the contours of ten simulated "circular notched" parts, each with $k = 200$ (labeled) landmarks: **a** original, unregistered shapes, and **b** shapes registered using the GPA

Example. Generalized Procrustes registration. Suppose ten cylindrical parts are manufactured. The parts have the same geometric specifications and were produced under homogeneous conditions. It is of interest to study the variability of the shapes (more on this below). The part design has a "notch", typical in parts that are used for assemblies. The ten two-dimensional shapes measured correspond to orthogonal contours obtained using a CMM at a fixed distance from the cylinder's origin. Each shape has $k = 200$ landmark measurements. We assume the landmark matching problem does not exist, so the labels between shapes correspond to each other. The original orientation of the parts, however, differs, and registration is necessary. Figure 7.5 shows ten such simulated contours before and after registration using the GPA (evidently, the noise has been exaggerated with respect to what actual measurements of real parts would look like).

7.3.3 Tangent Space Coordinates

Once n configurations or profiles have been registered using the GPA, the mainstream of the SSA literature (see, *e.g.*, Dryden and Mardia 1998; Goodall 1991; Adams *et al.* 2004) recommends that further statistical analysis of shape variability and any desired inferences be made on the basis of the resulting registered shapes $\mathbf{X}_i^p$ using the full Procrustes distances from the mean shape (or *pole*), called the *tangent space coordinates*. This is suggested in contraposition to working with the Procrustes distances which are not linear. A principal component analysis (PCA) on the tangent space coordinates is then recommended to better understand the directions in which the shapes are varying the most.

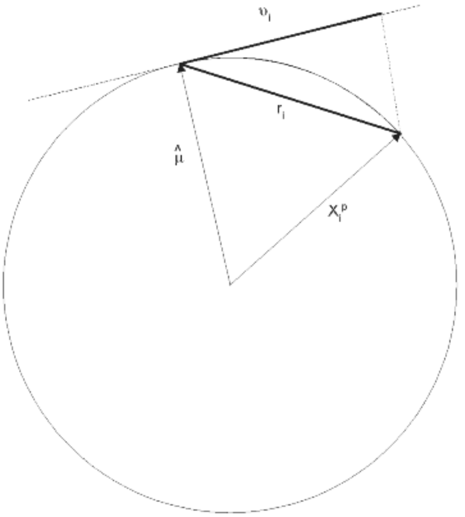


Figure 7.6 Tangent coordinates $\mathbf{v}_i$ and approximate tangent coordinates (secants) $\mathbf{r}_i = \mathrm{vec}(\mathbf{X}_i^p - \hat{\boldsymbol{\mu}})$

For a preshape $\mathbf{X}_i^p$ and mean shape $\hat{\boldsymbol{\mu}}$, the tangent coordinates v_i are the distances along the tangent at the mean shape corresponding to the projection of $\mathbf{X}_i^p$ on $\hat{\boldsymbol{\mu}}$ and are given by

$$\mathbf{v}_i = \left[\mathbf{I}_{(k-1)m} - \mathrm{vec}\left(\frac{\hat{\boldsymbol{\mu}}}{\|\hat{\boldsymbol{\mu}}\|} \right) \mathrm{vec}\left(\frac{\hat{\boldsymbol{\mu}}}{\|\hat{\boldsymbol{\mu}}\|} \right)' \right] \mathrm{vec}\left(\frac{\mathbf{X}_i^p}{\|\widehat{\mathbf{X}_i^p}\|} \right), \ i = 1, \ldots, n. \quad (7.11)$$

The tangent coordinates are $(k-1)m$ vectors. If the centered configurations are used, then $\mathbf{v}_i$ is a km dimensional vector.

An alternative approach which is close to the tangent coordinates if the preshapes are not too different from the mean shape (see Figure 7.6) is to use the *Procrustes residuals* $\mathbf{r}_i$ defined by

$$\mathbf{r}_i = \mathrm{vec}(\mathbf{X}_i^p - \hat{\boldsymbol{\mu}}), \ i = 1, \ldots, n,$$

that is, we work with the secants instead of the tangents. Again, for small differences about the mean, the conclusions of the analysis would be very similar. Regardless of how one computes the tangent coordinates, either using Equation 7.11 or using the approximate tangent coordinates $\mathbf{v}_i \approx \mathbf{r}_i$, the mainstream approaches to SSA recommend using a PCA on the $\mathbf{v}_i$'s (Goodall 1991; Adams *et al.* 2004). The theoretical justification for this recommendation comes from work by Kent and Mardia (2001), who have shown that an isotropic distribution of the landmarks results in an isotropic distribution in the tangent space (given that small changes in a configuration matrix $\mathbf{X}$ induce an approximately linear change in the tangent coordinates $\mathbf{v}$), and, hence, PCA in tangent space is valid for shape analysis.

Let $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n]$ be all the tangent coordinates of all n objects under study. An estimate of the covariance matrix $\mathrm{cov}(\mathbf{V}')$, giving the between-shape variances and covariances at the landmarks, is given by

$$\mathbf{S}_v = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}_i - \overline{\mathbf{v}})(\mathbf{v}_i - \overline{\mathbf{v}})',$$

where $\overline{\mathbf{v}}$ is the average of the $\mathbf{v}_i$'s. This is a $(k-1)m \times (k-1)m$ matrix if the preshapes are only scaled and a $km \times km$ matrix if the centered preshapes are used instead. In the first case, the rank of this matrix is $p = M = (k-1)m - 1 - m(m-1)/2$ and in the latter case the rank is $p = M + 1$, since the mean is not lost. Let $\{\lambda_j\}_{j=1}^{p}$ and $\{\mathbf{e}_j\}_{j=1}^{p}$ be the p eigenvalues and eigenvectors of $\mathbf{S}_v$. Dryden and Mardia (1998) suggested computing

$$\mathbf{v}(c, j) = \overline{\mathbf{v}} + c\sqrt{\lambda_j}\ \mathbf{e}_j,\ j = 1, \ldots, p$$

for several values of c, say, for $-6 < c < 6$.

One of the great advantages of shape analysis methods is visualization, as it takes place in a space that preserves the geometry of the objects. To visualize the principal components of the tangent coordinates, Dryden and Mardia (1998) suggested plotting

$$\mathrm{vec}(\mathbf{X}_I) = \begin{bmatrix} \mathbf{H}' & \mathbf{0} \\ \mathbf{0} & \mathbf{H}' \end{bmatrix} [\mathbf{v}(c, j) + \mathrm{vec}(\hat{\boldsymbol{\mu}})] \qquad (7.12)$$

for all principal components j and for all multiplies c. These are the coordinates of the original shapes where the (registered) objects exist, and indicate the directions in which the principal components indicate movement – variability – around the mean shape (if the $\mathbf{v}_i$'s are km-dimensional vectors, there is no need to premultiply the block matrix of Helmert matrices). Just as in regular PCA, the percentage of variation explained by the *j*th principal component is given by $100\lambda_j / \sqrt{\sum_{j=1}^{p} \lambda_j}$. *Once the tangent coordinates have been computed, multivariate analysis techniques can be applied in the usual way* until the point where visualization using Equation 7.12 is necessary.

Example. PCA of the circular notch shape data. Consider the ten shapes shown in Figure 7.5. These shapes were simulated by superimposing sinusoidal variability along the circle at a second harmonic (inducing a bilobed shape) and inducing

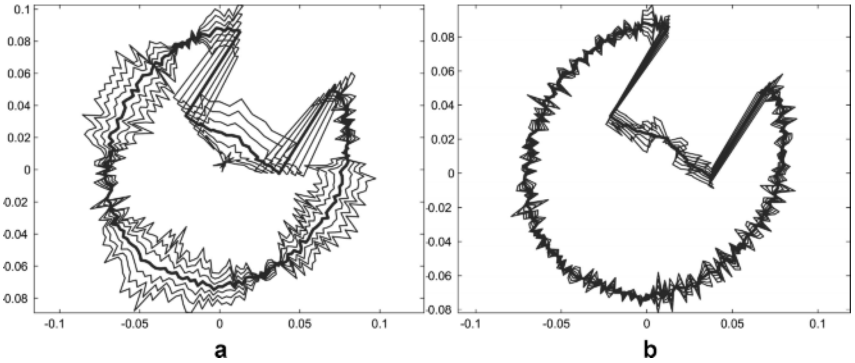


Figure 7.7 Example of principal component analysis applied to the data of the ten circular notched parts: **a** the first principal component detects a sinusoidal variation, and accounts for 56.7% of the variability, and **b** the second principal component, which corresponds to variation in the depth of the notch, accounts for 8.3% of the variability. In these plots, values of $c \in -(3,3)$ were used. The *dark line* is the mean shape

variability in the depth of the notch. Additional random normal bivariate variability was added at each landmark, which masks the first two sources of variability in such a way that they are not obvious to the eye. These sources of variability where added to demonstrate the power of PCA in the tangent space. Figure 7.7 shows the first two principal components in this example, which together account for more than 65% of the variability. Note how the first component is precisely the simulated bilobed shape and the second component refers to the depth of the notch. The remaining principal components do not show any obvious pattern.

Performing a PCA on the tangent coordinates is of value when one is interested in analyzing how the *variability* of the shapes behaves around the mean. To analyze the effect of factors (varied during an experiment) on the *mean* shape (as required when conducting manufacturing experiments that may improve the shapes of parts produced by a process) one needs to perform an analysis of variance. This was first discussed by Goodall (1991) (see also Dryden and Mardia 1998) for the one-way case and was studied in the two-way layout case, with an application in manufacturing, by Del Castillo and Colosimo (2009).

7.4 Further Work

The methods presented in this chapter assume all parts contain the *same* number of corresponding landmarks, or locations of interest. If there is a different number of landmarks between two objects, the iterative closest point (ICP)) algorithm (Besl and McKay 1992; Zhang 1998) has been proposed to obtain the same number of corresponding landmarks between the parts. If the landmarks do not correspond, a matching algorithm such as the context-labeling algorithm in Section 7.2 could be applied after the ICP algorithm. An alternative to the use of the ICP algorithm

(Belongie *et al.* 2002) is to simply add dummy landmarks to the smallest landmark matrix to get $\max(k_1, k_2)$ landmarks, and assign a large cost between these points and all others (this is also a mechanism to handle outlier landmarks, since they would be matched to the dummy points).

A matching algorithm notably different from the one presented in Section 7.2 was proposed recently by Green and Mardia (2006). It also applies to the case of two objects. Another possibility is to use the two-dimensional context histograms, but to use a statistic other than the χ^2 used here, to measure distances (costs) between two multivariate distributions, *e.g.*, a two-dimensional Kolmogorov–Smirnov test or other recent alternatives (*e.g.*, that in Rosenbaum 2005). Such an approach would still use the weighted matching linear programming formulation presented here, but with a different way of getting the cost matrix $\mathbf{C}$. Even in the Belongie *et al.* (2002) approach, it is not clear how to best scale the $\mathbf{X}$ matrices, how many bins to use in each dimension, or what is the best way to measure angles for differently oriented objects in order to achieve effective rotation invariance. An interesting embellishment of the landmark matching algorithm (Belongie *et al.* 2002) is to iterate the matching algorithm with an algorithm for the estimation of the registration transformation between the objects. This may result in better matching (and hence registration) because the initial matching may be sensitive to the different orientations of the parts due to the ambiguities mentioned earlier about defining the histogram resolution. These authors suggested using thin plate spline transformations, popular also in the area of morphometrics, as opposed to the GPA considered here. A similar iterative procedure could be attempted with the context labeling algorithm and the GPA applied iteratively. A recent description of the matching problem from a computer-vision perspective is given in the book by Davies *et al.* (2008).

As mentioned earlier, a generalization of two-object matching methods to the case of n objects is desirable, since once labeled (corresponding) landmarks are available (assuming there is the same number of landmarks in each object), the SSA methods presented herein can be implemented. The advantages and disadvantages of the "context labeling" method compared with the Green and Mardia (2006) approach need to be investigated.

In this chapter we did not discuss tests for comparing the mean shapes between two or more populations, which can be done using analysis of variance methods applied to the shapes. For more information on this topic, see Del Castillo and Colosimo (2009).

Most of the work on SSA has focused on two-dimensional shapes. Extensions to the three-dimensional case are evidently practical (but the landmark matching problem becomes more difficult). The context labeling approach presented here was extended recently by Frome *et al.* (2004) to the three-dimensional case using three-dimensional histograms. The implementation details mentioned above remain and need to be studied. For three-dimensional objects, the GPA and the PCA can be used without any change, but visualization of the PCAs is challenging if k (number of of landmarks) is large.

Finally, some authors (*e.g.*, Lele and Richtsmeier 2001) have proposed using the interlandmark Euclidean distance matrix $[d_{ij}]$ to make inferences on the shapes of objects, with application to testing for the difference between shapes. Lele and Richtsmeier (2001) suggested using the GPA to estimate the mean shape, but there is debate about how to estimate the covariance matrix of the landmarks in the nonisotropic case. This series of methods do not have an easy way to visualize the results, and require more information ($\binom{k}{2}$ distances instead of *km*), although this information is implicit in the $k \times m$ matrix $\mathbf{X}$. In addition, there seems to be no counterpart to the PCA of variability in distance-based methods. There is considerable debate about which method is more powerful to detect differences in shapes, and it is of interest to compare distance-based methods with those studied in Del Castillo and Colosimo (2009) for a variety of shapes of relevance in manufacturing, since the power of these methods appears to depend on the shape in question. Dryden and Mardia (1998) presented a good overview of distance-based methods.

Appendix: Computer Implementation of Landmark Matching and the GPA and PCA

MATLAB programs that perform the computations required for the context labeling algorithm in Section 7.2 and for the GPA, including visualization of PCAs, were written for this research and can be downloaded from

http://www2.ie.psu.edu/Castillo/research/EngineeringStatistics/software.htm.

The programs posted contain several programs for SSA. Two of the programs are related to what is discussed in the present paper: `ContextLabeling.m`, which implements the context labeling algorithm presented in Section 7.2 for two two-dimensional objects, and `GPA23.m`, which implements the GPA (assuming isotropic variance), and performs, if desired, the PCA in tangent space, including the corresponding visualization.

Acknowledgements This research was partially funded by the National Science Foundation grant CMI-0825786 and a grant from the Ministry of Science of Italy.

References

Adams DC, Rohlf FJ, Slice DE (2004) Geometric morphometrics: ten years of progress following the “revolution”. Ital J Zool 71:5–16

ASME Y14.5M (1994) Dimensioning and tolerancing. American Society of Mechanical Engineers, New York

Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24:509–522

Besl PJ, McKay ND (1992) A method for registration of 3-D shapes. IEEE Trans Pattern Anal Mach Intell 14:239–256

Chui H, Rangarajan A (2000) A new algorithm for non-rigid point matching Proc IEEE Conf Comput Vis Pattern Recognit 44–51

Colosimo BM, Pacella M, Semeraro Q (2008) Statistical process control for geometric specifications: on the monitoring of roundness profiles. J Qual Technol 40:1–18

Davies R, Twining C, Taylor C (2008), Statistical models of shape, optimisation and evaluation. Springer, London

Del Castillo E, Colosimo BM (2010) Statistical shape analysis of experiments for manufacturing processes. Technometrics (in press)

Dryden IL, Mardia KV (1998) Statistical shape analysis. Wiley, Chichester

Frome A, Huber D, Kolluri R, Bulow T, Malik J (2004) Recognizing objects in range data using regional point descriptors. In: Proceedings of the 8th European conference on computer vision, vol 3, pp 224–237

Gold S, Rangarajan A, Lu C-P, Pappu S, Mjolsness E (1998) New algorithms for 2D and 3D point matching: pose estimation and correspondence. Pattern Recognit 31:1019–1031

Goodall C (1991) Procrustes methods in the statistical analysis of shape. J R Stat Soc B,53:285–339

Gower JC (1975) Generalized Procrustes analysis. Psychometrika 40:33–51

Green PJ, Mardia KV (2006) Bayesian alignment using hierarchical models, with applications in protein bioinformatics. Biometrika 93:235–254

Horn BKP, Hilden HM, Negahdaripour S (1988) Closed-form solution of absolute orientation using orthonormal matrices. J Opt Soc Am A 5:1127–1135

Jackson JE (2003) A user's guide to principal components. Wiley, New York

Kang L, Albin SL (2000) On-line monitoring when the process yields a linear profile. J Qual Technol 32:418–426

Kendall DG (1984) Shape manifolds, Procrustean metrics, and complex projective spaces. Bull Lond Math Soc 16:81–121

Kendall DG (1989) A survey of the statistical theory of shape Stat Sci 4:87–89

Kent JT, Mardia KV (2001) Shape, Procrustes tangent projections and bilateral symmetry. Biometrika 88:469–485

Klingenberg CP, McIntyre GS (1998) Geometric morphometrics of developmental instability: analyzing patterns of fluctuating asymmetry with Procrustes methods. Evolution 52:1363–1375

Lele SR, Richtsmeier JT (2001) An invariant approach to statistical analysis for shapes. Chapman & Hall/CRC, Boca Raton

Papadimitrou CH, Steiglitz K (1982) Combinatorial optimization, algorithms and complexity. Prentice Hall, Englewood Cliffs

Rosenbaum PR (2005) An exact distribution-free test comparing two multivariate distributions based on adjacency. J R Stat Soc B 67:515–530

Ten Berge JMF (1977) Orthogonal Procrustes rotation for two or more matrices. Psychometrika 42(2):267–276

Woodall WH, Spitzner DJ, Montgomery DC, Gupta S (2004) Using control charts to monitor process and product quality profiles. J Qual Technol 36:309–320

Zelditch ML, Swiderski DL, Sheets HD, Fink WL (2004) Geometric morphometrics for biologists, a primer. Elsevier, San Diego

Zhang Z (1998) Iterative point matching for registration of free-form curves. Reports de recherche no. 1658. IRIA, Sophia Antipolis

Part III
Impact on Statistical Process Monitoring

Chapter 8
Statistical Quality Monitoring of Geometric Tolerances: the Industrial Practice

Bianca Maria Colosimo and Massimo Pacella

Abstract This chapter shows how traditional approaches for statistical process control can be used for monitoring geometric tolerances. These approaches can be useful to quickly detect changes in the manufacturing process. In particular, two simple methods are presented. The first one uses standard variable control charts for monitoring over time the error associated with the geometric tolerance at hand. The second approach designs a control band around the mean shape of the feature associated with the geometric tolerance. Both approaches are shown with reference to the problem of monitoring a roundness form tolerance. Given their ease of use, the approaches are viable solutions for industrial practice.

8.1 Introduction

Geometric features are machined by processes that may experience changes due to the material machined, improper setup, errors of the operator, wear or sudden changes of the machine conditions, *etc.* Usually, these changes cause deteriorated process performance, *i.e.*, the process may produce an increased number of nonconforming or defective items. This is why statistical quality monitoring – also

B.M. Colosimo
Dipartimento di Meccanica, Politecnico di Milano,
Via la Masa 1, 20156 Milan, Italy,
e-mail: biancamaria.colosimo@polimi.it

M. Pacella
Dipartimento di Ingegneria dell'Innovazione, Università del Salento,
Via per Monteroni, 73100 Lecce, Italy,
e-mail: massimo.pacella@unisalento.it

known as statistical process control (SPC) – can be effectively considered as a way to quickly detect unusual states of the manufacturing process by issuing an alarm.

Traditional approaches for quality monitoring consists in designing and using control charts when the quality characteristic of interest is well represented by one (or more) variable. As an example, a dimensional tolerance, such as a diameter, can be modeled as a random variable and then monitored by using a control chart.

However, when the quality of a manufactured product is related to a geometric tolerance, instead of a dimensional one, the problem is how quality monitoring can be implemented. This chapter presents two simple approaches that can be used for this aim. The first approach consists in summarizing all the information contained in the points measured on the shape of interest in just one synthetic variable. This variable usually measures the geometric form error as the (maximum) distance between the actual profile and the ideal geometry. Then a control chart of the estimated geometric error is implemented for quality monitoring.

A second approach, which is called a location control chart, consists in designing a control region around the ideal or mean shape observed on a set of profiles. This control region is defined by two limits that can be computed by virtually designing a control chart for points observed at any given location and controlling the whole false-alarm rate of this set of control intervals by using Bonferroni's inequality. An alarm is issued when at least one point, in the whole set of data observed in a profile, exceeds the control limits.

Both methods presented are based on Shewhart's traditional control chart, which is briefly summarized in Section 8.2. This section can be skipped by a reader who is already familiar with SPC tools. In Section 8.3, the issue of roundness form error is discussed. In Section 8.4, the control chart for the geometric form error is presented, while in Section 8.5 the location control chart is discussed. In Sections 8.4 and 8.5, a real case study concerning roundness form error is used as a reference. This case study is described in detail in Chapter 11.

8.2 Shewhart's Control Chart

SPC may be considered to have begun with the pioneering work of Walter A. Shewhart, an engineer at Bell Telephone Laboratories, where he was faced with the issue of obtaining good quality in the mass production of interchangeable equipment for the rapidly expanding telephone system. Shewhart's ideas (Shewhart 1931) are still relevant today. The most important technique he developed, *i.e.*, the control chart, is nowadays widely used for quality control of manufacturing processes and services. This section provides a brief overview of Shewhart's approach, while a more general description can be found in standard texts (Alwan 2000; Montgomery 2004; Ryan 2000).

The baseline idea of Shewhart's approach is that in any production process, no matter how well designed it is, there exists a certain amount of natural variation in the outcomes. This variability is always present as it results from a large number of so-called *common causes* (*i.e.*, natural causes) which are to some extent unavoidable (or can be removed but at prohibitive costs).

The process may also be affected by external sources of variation, which are upsetting its natural functioning. Sources of variability that are not part of the process and that can occur only accidentally are called *special causes* (*i.e.*, assignable causes) of variation. The presence of special causes may lead to excessive variation in process outcomes, possibly resulting in quality loss and customer complaints. In such cases, quality improvement is possible by detection and removal of special causes of variation. Since a special cause is not inherently part of the process, it can usually be eliminated without revising the process itself. In many cases, the removal of a special cause of variation is possible. An operator can be instructed to recognize and remove it.

Therefore, it is essential to be able to distinguish between situations where only common causes of variation affect the outcomes of a process and situations where special causes are also present. A tool for supporting the operator's decision is needed for this purpose, since the effect of a possible special cause can be hidden in the variation due to common causes. Shewhart developed the *control chart* for this purpose and gave a rationale for using such a tool in process monitoring. The control chart is intended to monitor a process by issuing an alarm signal when it is suspected of going out of control.

The baseline idea of Shewhart (1931) is that if only common causes of variation are present, the manufacturing process should be statistically in control, *i.e.*, the outcomes should be predictable according to a given statistical model. The statistical predictability of a process that is in control is the basis for the control chart. It does not mean that there is no variation, or that there is a small variation. Simply, it means that the outcomes are predictable in a statistical sense. For example, based on previous observations, it is possible for a given set of limits to determine the probability that future observations will fall within these limits.

In order to apply control charting, data are collected in samples, usually referred to as *subgroups*. Statistics of interest (such as the mean or the standard deviation) are computed for each sample in order to summarize the information contained within each subgroup. These statistics are then plotted on a graph and compared with limits, which represent the bandwidth of the natural variation due to common causes. The idea is that as long as all statistics are within the control limits, it is reasonable to assume that the underlying process is statistically in control.

Given this monitoring strategy, a control chart involves measuring a quality characteristic of interest at regular intervals, collecting n items each time and plotting one or more sample statistics of the quality parameter against time. In practice, a new point is plotted each time a subgroup of items is measured.

Figure 8.1 shows an example of a control chart for individual measurements, *i.e.*, when the sample size n is equal to 1 and hence the single datum observed at each time period is considered in the control chart.

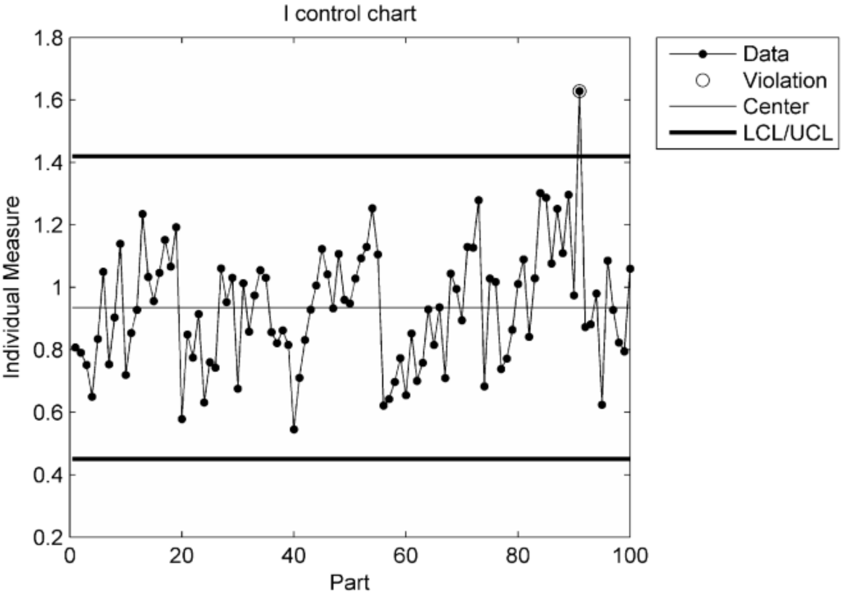


Figure 8.1 An example of a control chart for subgroups of size $n = 1$, *i.e.*, individual measurements. *LCL* lower control limit, *UCL* upper control limit

Points outside the control limits are called out-of-control signals and indicate the likely presence of special causes of variation that are affecting the process. The presence of special causes implies that there could be some source of variation that causes the measurements to be variable, differently from what can be attributed to the effect of natural causes only. If an out-of-control signal is observed, action is required to track down the special cause that is responsible. Owing to the random nature of the observations, there is also the possibility that an out-of-control signal is encountered while the process is statistically in control and hence when there is no special cause to be identified and removed. This is called a false out-of-control signal or false alarm.

It is worth noting that control limits are not directly related to customer's specifications (*i.e.*, quality requirements) for the process outcomes. In fact, control limits just indicate the magnitude of the natural variability of the sample characteristic used to monitor the process. They are based on the relevant sampling distributions of process outcomes when only natural causes are present, while, in general, quality requirements are not related to the actual performance of the process. Besides, quality specifications are always related to single products, whereas control limits can be computed for any arbitrary statistic of the process outcome (such as the sample mean, range, or standard deviation).

Even when a control chart for individual measurements is considered, *i.e.*, a single process outcome at each time period is plotted on the chart, if the variation due to common causes is relatively large, all points on the control chart may be within the control limits, but the process outcomes might fail to meet quality specifications. Furthermore, the presence of special causes does not necessarily mean that there is large variation, or that the specifications are not met.

A control chart must be sensitive enough to detect the effect of special causes of variation, but also must not generate too many false alarms. In practice, a balance between these two objectives can be achieved by a proper setup of the control chart, *i.e.*, by determining the appropriate width of the control limits. In Shewhart's approach, the control limits are usually set equal to three standard deviations from the center line, which represent in turn the expected value of the plotted statistics. Considering the assumption of normally and independently distributed points plotted on the chart, this control limit position corresponds to an expected false-alarm signal in about 370 points on average.

8.2.1 Two Stages in Control Charting

Consider a process which has been set up in order to operate stably and properly, *i.e.*, in its (presumable) in-control state. A set of observations collected on this process are first analyzed and then used to design a control chart. The aim is both to evaluate the stability of the process and to estimate the in-control-state's parameters. In this phase, also called phase I, control charts are used offline to determine retrospectively from the set of collected data whether the process is indeed in a state of statistical control.

In phase I, the control chart is used as an aid to the analyst to screen out out-of-control data from the set of collected observations so that a set of presumably in-control process data can be obtained to model the distribution of the monitoring statistic (such as the mean, range, or standard deviation). To this aim, the collected data are used to set up a set of initial trial control limits for the monitoring statistic. If out-of-control signals are observed, the process is investigated to see if there exist any special causes to explain these out-of-control signals. If indeed some special causes are found, then the samples which produce the out-of-control signals are removed from the data set. Then, the remaining samples are used to re-estimate control limits. This procedure should be repeated until no out-of-control signals are generated, or when underlying assignable causes cannot be found. In the latter case, even if these points exceed the limits simply by chance or because of some uncovered assignable causes, to be conservative, one may choose to discard them to avoid potential contamination in the data set.

When no out-of-control signals are eventually generated, at the end of phase I a data set is available which provides information concerning the variability that can be attributed to common causes of variation of the process in its in-control state. Thus, reliable control limits of the control chart are established for online process monitoring in subsequent phase II.

In phase II (also called the operating phase), control charts are used for testing whether the process remains in control when future subgroups are drawn. During this phase, the goal is to monitor the online data and quickly detect changes in the process from the baseline model established in phase I.

Since the target of control charting is detecting process changes as quickly as possible, the performance of a monitoring approach is usually described by the run-length distribution, where the run length is the number of samples taken before an out-of-control signal is given. In particular, the average run length is often used as a performance index.

8.3 Geometric Tolerances: an Example of a Geometric Feature Concerning Circularity

In order to exemplify the issue of geometric tolerance monitoring, the roundness of a circular feature is assumed as a reference from now on. A circular feature in a component such as a shaft or a hole is one of the most frequently encountered features in manufacturing, because functionality of mechanical parts is very often related to a proper rotation. According to the ISO/TS 12181 (2003) standard, a circular profile is the line extracted on a cross section of a surface of revolution. For a circular feature, the circularity error attained is represented by the so-called *out-of-roundness* (OOR) value. This important geometric characteristic can be estimated in a number of different ways.

Conventionally, the profile of a circular feature can be measured diametrically by an operator. The difference between the maximum and the minimum diameters measured is used as an indication of the OOR for that circular feature. Obviously, this diametric measurement of the OOR can be deceptive in actual applications. Alternatively, the profile of a circular feature can be measured by tracing its perimeter with a stylus or a probe. One approach is to revolve the item against a displacement transducer. This is the basic approach of the so-called rotondimeter, *i.e.*, a measuring device equipped with a stylus which is able to measure a circular profile with a sufficiently high number of points in a relatively short time. Such a device can be exploited to measure axially symmetric workpieces, *i.e.*, workpieces which can be easily rotated around their axes. A different practice may often be encountered in industry. It consists in measuring a circular profile by sampling several points on its perimeter by either a mechanical probe or an optical probe on a coordinate measuring machine (CMM). CMMs are the most-general-purpose devices nowadays available in industry for measuring any type of physical geometric characteristic of an object.

Irrespective of the specific device exploited, the OOR value must be computed as the difference between the maximum and the minimum radial distances of the manufactured feature with respect to a predetermined center, which is the center of the so-called *substitute geometry*. Geometrically, this corresponds to finding two concentric circles, one circumscribing and one inscribing the profile sampled on the manufactured feature. The OOR is then estimated by the width of the annulus determined by these two concentric circles. CMMs, which are controlled by a computer, estimate algorithmically the geometric error on the basis of a finite number of points sampled on the manufactured feature. The geometric best fit is a numerical transformation between measurements and their substitute geometry, *i.e.*, the actual center of the substitute geometry.

Several methods can be implemented in practice. The differences in methods (and results) essentially depend upon the algorithm used to determine the common center of the circumscribed circle and of the inscribed one. Four procedures are commonly used: the minimum zone (MZ), the least squares (LS), the maximum-inscribed circle (MIC), and the minimum-circumscribed circle (MCC).

1. The MZ center is that for which the radial difference between two concentric circles that just contain the measured points is a minimum. The mathematical problem of finding the MZ center can be stated as a Chebychev problem. This is a rather complicated nonlinear problem and exact algorithms for solving it are not readily available.
2. The LS center is that of a circle from which the sum of the squares of the radial ordinates between this circle and the measured points is a minimum.
3. The MIC center is that of the largest circle that can be inscribed within the set of measured points.
4. The MCC center is that of the smallest circle that just contains all the measured points.

Algorithms for calculating these centers exist, with varying computational complexity (Carr and Ferreira 1995a, b; Gass *et al.* 1998). As an example, Figure 8.2 depicts a sampled roundness profile, the corresponding MZ OOR value (the radius distance between two concentric circles which contain the sampled profile), as well as the resulting substitute geometry.

Two procedures for determining the center of the substitute geometry are the MZ and the LS methods. These procedures are assumed as a reference henceforth since the former best conforms to the ISO standards for form tolerances, while the latter is the one most frequently encountered in actual applications. The MZ algorithm looks for a couple of geometric nominal features (*e.g.*, a couple of concentric circles for roundness) at the minimum distance that includes the whole set of measurement points. The method always consists of the minimum deviation, given a set of measurement data. However, it is very sensitive to asperities.

Different procedures to estimate the couple of minimum-distance nominal features have been reported in the literature. In particular, Carr and Ferreira (1995a, b) formulated roundness as nonlinear optimization problems, which are then transformed into a series of linear problems. Their linear program models are derived from the original optimization problem modeled as a constrained nonlinear programming problem with a linear objective function. This model can be effectively implemented by the library of optimization subroutines available in a

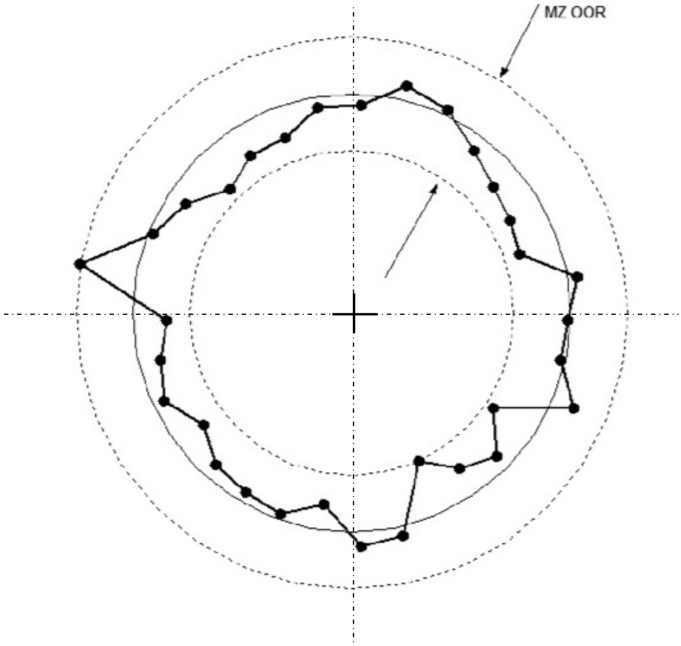


Figure 8.2 Minimum zone (*MZ*) out-of-roundness (*OOR*) for a sampled roundness profile (*bold line*). The *thin continuous line* represents the ideal geometry

MATLAB® environment (MathWorks 1991). The efficiency of these algorithms with respect to computation time is reported as a linear function of the number of data points. Gass *et al.* (1998) and Moroni and Pacella (2008) further improved the solution proposed by Carr and Ferreira (1995a, b) for MZ form error computation.

The LS algorithm, which represents a widely used procedure for form tolerance evaluation, minimizes the sum of squared deviations of measured points from the fitted feature. The LS method associates one substitute feature with measurement points (*e.g.*, one circle for roundness) and calculates the maximum peak-to-valley distance of the measurement points from the substitute feature. The LS method requires that the sum of the squared errors be minimized. Although the form error computed from the extreme points can be slightly higher than that obtained from the MZ method, the LS-fitted feature is very stable and much less sensitive to the effects of asperities, making it suitable for many practical applications. The implementation developed by NPL (the UK's national measurement laboratory) and based on the Least Squares Geometric Elements (LSGE) library for MATLAB® is used henceforth. The LSGE library consists of functions to find the LS fit of geometric shapes to data, implementing a number of geometric fitting routine key functions. It is based on a general-purpose, nonlinear LS solver that takes as the input function-and-gradient routines and these routines are implementations of the geometric evaluation key functions. This library was tried out on several benchmark data sets and was found to give correct results (Moroni and Pacella 2008).

Irrespective of the specific method used, a roundness profile is usually considered as conforming to the requirements when the OOR value is lower than the specified tolerance. It should be noted, however, that many different circular profiles, which can induce different functional properties of the machined items, can be characterized by the same OOR value, resulting in important differences in the characteristics of the profile, *e.g.*, in the assembly precision (Cho and Tu 2002).

8.4 Control Chart of Geometric Errors

When the quality of the manufactured product is related to surface texture (roughness and waviness) or to geometric form errors (*e.g.*, straightness, roundness, cylindricity, planarity), Shewhart's control chart can be used for process monitoring if the information related to the texture or to the form is summarized in just one (or few) synthetic variable. As an example, the roughness obtained on a machined surface is usually represented by the average roughness R_a parameter, or the roundness characterizing a given item can be summarized by the OOR.

This section reviews the approach based on Shewhart's control chart for monitoring a synthetic measure of the error between the actual profile and the ideal geometry, which can be considered representative of industrial practice.

8.4.1 Control Limits of the Individuals Control Chart

If the distribution function of the measured data can be assumed to be normal, then the individuals control chart can be used. Given a nominal false alarm probability α', the upper and lower control limits of the individuals control chart can be computed as follows.

$$\begin{aligned} \text{UCL} &= \mu + Z_{\alpha'/2}\sigma, \\ \text{CL} &= \mu, \\ \text{LCL} &= \mu - Z_{\alpha'/2}\sigma, \end{aligned} \tag{8.1}$$

where UCL is the upper control limit, μ is the mean of the individual measurements, σ represents the standard deviation, $Z_{\alpha'/2}$ is the $(1-\alpha'/2)$ percentile of the standard normal distribution, CL is the center line, and LCL is the lower control limit. Typically, μ and σ are unknown and hence they have to be estimated via a phase I sample of independently and identically distributed measurements.

Assume we collect a sample of n profiles observed from the in-control manufacturing process and let o_j denote the synthetic variable which summarizes the form error for the j th profile, where $j = 1, 2, \ldots n$. Classic estimators of μ and σ

are the sample mean and the sample standard deviation, *i.e.*, $\bar{o} = \frac{1}{n}\sum_{j=1}^{n} o_j$ and $s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left(o_j - \bar{o}\right)^2}$, respectively.

The sample standard deviation is asymptotically efficient for independently and identically distributed normal random variables. The disadvantage is that it may be sensitive to outliers. When outliers might occur, a different estimator of the standard deviation should be used which is less sensitive to these deviations. The average of the moving ranges, which are the absolute values of the difference of two successive observations, can be used to measure the process variability. The average moving range, say, $\overline{\text{MR}}$, scaled by $2/\sqrt{\pi}$ can be used to obtain a more robust estimator than the sample standard deviation. Let $\overline{\text{MR}} = \frac{1}{n-1}\sum_{j=2}^{n}\left|o_j - o_{j-1}\right|$, the individuals control chart is defined as follows. (Note that in a set of n observations sequentially sampled on a process, the number of moving ranges is $n-1$).

$$\begin{aligned} \text{UCL} &= \bar{o} + Z_{\alpha'/2}\frac{\sqrt{\pi}}{2}\overline{\text{MR}}, \\ \text{CL} &= \bar{o}, \\ \text{LCL} &= \bar{o} - Z_{\alpha'/2}\frac{\sqrt{\pi}}{2}\overline{\text{MR}}. \end{aligned} \tag{8.2}$$

When the underlying distribution function is not normal, the approach described can be still used after a suitable transformation of the OOR values has been implemented to achieve normality of the transformed data. The transformation is obtained by applying a single mathematical function to the raw data values. Depending on the distribution of sample data, there are many different functions, such as square root, logarithm, power, reciprocal, and arcsine, which one could apply to transform sample data. The Box–Cox power approach (Box and Cox 1964) finds an optimal power transformation that can be useful for correcting nonnormality in process data.

8.4.2 An Example of Application to the Reference Case Study

As previously observed, the information contained in a roundness profile can be summarized in the OOR value. With reference to the tolerance value *t* characterizing the circular profile, this OOR value is usually computed to decide whether the machined item has to be scrapped/reworked (if OOR ≥ *t*) or can be considered to conform to the requirements (if OOR < *t*).

The simplest approach for quality control of roundness profiles consists in monitoring the OOR values by an individuals control chart. In order to exemplify this approach, the case study described by Colosimo *et al.* (2008) is considered as a reference. It consists of 100 roundness profiles obtained from lathe turning, where each profile consists of a set of 748 points representing the deviations from the nominal radius (equal to 13 mm) at equally distributed angle locations. This case study is described in detail in Chapter 11. With reference to this case study of 100 roundness profiles, Table 8.1 reports the OOR values obtained for each item estimated by both the MZ algorithm and the LS algorithm.

Table 8.1 Out-of-roundness (*OOR*) values (mm) based on minimum-zone (*MZ*) and least-squares (*LS*) algorithms

	MZ	LS		MZ	LS		MZ	LS		MZ	LS
Sample	OOR	OOR	Sample	OOR	OOR	Sample	OOR	OOR	Sample	OOR	OOR
1	0.0217	0.0231	26	0.0181	0.0184	51	0.0241	0.0258	76	0.0168	0.0178
2	0.0182	0.0208	27	0.0166	0.0178	52	0.0097	0.0101	77	0.0175	0.0193
3	0.0165	0.0178	28	0.0100	0.0109	53	0.0154	0.0184	78	0.0156	0.0160
4	0.0125	0.0141	29	0.0215	0.0251	54	0.0121	0.0128	79	0.0133	0.0141
5	0.0168	0.0178	30	0.0131	0.0145	55	0.0179	0.0189	80	0.0215	0.0237
6	0.0133	0.0135	31	0.0149	0.0169	56	0.0168	0.0178	81	0.0192	0.0207
7	0.0251	0.0259	32	0.0115	0.0122	57	0.0207	0.0212	82	0.0133	0.0148
8	0.0192	0.0213	33	0.0120	0.0128	58	0.0191	0.0196	83	0.0222	0.0230
9	0.0108	0.0118	34	0.0145	0.0155	59	0.0146	0.0157	84	0.0217	0.0227
10	0.0188	0.0209	35	0.0172	0.0193	60	0.0182	0.0191	85	0.0170	0.0176
11	0.0110	0.0120	36	0.0206	0.0221	61	0.0185	0.0193	86	0.0181	0.0191
12	0.0193	0.0203	37	0.0155	0.0180	62	0.0100	0.0106	87	0.0163	0.0171
13	0.0177	0.0183	38	0.0188	0.0218	63	0.0133	0.0141	88	0.0188	0.0207
14	0.0204	0.0224	39	0.0105	0.0119	64	0.0131	0.0136	89	0.0117	0.0117
15	0.0188	0.0214	40	0.0134	0.0145	65	0.0132	0.0139	90	0.0130	0.0139
16	0.0179	0.0189	41	0.0171	0.0181	66	0.0138	0.0151	91	0.0166	0.0181
17	0.0164	0.0171	42	0.0136	0.0141	67	0.0156	0.0167	92	0.0197	0.0207
18	0.0117	0.0119	43	0.0176	0.0183	68	0.0120	0.0124	93	0.0192	0.0199
19	0.0235	0.0256	44	0.0206	0.0217	69	0.0165	0.0177	94	0.0178	0.0187
20	0.0124	0.0129	45	0.0255	0.0280	70	0.0143	0.0146	95	0.0186	0.0204
21	0.0148	0.0155	46	0.0141	0.0149	71	0.0176	0.0191	96	0.0134	0.0149
22	0.0131	0.0146	47	0.0174	0.0185	72	0.0103	0.0111	97	0.0209	0.0214
23	0.0188	0.0197	48	0.0130	0.0146	73	0.0197	0.0206	98	0.0167	0.0181
24	0.0140	0.0150	49	0.0167	0.0182	74	0.0128	0.0133	99	0.0143	0.0158
25	0.0132	0.0146	50	0.0150	0.0159	75	0.0123	0.0130	100	0.0146	0.0149

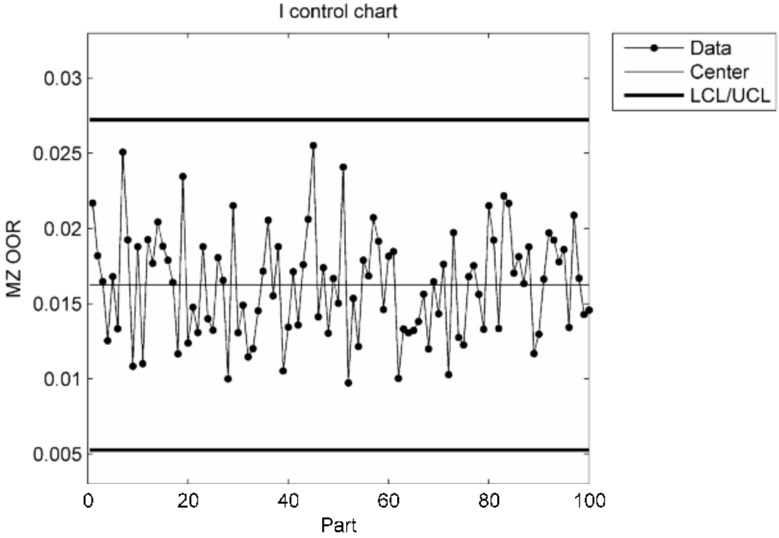


Figure 8.3 Individuals control chart of the transformed OOR values based on the MZ algorithm

From Table 8.1 it can be observed that for each profile the LS algorithm overestimates the roundness error calculated by the MZ algorithm. In fact, the sample mean of the OOR values based on the LS algorithm is 0.0174 mm, while the corresponding value based on the MZ algorithm is 0.0162 mm (a *t* test of the mean difference is significant with a *p* value less than 0.0005).

With reference to the 100 samples in phase I, summarized in Table 8.1, two individuals control charts of the OOR values are designed. The Anderson–Darling test (Anderson and Darling 1952) was implemented in order to detect any departures from normality for the distribution function of the OOR values in the case of both the MZ algorithm and the LS algorithm. From numerical computation on the data reported in Table 8.1, the set of 100 OOR values is normally distributed (with $p = 0.149$ in the case of the MZ algorithm and $p = 0.150$ in the case of the LS algorithm for the test). Therefore, an individuals control chart is designed for the sequence of sample OOR values, resulting in the control charts shown in Figure 8.3 and Figure 8.4. In particular, Figure 8.3 depicts the individuals control chart with reference to the OOR values based on the MZ algorithm, while a similar control chart with reference to the OOR values based on the LS algorithm is depicted in Figure 8.4. The two control charts depicted in Figures 8.3 and 8.4 are based on different center lines (*i.e.*, 0.0162 mm in the case of the MZ algorithm and 0.0174 mm in the case of the LS algorithm) as well as on a different estimated standard deviation for the data plotted (*i.e.*, 0.0037 mm in the case of the MZ algorithm and 0.0041 mm in the case of the LS algorithm, where the estimated value is based on the average moving range of data).

It can be observed that no out-of-control signals are detected in the two control charts in this design phase. In fact, there are no points exceeding the control limits

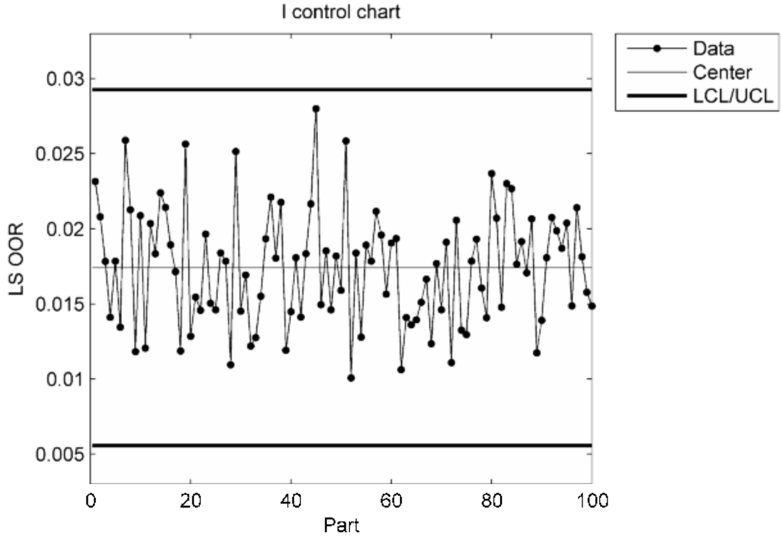


Figure 8.4 Individuals control charts of the transformed OOR values based on the least-squares (*LS*) algorithm

in the control charts of the OOR values obtained by the MZ algorithm and by the LS algorithm. As expected, irrespective of the specific method used, the patterns of OOR values in the two control charts depicted in Figures 8.3 and 8.4 are similar. Indeed, a computer simulation not reported here showed that the performance of an individuals control chart of the OOR values, measured in terms of a type I error rate for in-control profiles, is typically not affected by the specific algorithm used to estimate the geometric form error (*i.e.*, MZ algorithm or LS algorithm).

8.5 Monitoring the Shape of Profiles

Given the case study of 100 roundness profiles obtained from lathe turning (Colosimo *et al.* 2008), where each profile consists of a set of 748 points representing the deviations from the nominal radius at equally distributed angle locations (refer to Chapter 11 for details), it can be observed that the data can be stored as 748-length vectors.

Hence, each profile can be considered as a realization of a multivariate process. A possible approach for profile monitoring consists in studying all the points of the profile simultaneously by means of a specific multivariate technique, which reduces the information contained within the vector down to a single metric. An example of such multivariate metrics is the T^2 statistic (Montgomery 2004). However, the use of this technique is frequently ill-advised for profile monitoring. When the number of monitored points exceeds the number of collected samples (as in the reference case), the sample covariance matrix of data is not invertible and the usual statistical inference is not possible. This condition can often be encountered in actual applications, especially when machined profiles subject to geometric specification are considered, where in order to have an accurate estimate of the form error the number of observations sampled is usually on the order of hundreds.

A different approach, aimed at combining simplicity with the need of keeping all the information of the data observed at each location of the machined feature is the *location control chart* proposed by Boeing (1998) and described in the following subsections.

8.5.1 The Location Control Chart

The location control chart was presented in Boeing (1998, pp. 89–92) with reference to applications in which numerous measurements of the same variable (*e.g.*, a dimension such as thickness) are made at several locations on each manufactured part, *i.e.*, in the context of profile monitoring (Woodall *et al.* 2004). In practice, the location control chart consists in applying a traditional Shewhart control chart separately to each data point observed at a given location of the part, *i.e.*, it consists in designing a control interval at each different position of the point observed on the shape of interest. The rationale behind this approach is that, if the observed shape is in control, the data observed at that specific location should stay within that interval with a given probability. On the other hand, when the process goes out of control, it is likely that the control interval will be violated at one or more locations.

In order to design the location control chart, the first step consists in identifying the center of each interval, *i.e.*, the systematic pattern of the in-control shape. This reference for the in-control shape is usually estimated as the average of all the in-control data observed at each location.

Starting from the mean shape, generation of the location control chart consists in computing the upper and lower control limits at each location, using the standard approach that places the limits at $\pm K$ standard deviations from the sample mean. According to this method, an alarm is issued when at least one point, in the whole set of data observed, exceeds the control limits. The actual value of constant K depends on the required false-alarm rate for the monitoring approach. As in standard control charting, a greater value of K implies a larger control band and hence a lower false-alarm rate.

Owing to its inner simplicity, this chart can be easily applied in industrial practice (and in fact its origin is in Boeing 1998). However, since the control limits used at each location depend on the responses at that specific position only, the main disadvantage with this method is that the multivariate structure of data is ignored. The only form of the relationship between control intervals at each location is a constraint on the false alarm, as discussed in the next subsection.

8.5.2 Control Limits of the Location Control Chart

Assume we collect a group of n profiles, where each profile is a vector of p measurements observed at a fixed set of locations. The location control chart consists of limits computed separately for each location by means of Shewhart's approach, *i.e.*, by considering the mean and the standard deviation of the n data observed at that location and by computing the common $\pm K$ standard deviations from the sample mean. Given a profile, an alarm should be considered when at least one point, in the set of p observations, exceeds either the upper or the lower control limit.

Let y_j (k) denote the data measured at a specific location of index k on the jth profile, where $k = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, n$. The control limits for the location of index k are as follows:

$$\begin{aligned} &\mathrm{UCL}(k) = \bar{y}(k) + Z_{\alpha/2}s(k), \\ &\mathrm{CL}(k) = \bar{y}(k), \qquad\qquad\quad , \\ &\mathrm{LCL}(k) = \bar{y}(k) + Z_{\alpha/2}s(k), \end{aligned} \tag{8.3}$$

where $\bar{y}(k) = \frac{1}{n}\sum_{j=1}^{n} y_j(k)$ and $s(k) = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}\left[y_j(k) - \bar{y}(k)\right]^2}$ are, respectively, the sample mean and the sample standard deviation of the data observed at location k, while $Z_{\alpha/2}$ represents the $(1-\alpha/2)$ percentile of the standard normal distribution.

Given that p dependent control rules are simultaneously applied, Bonferroni's rule for dependent events should be used to attain an actual false-alarm rate not greater than a predefined value. Therefore, let α' denote the upper bound of the first type of probability error (false-alarm probability); the value $\alpha = \alpha'/p$ is used to design the p control limits of Equation 8.3.

In other words, the constant K of the location control chart is computed as a function (percentile of the standardized normal distribution) of the required false-alarm rate (type I error) corrected by Bonferroni's method. However, it is worth noting that different procedures can also be used, for instance the Simes modified Bonferroni procedure. Colosimo and Pacella (2010) showed that when compared with the standard Bonferroni method, the Simes procedure does not produce significant effects on the performance obtained by the location control chart. Furthermore, since the Simes procedure does not allow the graphical representation of the control region as the Bonferroni procedure does, the latter is usually considered for designing the location control chart in industrial practice.

8.5.3 An Example of Application of the Location Control Chart

The location control chart consists of a center line, an upper control limit, and a lower control limit. With reference to the roundness case study previously consid-

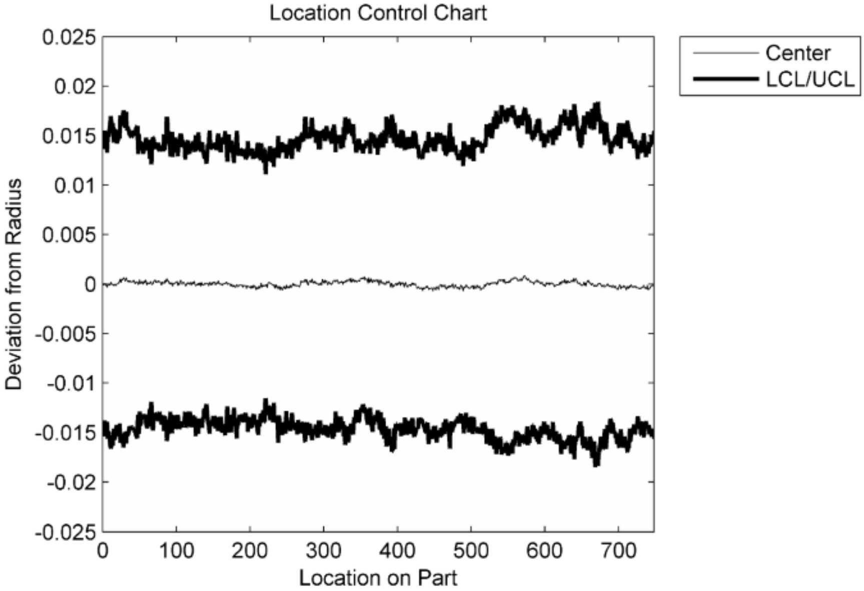


Figure 8.5 Control limits of the location control chart (748 locations) with reference to the 100 samples of the case study. Actual false-alarm rate not greater than 0.0027

ered (and detailed in Chapter 11), Figure 8.5 shows the average in-control shape together with the corresponding control limits. From a visual inspection of Figure 8.5 it seems that no systematic pattern characterizes the roundness profiles of the reference case study (*i.e.*, the Cartesian plot of the average profile is almost flat and equal to zero, *i.e.*, the mean profile is a perfect circle).

This appearance hides a common problem of shape analysis, which consists in feature registration or alignment. In fact, the average in-control shape and the corresponding control band are correctly computed by using the limits in Equation 8.3, but on profiles which are actually misaligned. After applying a registration procedure on the set of roundness data (refer to Chapter 11 for details), Figure 8.6 shows the average profile and the control band of the roundness profiles under study. From a visual inspection, it can be easily observed that the roundness profiles share a common shape (pattern), *i.e.*, the turning process leaves a specific signature on the machined components (Colosimo *et al.* 2008).

Given that 748 dependent control rules are simultaneously applied, Bonferroni's rule for dependent events is used to attain an actual false-alarm rate not greater than a predefined value. In particular, under the assumption of the standard value $\alpha' = 0.0027$ for the upper bound of the type I probability error (false-alarm probability), the value $\alpha = 3.61 \cdot 10^{-6}$ is used to design the 748 control limits in Equation 8.3. The corresponding percentile of the standard normal distribution is about equal to $Z_{\alpha/2} = 4.63$.

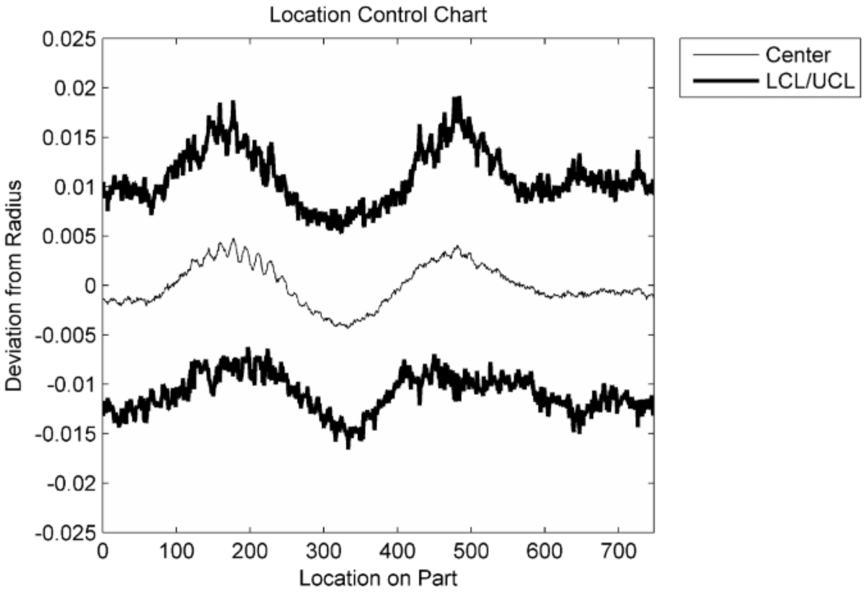


Figure 8.6 Location control chart (748 locations) with reference to the 100 samples of the case study (after the registration step). Actual false-alarm rate not greater than 0.0027

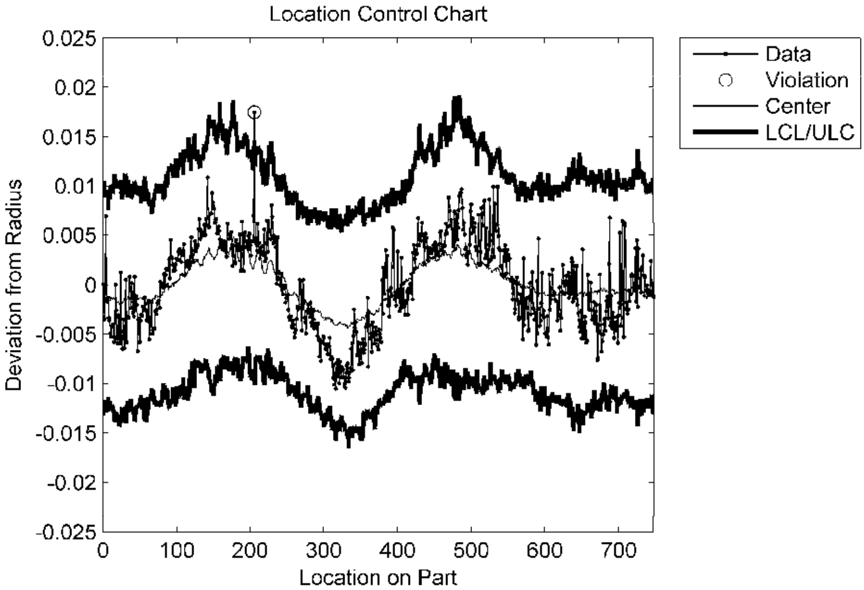


Figure 8.7 One of the 100 profiles of the reference case study depicted against the control limits of the location control chart (after the registration step)

Figure 8.7 shows one out of the 100 profiles of the reference case study depicted against the control limits of the location control chart. The profile is plotted against this control region, with the advantage of allowing a simple identification of the locations where problems arise. In this specific case, even if an alarm is issued at a specific location (location no. 206), the profile is considered to be in control. Indeed, from the visual inspection of the location control chart, it appears that the behavior of this profile is close to the average common profile and there is no apparent discrepancy in the shape of the profile when compared with the center line of the location control chart.

8.6 Conclusions

This chapter reviewed industrial practice and the contributions of statistical methods to quality monitoring when a geometric tolerance is the quality characteristic of interest.

In particular, two methods were described. First, a control chart for monitoring a synthetic measure of the geometric error, (*i.e.*, a scalar that measures the error between the actual profile and the ideal geometry). Second, a control region for monitoring the whole profile observed (where bound limits of this region are computed by applying a control chart separately to each set of data points observed at a given location).

Given the ease of the approaches presented in this chapter, they have been considered as industrial benchmarks (Colosimo and Pacella 2007, 2010; Colosimo *et al.* 2008, 2010) when different and more complex methods, such as the ones presented in the subsequent chapters, are considered for geometric tolerance monitoring.

A comparison of the performance presented by the two approaches detailed in this chapter, along with those of the methods presented in the subsequent two chapters for profile monitoring, is left to Chapter 11. The aim is to allow practitioners to select a specific method in a given production scenario that mimics the actual case study of roundness profiles.

Acknowledgements This research was partially funded by the Italian Ministry of University and Research grant PRIN 2005 – "Characterization of the multiscale geometry of manufactured surfaces for process control and tolerance analysis".

References

Alwan LC (2000) Statistical process analysis. Irwin/McGraw-Hill, New York

Anderson TW, Darling DA (1952) Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Ann Math Stat 23:193–212

Boeing (1998) Advanced quality system tools, AQS D1-9000-1. Boeing Commercial Airplane Group, Quality Assurance Department. http://www.boeing.com/companyoffices/doingbiz/supplier/d1-9000-1.pdf

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc Ser B 26:211–252

Carr K, Ferreira P (1995a) Verification of form tolerances. Part I: basic issues, flatness, and straightness. Precis Eng 17:131–143

Carr K, Ferreira P (1995b) Verification of form tolerances. Part II: cylindricity and straightness of a median line. Precis Eng 17:144–156

Cho NW, Tu JF (2001) Roundness modeling of machined parts for tolerance analysis. Precis Eng 25:35–47

Cho NW, Tu JF (2002) Quantitative circularity tolerance analysis and design for 2D precision assemblies. Int J Mach Tools Manuf 42:1391–1401

Colosimo BM, Pacella M (2007) On the use of principal component analysis to identify systematic patterns in roundness profiles. Qual Reliab Eng Int 23:707–725

Colosimo BM, Pacella M (2010) Control Charts for Statistical Monitoring of Functional Data, Int J Prod Res 48(6):1575–1601

Colosimo BM, Pacella M, Semeraro Q (2008) Statistical process control for geometric specifications: on the monitoring of roundness profiles. J Qual Technol 40:1–18

Colosimo BM, Mammarella F, Petrò S (2010) Quality control of manufactured surfaces. In: Lenz HJ, Wilrich PT (eds) Frontiers of statistical quality control, vol 9. Springer, Vienna

Gass SI, Witzgall C, Harary HH (1998) Fitting circles and spheres to coordinate measuring machine data. Int J Flex Manuf Syst 10:5–25

ISO/TS 12181 (2003) Geometrical product specification (GPS) roundness. http://isotc213.ds.dk/

MathWorks (1991) MATLAB user's guide. The MathWorks, Natick

Montgomery DC (2004) Introduction to statistical quality control, 5th edn. Wiley, New York

Moroni G, Pacella M (2008) An approach based on process signature modeling for roundness evaluation of manufactured items. J Comput Inf Sci Eng 8:021003

Ryan TP (2000) Statistical methods for quality improvement. Wiley, New York

Shewhart WA (1931) Statistical method from an engineering viewpoint. J Am Stat Assoc 26:262–269

Woodall WH, Spitzner DJ, Montgomery DC, Gupta S (2004) Using control charts to monitor process and product quality profiles. J Qual Technol 36:309–320

Chapter 9
Model-based Approaches for Quality Monitoring of Geometric Tolerances

Bianca Maria Colosimo and Massimo Pacella

Abstract A relatively new area of research in the field of statistical process control has been named *profile monitoring*. It includes a collection of methods and techniques used to check the stability of a functional relationship over time. This chapter shows how this approach can be usefully considered as a viable solution to form error monitoring when geometric tolerances are of interest. In this case, quality monitoring consists in detecting deviations of the shape from its nominal or in-control state. This task is accomplished by firstly modeling the functional relationship representing the manufactured shape and then checking whether or not the estimated model is stable over time. The goal of this chapter is to introduce profile monitoring, show how it works, and then illustrate how this approach can be effectively used for quality control of geometric form errors.

9.1 Introduction

With the development of computerized data-acquisition systems and of modern measuring equipment, the quality of products or processes is more and more often related to *functional data* (Ramsay and Silverman 2005). In the simplest form, func-

B.M. Colosimo
Dipartimento di Meccanica, Politecnico di Milano,
Via la Masa 1, 20156 Milan, Italy,
e-mail: biancamaria.colosimo@polimi.it

M. Pacella
Dipartimento di Ingegneria dell'Innovazione, Università del Salento,
Via per Monteroni, 73100 Lecce, Italy,
e-mail: massimo.pacella@unisalento.it

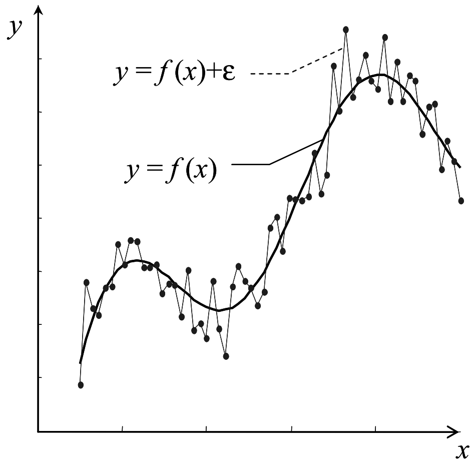


Figure 9.1 A simple example of functional data

tional data refer to information summarized in the form of profiles where the response variable y can be modeled as a function of one (or more) independent variable x plus random noise ε, *i.e.*, $y = f(x) + \varepsilon$ (Figure 9.1). Usually, the independent variable is used to define a spatial or temporal location. Very often, functional data are related to quality characteristics of the product or process and hence monitoring the stability of the functional data allows one to detect out-of-control states that are usually associated with deteriorated process performances (quality loss).

Since functional data are usually observed only at a finite set of locations, a vector can be used to store the observed set of responses. Therefore, the most direct way to deal with these data vectors consists in using multivariate control charting (Montgomery 2000), thus treating each profile as a realization of a multivariate process. The use of standard multivariate charts for monitoring profile data is not recommended. When the number of points of the profile exceeds the number of profiles used during the design stage, the common multivariate statistics cannot be estimated. In actual applications, the number of monitored points on each profile is usually greater than the number of samples collected during the design stage. This is especially true in the case of machined profiles subject to geometric specification (*e.g.*, roundness, straightness, free-form tolerance) where, in order to have an accurate estimate of the form error, the number of observations sampled can be on the order of hundreds. Hence, when the quality of a process or product is characterized by functional data, new approaches are required.

Recently, there has been much research activity in a new area of statistical process control, named *profile monitoring*. Woodall *et al.* (2004) and Woodall (2007) discussed general issues to be addressed when monitoring quality profiles, and presented a complete review of the literature on the topic of profile monitoring. The approaches for profile monitoring proposed in the literature share a common structure which consists of:

1. identifying a parametric model of the functional data;
2. estimating the model parameters; and
3. designing a multivariate control chart of the estimated parameters and a univariate control chart of the residual variance.

The proposed approaches can then be classified with reference to the type of application faced (*i.e.*, calibration study, process signal, or geometric specification monitoring) or to the modeling approach considered [linear or nonlinear regression, nonparametric regression, or approaches for multivariate data reduction such as principal component analysis (PCA)/independent component analysis].

With reference to the type of application faced, most of the studies on profile monitoring have dealt with calibration studies (Stover and Brill 1998; Kang and Albin 2000; Kim *et al.* 2003; Mahmoud and Woodall 2004; Chang and Gan 2006; Gupta *et al.* 2006; Zou *et al.* 2006; Mahmoud *et al.* 2007) where the profiles that have to be monitored are straight lines. A second stream of applications concerns monitoring of signals coming from machines with sensors. For instance, Jin and Shi (1999, 2001) and Ding *et al.* (2006) referred to profiles representing force and torque signals collected from online sensors on a press in a stamping process. The third stream of applications which is of interest in this book concerns the use of profile monitoring for quality control of geometric specifications (Colosimo and Pacella 2007, 2010; Colosimo *et al.* 2008, 2010). In fact, machined profiles and surfaces can be thought of as functional data if one of the spatial coordinates describing the machined surface can be represented as a function of the other two spatial coordinates. For instance, in roundness profiles where the radius is modeled as a function of the angle, the observed data can be represented as functional data (Figure 9.2). Similarly, if the cylindricity tolerance is of interest, the cylindrical surface can be modeled by representing the radius as a function of the angle and vertical position of each measured point, as shown in Figure 9.3. Therefore, a three-dimensional surface can be modeled as functional data as well and surface monitoring can be shown to be a generalization of profile monitoring (Colosimo *et al.* 2010).

Colosimo and Pacella (2007) and Colosimo *et al.* (2008) dealt with the roundness profile obtained by lathe-turning, showing that both PCA and spatial autoregressive regression (SARX) models can be used for modeling and then monitoring the geometric profile. By combining these models with control charting, the papers show how out-of-control states of the manufactured profile can be easily and quickly detected. Colosimo and Pacella (2010) compared the performances of different approaches (such as the simplest methods described in the previous chapter and the model-based approaches described in this chapter) to outline scenarios in which a specific approach outperforms the others for geometric error monitoring. More recently, Colosimo *et al.* (2010) extended the proposed method to three-dimensional surface monitoring, using as a case study the cylindricity of lathe-turned items.

This chapter shows how profile monitoring can be used effectively for geometric form error monitoring. In particular, the simplest case of straight profiles is used in Section 9.2 to show how profile monitoring works. The following sections describe approaches aimed at monitoring geometric tolerances by using spatial regression models and PCA-based ones. Numerical examples are used to illustrate all the approaches presented.

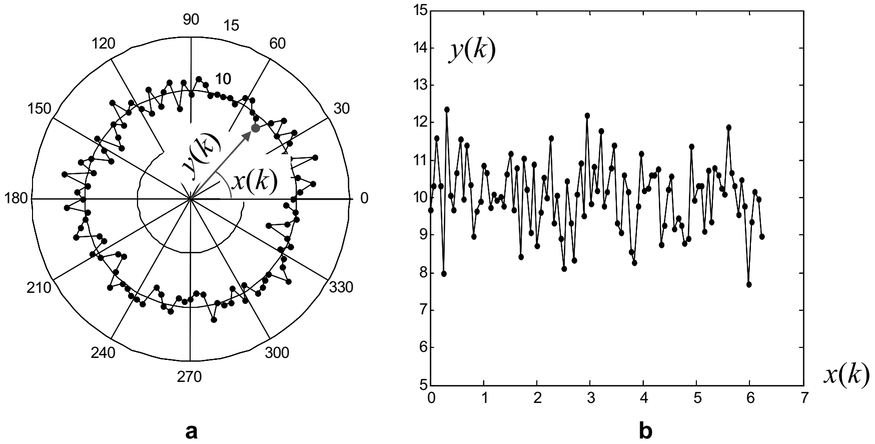


Figure 9.2 An example of a roundness profile: **a** the polar graph shows the radius $y(k)$ modeled as a function of the angle $x(k)$, and **b** the Cartesian diagram describes the roundness profile as functional data

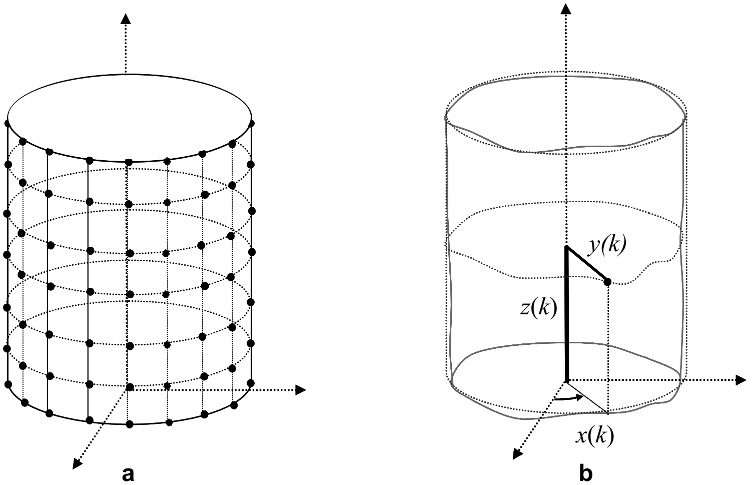


Figure 9.3 The equally spaced grid of locations where the radius has to be measured to model **a** a cylindrical surface, and **b** the coordinates of the *k*th point [radius $y(k)$, angle $x(k)$, and vertical position $z(k)$]

9.2 Linear Profile Monitoring

Linear profile monitoring is a broad topic applied for a wide variety of applications. Most of the work on profile monitoring proposed in the literature is related to the case in which the profile can be adequately represented by a simple straight line (Woodall *et al.* 2004; Woodall 2007).

Calibration processes are often characterized by such linear functions. For instance, Stover and Brill (1998) studied multilevel ion chromatography linear calibrations to determine the instrument response stability and the proper calibration frequency. Furthermore, Kang and Albin (2000) presented two examples of process profiles. One of them was a semiconductor manufacturing application in which the process is represented by a linear calibration function. Several other case studies and approaches for linear profile monitoring can be found in the literature (Kim *et al.* 2003; Mahmoud and Woodall 2004; Chang and Gan 2006; Gupta *et al.* 2006; Zou *et al.* 2006; Mahmoud *et al.* 2007). These approaches share a common basic idea for implementing linear profile monitoring, *i.e.*, to use control charts based on the estimated regression parameters.

Assume we collect a group of n profiles, where each profile consists of p measurements observed at a fixed set of locations. Let $y_j(k)$ denote the dependent variable measured at a specific location of index k on the j th profile, and let $x(k)$ represent the value of the independent variable at the same location ($k = 1, 2, \ldots p$ and $j = 1, 2, \ldots n$). We also assume that, as in many profile monitoring applications, the x-values are known constants and have the same values in all samples.

The observed data collected over time are n random samples, with each sample consisting of a sequence of p pairs of observations $\left[x(k), y_j(k)\right]$. For each sample of index $j = 1, 2, \ldots n$, it is assumed that the model which relates the independent variable x to the response y is the following:

$$y_j(k) = b_{0j} + b_{1j}x(k) + \varepsilon_j(k), \tag{9.1}$$

where the $\varepsilon_j(k)$'s are assumed to be independent, identically distributed $N\left(0, \sigma_j^2\right)$ random variables.

The in-control values of the parameters b_{0j}, b_{1j}, and σ_j^2 in Equation 9.1 are unknown. If $b_{0j} = \beta_0$, $b_{1j} = \beta_1$, and $\sigma_j^2 = \sigma^2$, $j = 1, 2, \ldots n$, then the model in Equation 9.1 is called a "fixed-effects" model. Furthermore, the process monitored is considered statistically in control if the function used to represent the observed profiles is stable over time and the profile-to-profile variability is also stable over time.

In ordinary linear regression, it is well known that the least-squares estimates of b_0 and b_1 for a sample of index j are the following:

$$\hat{b}_{0j} = \bar{y}_j - \hat{b}_{1j}\bar{x} \qquad \text{and} \qquad \hat{b}_{1j} = S_{xy(j)} \big/ S_{xx}\,, \tag{9.2}$$

where $\bar{y}_j = \sum_{k=1}^{p} y_j(k) \big/ p$, $\bar{x} = \sum_{k=1}^{p} x(k) \big/ p$, $S_{xx} = \sum_{k=1}^{p} \left[x(k) - \bar{x}\right]^2$, and $S_{xy(j)} = \sum_{k=1}^{p} \left[x(k) - \bar{x}\right]\left[y_j(k) - \bar{y}_j\right]$. Furthermore, σ_j^2 can be estimated by the jth mean square error MSE_j, where $\mathrm{MSE}_j = \sum_{k=1}^{p} e_j^2(k) \big/ (p-2)$. Here $e_j(k)$ is the residual error at the location of index $k = 1,2,\ldots p$ on the profile of index $j = 1,2,\ldots n$, in the formula $e_j(k) = y_j(k) - \hat{b}_{0j} - \hat{b}_{1j}x(k)$.

Assume the linear profile process is in control, with fixed effects and independent, identically distributed residual errors $N\left(0,\sigma^2\right)$, then the least-squares estimators of the intercept $\hat{b}_{0j}$ and slope $\hat{b}_{1j}$ are distributed as a bivariate normal distribution with the mean vector and the variance–covariance matrix, respectively:

$$\boldsymbol{\beta}' = [\beta_0 \ \beta_1], \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}, \tag{9.3}$$

where $\sigma_0^2 = \sigma^2\left(1/p + \bar{x}^2/S_{xx}\right)$, $\sigma_1^2 = \sigma^2/S_{xx}$, and $\sigma_{01} = -\sigma^2\bar{x}^2/S_{xx}$ are the variance of $\hat{b}_{0j}$, the variance of $\hat{b}_{1j}$, and the covariance between $\hat{b}_{0j}$ and $\hat{b}_{1j}$, respectively, $j = 1,2,\ldots n$.

Kim *et al.* (2003) proposed a method for linear profile monitoring which is based on the fact that the estimators of the intercept and the slope are statistically uncorrelated ($\sigma_{01} = 0$) when the independent variable $x(k)$ is coded so that the average coded value is zero ($\bar{x} = 0$). In this case, Kim *et al.* (2003) recommended monitoring the two regression coefficients (intercept and slope) using separate control charts. They also recommended using an additional univariate control chart to monitor the residual variation about the regression line (*i.e.*, a statistic related to the residual variance). According to the approach proposed by Kim *et al.* (2003), a signal is produced as soon as any of the three control charts for the intercept, the slope, and the variation about the regression line produce an out-of-control signal. This method provides easier interpretation of an out-of-control signal than other methods since each parameter in the model is monitored using a separate control chart.

The following subsection details the approach proposed by Kim *et al.* (2003) for linear profile monitoring and shows how it can be effectively used. A numerical example is also given.

9.2.1 A Control Chart Approach to Linear Profile Monitoring

As in any statistical process control method, profile monitoring approaches can be referred to two different phases, namely, phase I and phase II of control charting. The purpose of the analysis in phase I is to analyze a historical set of a fixed number of process samples collected over time to understand the process variation, determine the stability of the process, and remove samples associated with any assignable causes. Having removed those samples, one estimates the in-control values of the process parameters to be used in designing control charts for the phase II analysis. The main interest in phase II monitoring of profile data is to quickly detect parameter changes from the in-control parameter values established in phase I.

Kim *et al.* (2003) proposed a method for monitoring a linear profile process in phase II and recommended applying this method also in phase I. This method consists in using three separate Shewhart-type control charts for monitoring the intercept, slope, and residual variance. The performance of the method originally proposed by Kim *et al.* (2003) was thoroughly analyzed by Mahmoud and Woodall (2004), who demonstrated how it can be much more effective than several others in signaling shifts affecting data. Mahmoud and Woodall (2004) recommended using the Kim *et al.* (2003) method, not only as it outperforms competing approaches, but also because it is simple and much more interpretable than other methods.

The first step in the Kim *et al.* (2003) method is to code the independent variable $x(k)$ so that the average coded value is zero. Coding the independent variable in this way yields another form of the model in Equation 9.1, *i.e.*,

$$y_j(k) = a_{0j} + a_{1j}x'(k) + \varepsilon_j(k), \tag{9.4}$$

where $a_{0j} = b_{0j} + b_{1j}\bar{x}$, $a_{1j} = b_{1j}$, and $x'(k) = x(k) - \bar{x}$. The least-squares estimators for the regression parameters for sample j are calculated as $\hat{a}_{0j} = \bar{y}_j$ and $\hat{a}_{1j} = \hat{b}_{1j} = S_{xy(j)} / S_{xx}$. For an in-control process, the slope $\hat{a}_{0j}$ and the intercept $\hat{a}_{1j}$ are statistically uncorrelated random variables. The means are a_{0j} and a_{1j} and the variances are σ^2/p and σ^2/S_{xx}, respectively.

As a second step, Kim *et al.* (2003) proposed applying a separate Shewhart-type control chart for each of the three parameters in the model (the slope and the intercept of the regression line and the residual variation about the regression line). In particular, with reference to the intercept a_{0j}, as it can be demonstrated that the quantity $\left(a_{0j} - \bar{a}_0\right) / \sqrt{(n-1)\mathrm{MSE}/pn}$ follows a *t* distribution with $n(p-2)$ degrees of freedom, a Shewhart-type control chart for monitoring this parameter during phase I has the following control limits:

$$\begin{aligned} \mathrm{UCL} &= \bar{a}_0 + t_{n(p-2),\alpha/2}\sqrt{(n-1)\mathrm{MSE}/pn}, \\ \mathrm{CL} &= \bar{a}_0, \\ \mathrm{LCL} &= \bar{a}_0 - t_{n(p-2),\alpha/2}\sqrt{(n-1)\mathrm{MSE}/pn}, \end{aligned} \tag{9.5}$$

where UCL is the upper control limit $\bar{a}_0 = \sum_{j=1}^{n} a_{0j}/n$, $t_{n(p-2),\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the *t* distribution with $n(p-2)$ degrees of freedom, CL is the center line, and LCL is the lower control limit. Similarly, it can be shown that the quantity $(a_{1j} - \bar{a}_1)/\sqrt{(n-1)\mathrm{MSE}/nS_{xx}}$ follows a *t* distribution with $n(p-2)$ degrees of freedom. Therefore, a Shewhart-type control chart with the following control limits can be used for monitoring the slope a_{1j}:

$$\begin{aligned} \mathrm{UCL} &= \bar{a}_1 + t_{n(p-2),\alpha/2}\sqrt{(n-1)\mathrm{MSE}/nS_{xx}}, \\ \mathrm{CL} &= \bar{a}_1, \\ \mathrm{LCL} &= \bar{a}_1 - t_{n(p-2),\alpha/2}\sqrt{(n-1)\mathrm{MSE}/nS_{xx}}, \end{aligned} \tag{9.6}$$

where $\bar{a}_1 = \sum_{j=1}^{n} a_{1j}/n$ and $t_{n(p-2),\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the *t* distribution with $n(p-2)$ degrees of freedom.

Finally, it can also be shown that $F_j = \mathrm{MSE}_j/\mathrm{MSE}_{-j}$ has an *F* distribution with $(p-2)$ and $(n-1)(p-2)$ degrees of freedom, where $\mathrm{MSE}_{-j} = \sum_{k\neq j}^{p} \mathrm{MSE}_k/(p-1)$. Therefore, a Shewhart control chart for monitoring the residual variation about the regression line involves plotting the quantities F_j, $j = 1,2,\ldots,n$, on a chart with the following upper and lower limits:

$$\begin{aligned} \mathrm{UCL} &= F_{(p-2),(n-1)(p-2),(1-\alpha/2)}, \\ \mathrm{LCL} &= F_{(p-2),(n-1)(p-2),\alpha/2}. \end{aligned} \tag{9.7}$$

9.2.2 A Numerical Example for Linear Profile Monitoring

Consider a set of $n = 12$ profiles, where each profile consists of $p = 10$ measurements observed at a fixed set of locations. Without loss of generality, assume we have the independent variable x of $p = 10$ discrete values equally spaced in the interval [–0.5, 0.5]. The x-values take the same values in all samples.

Figure 9.4 graphically depicts the data set. In particular, each graph represents the points on each profile. Each point represents a functional relationship between an independent variable $x(k)$ and a dependent variable $y(k)$. Also in the same figure, the fitted linear profile, which links the independent variable $x(k)$ to the dependent one $y(k)$ according to Equations 9.1 and 9.2, is depicted. The least-squares estimates of the intercept ($\hat{b}_{0j}$) and the slope ($\hat{b}_{1j}$) for each sample j are obtained by Equation 9.2 and are summarized in Table 9.1. The residuals $e_j(k)$ for each sample j at the location of index k are depicted in Figure 9.5.

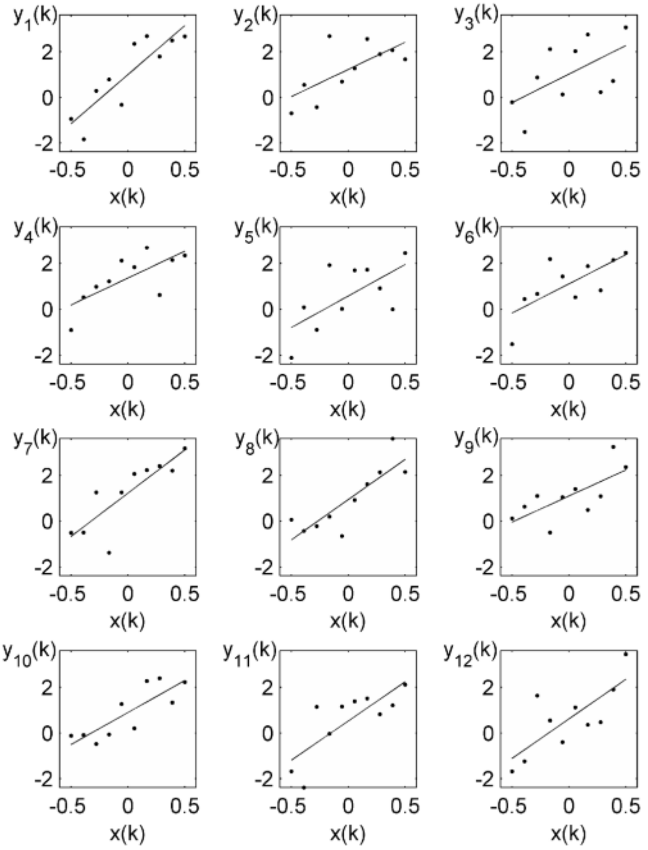


Figure 9.4 A linear profile data set consisting of 12 samples each with ten points. Each point represents a functional relationship between an independent variable (*abscissa*) and a dependent one (*ordinate*). The independent variable assumes the values are equally spaced on the abscissa. The *continuous line* in each graph represents the least-squares fitted line where the intercept and slope are based on Equation 9.2

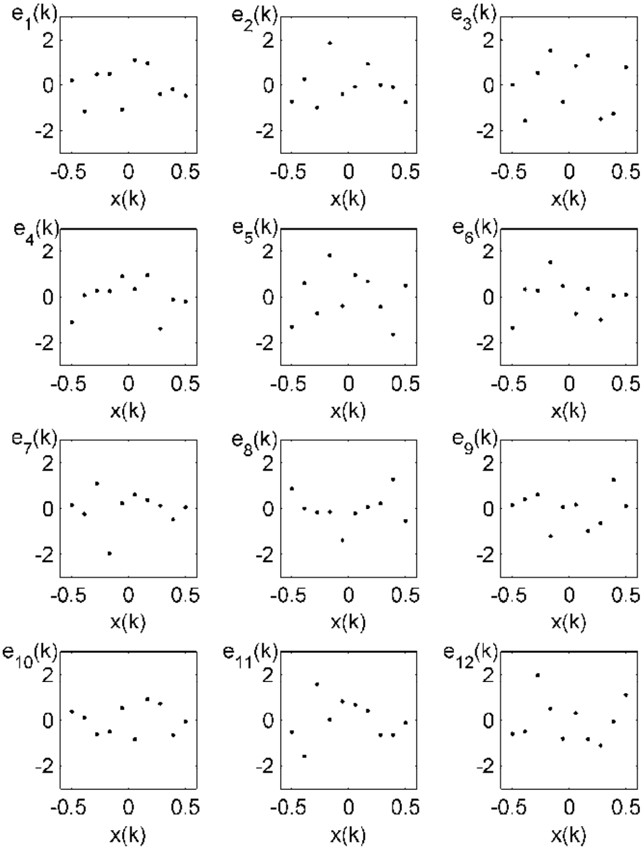


Figure 9.5 Residual errors between observed points depicted in Figure 9.4 and the least-squares fitted line. A standard assumption in the monitoring of simple linear regression profiles is that the errors are independent and identically distributed, usually with an assumed normal distribution

Table 9.1 Least-squares estimation of the intercept and slope for 12 samples of ten points each

j	$\hat{b}_{0j}$	$\hat{b}_{1j}$
1	1.0013	4.2896
2	1.2310	2.3770
3	1.0253	2.4904
4	1.3588	2.3424
5	0.5803	2.7451
6	1.1055	2.5346
7	1.2253	3.7798
8	0.9457	3.5215
9	1.1022	2.2775
10	0.9039	2.8027
11	0.5335	3.4190
12	0.6243	3.4545

The first step in the Kim *et al.* (2003) approach is to code the independent variable $x(k)$ so that the average coded value is zero. In the numerical example at hand, as the independent variable x assumes discrete values equally spaced in the interval $[-0.5, 0.5]$, we have $\bar{x} = 0$. Therefore, $a_{0j} = b_{0j}$, $a_{1j} = b_{1j}$, and $x'(k) = x(k)$, *i.e.*, Equation 9.4 is equal to Equation 9.1. For an in-control process with fixed effects (*i.e.*, $b_{0j} = \beta_0$, $b_{1j} = \beta_1$), the slope $\hat{b}_{0j}$ and the intercept $\hat{b}_{1j}$ are statistically uncorrelated, normally distributed random variables [as $\varepsilon_j(k) \sim N(0, \sigma^2)$], with means β_0 and β_1 and variances σ^2/p and σ^2/S_{xx}, respectively.

In this numerical example, the in-control parameters of the simulated linear profile process are $\beta_0 = 1$, $\beta_1 = 3$, and $\sigma^2 = 1$. Therefore, the estimated least-square estimate slopes $\hat{b}_{0j}$ have mean equal to 1 and variance equal to 0.1. On the other hand, the estimated least-squares intercepts $\hat{b}_{1j}$ have mean equal to 3 and variance approximately equal to 0.9818. These true values of the process monitored can be used to implement two control charts in order to monitor the intercept and the slope separately (Kim *et al.* 2003).

When the in-control process parameters β_0, β_1, and σ^2 are not known, the control limits of Equations 9.5 and 9.6 can be used for phase I control charting. In particular, the limits computed for the numerical example at hand are reported in Table 9.2. The control chart for the intercept is depicted in Figure 9.6a, while the chart of the slopes is shown in Figure 9.6b. Similarly, the control limits in Equation 9.7 are used to monitor the statistics F_j. The numerical values of such limits are also reported in Table 9.2, while the corresponding control chart is depicted in Figure 9.6c.

A nominal false-alarm probability (type I error rate) equal to $\alpha' = 5\%$ was used to compute the control limits of the three charts. Given three control charts are contemporarily used, Bonferroni's rule for independent events was used to design the control limits of the three charts. Therefore, given a nominal false-alarm probability $\alpha' = 5\%$, the control limits of each of the three control charts in Equations 9.5–9.7 were set by assuming a false-alarm probability $\alpha = 1 - \sqrt[3]{1 - \alpha'}$. From Figure 9.6, it can be noted that there are no profiles, in the set of 12 samples, which produce an alarm in any of the three control charts.

Table 9.2 Phase I control limits for each parameter of the linear profile model for the numerical example. Type I error rate 5%

Parameter	Upper control limit	Central line	Lover control limit
Intercept	0.2902	0.9698	1.6493
Slope	0.8734	3.0029	5.1323
Variance	0.1912		2.7882

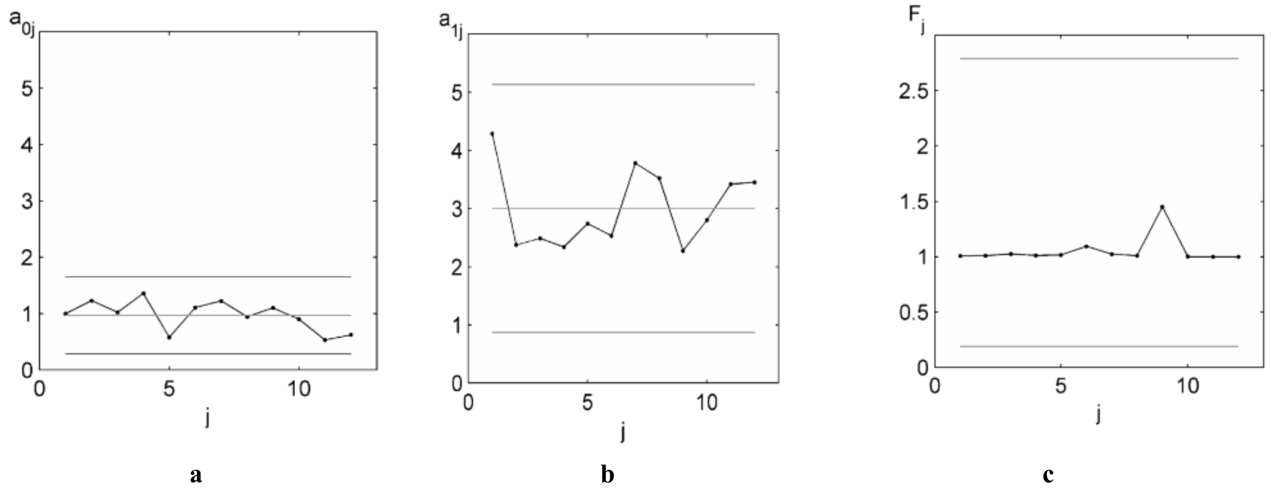


Figure 9.6 **a** Control chart of the least squares estimates of the intercept (phase I), **b** control chart of the least-squares estimates of the slope (phase I), and **c** control chart for monitoring the residual variation about the regression line (phase I)

9.3 Profile Monitoring for Geometric Tolerances

Although most approaches for profile monitoring focus on linear profiles, when geometric tolerances are of interest, *nonlinear* profiles are usually necessary. Different authors have focused on nonlinear profile monitoring (Williams *et al.* 2007; Walker and Wright 2002; Jin and Shi 1999, 2001; Young *et al.* 1999; Ding *et al.* 2006; Zou *et al.* 2008; Zhang and Albin 2009; Jensen and Birch 2009).

Walker and Wright (2002, p. 124) mentioned that "autocorrelation is frequently present in data that are observed within small intervals of time or space". Despite the specific mention of the autocorrelation problem, the approach proposed by Walker and Wright (2002) was based on independent data, whereas most of the other approaches focused on profile monitoring. The only exceptions to this general rule are the papers by Jensen *et al.* (2008) and Colosimo *et al.* (2008, 2010). In particular, while the first paper dealt with autocorrelated profile data, the second ones focused on spatially autocorrelated data and are more suitable for modeling geometric form features (profiles and surfaces). When data refer to surface texture or profile tolerances, measurements are often spatially correlated because they are obtained in similar conditions of the machining process and are related to similar (local) properties of the machined material. Spatial autocorrelation is different from temporal autocorrelation, which is usually represented through time-series models. Spatial autocorrelation models allow one to represent contiguity in space rather than in time. With reference to profiles, contiguity in space implies that the dependency among data on a profile is bidirectional (*i.e.*, a given point is, in principle, correlated to points located on its left and on its right, regardless of the specific direction), while time-series models are suitable for representing just a one-direction dependency (*i.e.*, past data influence future ones) (Whittle 1954). Second, spatial models allow one to represent a specific type of relationship among points observed in closed profiles (*e.g.*, roundness profile). In fact, when data on a closed or circuit profile are sequentially numbered (by defining an arbitrary starting point), observations at the beginning and at the end of the profile are spatially correlated (Colosimo *et al.* 2008).

With reference to the vector of points measured on a machined profile which is subject to geometric specification, the predictable behavior can be referred to as the process *signature* and is defined as the systematic pattern that characterizes all the features machined with a given process. Knowledge of this signature can be used to design proper tools for profile monitoring. Colosimo *et al.* (2008, 2010) presented an approach for modeling the manufacturing signature based on fitting a SARX model (Cressie 1993). The regression model presented in these papers is summarized below.

9.3.1 Regression Model with Spatially Correlated Errors

Assume we collect a group of n profiles, where each profile consists of p measurements observed at a fixed set of locations. Let $y_j(k)$ denote the dependent variable measured at a specific location of index k on the jth profile, and let $x_l(k)$ represent the values of r independent variables at the same location ($k = 1,2,\dots p$, $j = 1,2,\dots n$, and $l = 1,2,\dots r$). As in many profile monitoring applications, we assume that the x-values are known constants and have the same values in all samples. If we assume that we organize the p data observed on the jth profile into a column vector $\mathbf{y}'_j = \left[y_j(1)\dots y_j(k)\dots y_j(p)\right]$, the general SARX model can be written in matrix notation as follows:

$$\begin{aligned} \mathbf{y}_j &= \mathbf{X}\mathbf{b}_j + \boldsymbol{\upsilon}_j, \\ \left(\mathbf{I} - \mathbf{R}_j\right)\boldsymbol{\upsilon}_j &= \boldsymbol{\varepsilon}_j, \\ \mathbf{R}_j &= \sum_{s=1}^{q} a_{sj}\mathbf{W}^{(s)}. \end{aligned} \tag{9.8}$$

The first expression in Equation 9.8 describes the $p \times 1$ vector of the response for the jth profile $\mathbf{y}_j$ as formed by a large-scale and a small-scale component (Cressie 1993). The large-scale component is given by $\mathbf{X}\mathbf{b}_j$ where $\mathbf{X}$ is a $p \times r$ matrix of r regressor variables that are assumed to be known and constant and $\mathbf{b}'_j = \left[b_{1j}\dots b_{lj}\dots b_{rj}\right]$ is the $r \times 1$ vector of regression parameters which are normally distributed with mean $\boldsymbol{\beta}' = [\beta_1\dots\beta_l\dots\beta_r]$ and covariance matrix $\mathbf{B}$ $\left[\mathbf{b}_j \sim N(\boldsymbol{\beta},\mathbf{B})\right]$. With reference to the large-scale component, fixed-effect models are usually assumed in traditional approaches for profile monitoring. This situation can be seen as a special case of the general model in Equation 9.8, where $\mathbf{B} = \mathbf{0}$ is the variance matrix of the large-scale model coefficients $\mathbf{b}_j$. Alternatively, when $\mathbf{B} \neq \mathbf{0}$ is considered, random effects are included in the large-scale component of the model.

The small-scale component is the $p \times 1$ vector of error terms $\boldsymbol{\upsilon}_j$ in Equation 9.8. Error terms are assumed to be spatially correlated and are represented as a generic spatial autoregressive process (SAR) of order q. The $\mathrm{SAR}(q)$ model expression is given in the last two expressions in Equation 9.8, where $\mathbf{I}$ is the $p \times p$ identity matrix, $\boldsymbol{\varepsilon}_j$ is a $p \times 1$ vector of independently and normally distributed errors $\left[\boldsymbol{\varepsilon}_j \sim MN\left(\mathbf{0},\sigma^2\mathbf{I}\right)\right]$ and $\mathbf{a}'_j = \left[a_{1j}\dots a_{sj}\dots a_{qj}\right]$ is the vector of the coefficients of the $\mathrm{SAR}(q)$ model for the jth profile, which is assumed to be normally distributed with mean $\boldsymbol{\alpha}' = \left[\alpha_1 \dots \alpha_s \dots \alpha_q\right]$ and covariance matrix $\mathbf{A}$ $\left[\mathbf{a}_j \sim MN\left(\boldsymbol{\alpha}, \mathbf{A}\right)\right]$.

The $p \times p$ matrix $\mathbf{W}^{(s)}$ of elements $w^{(s)}(k,t)$ $(k,t = 1,2,\dots p)$ represents the core of the small-scale model, since it is the sth-order adjacency matrix $w^{(1)}(k,t)$. For example, for $s = 1$, $\mathbf{W}^{(1)}$ is the first-order adjacency matrix and the element $w^{(1)}(k,t)$ is set equal to 1 if the tth point is the neighbor of the kth point and is set to 0 otherwise. Analogously, the element of the second-order adjacency matrix, $w^{(2)}(k,t)$, is set equal to 1 if the tth point is a neighbor of the original first-generation neighbors of the kth point, and so on. By definition, all the adjacency matrices are binary and symmetric matrices whose diagonal elements are zero (Cressie 1993).

For each profile, two vectors of coefficients $\mathbf{b}_j$ and $\mathbf{a}_j$ are considered. In order to let the model have the most general form, we further assume that these two vectors could also be correlated, *i.e.*, $\operatorname{cov}\left(\mathbf{b}_j, \mathbf{a}_j\right) = \mathbf{D}$ (Colosimo and Pacella 2010). In other words, with reference to the parametric model structure given in Equation 9.8, we merge the two vectors characterizing the observed pattern into a single coefficient vector related to the j th profile:

$$\begin{aligned} &\mathbf{c}'_j = \left[\mathbf{b}'_j \quad \mathbf{a}'_j\right] = \left[b_{1j} \dots b_{lj} \dots b_{rj} \quad a_{1j} \dots a_{sj} \dots a_{qj}\right], \\ &\mathbf{c}_j \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \qquad \text{where} \quad \boldsymbol{\mu}' = \left[\boldsymbol{\beta}' \quad \boldsymbol{\alpha}'\right], \ \boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{B} & \mathbf{D} \\ \mathbf{D}' & \mathbf{A} \end{bmatrix}. \end{aligned} \tag{9.9}$$

9.3.1.1 Control Limits of the SARX-based Control Chart

The signature model for the j th profile shown in Equation 9.8 is completely defined by a SARX model which requires one to estimate the $d = r + q$ parameters which are components of the vector $\mathbf{c}_j$, besides the residual variance σ^2. Let $\hat{\mathbf{c}}'_j = \left[\hat{\mathbf{b}}'_j \quad \hat{\mathbf{a}}'_j\right]$ represent the vector of the estimates of the d parameters for the j th profile. A T^2 control chart can be designed with reference to the statistics,

$$T_j^2 = \left(\hat{\mathbf{c}}_j - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\hat{\mathbf{c}}_j - \boldsymbol{\mu}\right), \tag{9.10}$$

where $j = 1,2,\dots$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are, respectively, the mean vector and the covariance matrix of the d-dimensional vectors of the coefficients. Here, the control chart parameters are assumed to be known or estimated from a large data set of in-control profiles.

Williams *et al.* (2006) studied the performance of different control limits to be used for T_j^2. In particular, when the number of samples *n* is at least twice the number of parameters estimated $\left(d+\left[d\left(d+1\right)/2\right]\right)$, the upper control limit based on the asymptotic distribution of T^2 can be used:

$$\mathrm{UCL} = \chi^2_{\alpha,d} , \tag{9.11}$$

where $\chi^2_{\alpha,d}$ is the $100(1-\alpha)$ percentile of the χ^2 distribution with *d* degrees of freedom.

The $p \times 1$ vector of estimated residuals associated with the *j* th profile can be described as

$$\mathbf{e}_j = \left(\mathbf{I}-\mathbf{R}_j\right)\left(\mathbf{y}_j - \mathbf{X}\hat{\mathbf{b}}_j\right), \tag{9.12}$$

where $\mathbf{R}_j = \sum_{s=1}^{q} \hat{a}_{sj}\mathbf{W}^{(s)}$. Hence, the estimated variance of residuals s_j^2 is given by

$$s_j^2 = \frac{\mathbf{e}_j'\mathbf{e}_j}{p-1}. \tag{9.13}$$

In order to monitor the residual variance, a traditional Shewhart-type control chart can be used. In particular, the control limits used to monitor the residual variance can be based on the χ^2 distribution with $p-1$ degrees of freedom, where p is the number of points monitored. If we denote by σ^2 the variance of the residuals, the control limits can be computed as follows:

$$\begin{aligned} \mathrm{UCL} &= \frac{\sigma^2}{p-1}\chi^2_{\alpha/2,p-1} , \\ \mathrm{CL} &= \sigma^2, \\ \mathrm{LCL} &= \frac{\sigma^2}{p-1}\chi^2_{1-(\alpha/2),p-1} , \end{aligned} \tag{9.14}$$

where $\chi^2_{\alpha/2,p-1}$ and $\chi^2_{1-(\alpha/2),p-1}$ are the upper and lower $\alpha/2$ percentage points of the χ^2 distribution with $p-1$ degrees of freedom associated with the residuals (Montgomery 2000).

9.3.1.2 A Numerical Example

The case study used as a reference was described in detail by Colosimo *et al.* (2008). It consists of items obtained by turning (cutting speed, 163 m/min; feed rate, 0.2 mm/rev; two steps of cutting depth 1 mm) C20 carbon steel cylinders, where each item was characterized by a roundness profile of 748 evenly distributed measurements of its radius. The original measurements sampled using a coordinate measuring machine were scaled by subtracting the least-squares estimate of the radius and centered at the least-squares estimate of the center. A further step was eventually applied to register all the sampled profiles by minimizing the phase delay caused by the random contact angle (Colosimo and Pacella 2007) (see Chapter 11).

Starting from actual measurements, the general SARX model of Equation 9.8 was fitted to the data. In this case, $\mathbf{y}_j$ represents the column vector of the radial deviation from the nominal radius measured at the angular position $\theta_k = k(2\pi/p)$, where $k = 1,2,\ldots p$ $(p = 748)$ is the index of equally spaced observations on each profile.

Consider the lth column vector of matrix $\mathbf{X}$ in Equation 9.8, and denote such a $p \times 1$ vector by $\mathbf{x}_l$, where $l = 1,2,\ldots r$. In general, each element of vector $\mathbf{x}_l$ is described as a function of the index location k. In the specific case of roundness profiles, $x_l(k)$ can be expressed either as $x_l(k) = \cos\left[(k-1)f_h\right]$ or as $x_l(k) = \sin\left[(k-1)f_h\right]$, *i.e.*, as a sinusoidal function of frequency equal to $f_h = h(2\pi/p)$ rad per sample. h is the frequency $(h \in \{1,2,\ldots p/2\})$ measured in undulations per revolution, which is fixed for all the elements of $\mathbf{x}_l$.

Two harmonics were selected to model the radial deviations in the actual test case, namely, the second and the third harmonics. Indeed, the process signature was mainly affected by ovality and triangularity (Moroni and Pacella 2008). The oval contour was possibly due to a bilobe error motion affecting the spindle's lathe or to eccentricity caused by an improper setup, while the three-lobe pattern was due to a similar error motion of the spindle.

Therefore, the regressor matrix $\mathbf{X}$ in Equation 9.8 has four columns $(r = 4)$ since two sinusoidal functions are needed to model the amplitude and phase of each specific harmonic. The kth row of matrix $\mathbf{X}$ is equal to $\cos\left[(k-1)f_2\right]$ $\sin\left[(k-1)f_2\right]$ $\cos\left[(k-1)f_3\right]$ $\sin\left[(k-1)f_3\right]$, where $f_h = h(2\pi/p)$ represents the frequency in radians per sample of the hth harmonic $(h = 2,3)$.

The vector of random error $\boldsymbol{\upsilon}_j$ in Equation 9.8 was fitted as a SAR model of order 2 $(q = 2)$ (Colosimo *et al.* 2008), using the algorithm implemented in the Spatial Econometrics toolbox (LeSage 1999).

With reference to the model of Equation 9.8, each specific profile of index j is associated with the vector of $d = p + q = 4 + 2$ parameters $\mathbf{c}_j \sim N(\boldsymbol{\mu},\boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = [\boldsymbol{\alpha}' \quad \boldsymbol{\beta}']$ and $\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{B} & \mathbf{D} \\ \mathbf{D}' & \mathbf{A} \end{bmatrix}$. Values of the vector and the matrices for the actual roundness data are shown in Table 9.3 (Colosimo *et al.* 2008).

Table 9.3 Parameters characterizing the mean $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\alpha}' & \boldsymbol{\beta}' \end{bmatrix}$ and the variance $\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{B} & \mathbf{D} \\ \mathbf{D}' & \mathbf{A} \end{bmatrix}$ of the distribution of coefficients $\mathbf{c}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the actual roundness data (Colosimo *et al.* 2008)

$$\boldsymbol{\beta}' = \begin{bmatrix} -0.0341 & 0.0313 & 0.0080 & -0.0322 \end{bmatrix}$$

$$\boldsymbol{\alpha}' = \begin{bmatrix} 0.3021 & 0.2819 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 4.07\times10^{-4} & -2.02\times10^{-4} & 6.54\times10^{-5} & 2.65\times10^{-5} \\ -2.02\times10^{-4} & 3.90\times10^{-4} & 1.49\times10^{-4} & 6.10\times10^{-6} \\ 6.54\times10^{-5} & 1.49\times10^{-4} & 2.24\times10^{-4} & -1.07\times10^{-5} \\ 2.65\times10^{-5} & 6.10\times10^{-6} & -1.07\times10^{-5} & 3.12\times10^{-4} \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} -8.84\times10^{-5} & -2.41\times10^{-4} \\ -1.21\times10^{-4} & 1.96\times10^{-4} \\ -1.18\times10^{-4} & 5.96\times10^{-5} \\ -1.50\times10^{-4} & -3.72\times10^{-4} \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 3.80\times10^{-3} & 1.59\times10^{-3} \\ 1.59\times10^{-3} & 4.32\times10^{-3} \end{bmatrix}$$

In order to show how the approach works, consider a group of $n = 12$ roundness profiles, where the j th profile is simulated according to the SARX model of Equations 9.8 and 9.9, where the coefficients $\mathbf{c}_j$ are randomly generated by a multinormal with the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given in Table 9.3. In particular, each profile data set consists of $p = 748$ measurements observed at a fixed set of locations, as in the real case. Figure 9.7 graphically depicts this data set. The estimates of the SARX model parameters are summarized in Table 9.4. In the specific case of roundness profiles, sinusoidal functions are used to characterize the observed data. Furthermore, the correlation structure is modeled by fitting a SAR model of order 2 using the algorithm implemented in the Spatial Econometrics toolbox (LaSage 1999). The residuals $e_j(k)$ of the model for each sample j at the location of index k are depicted in Figure 9.8. The parameters in Table 9.4 are used to design a phase II control chart. The T_j^2 statistics are computed for each of the 12 samples depicted in Figure 9.9.

Since T_j^2 follows a central χ^2 distribution with six degrees of freedom, the upper control limit of this chart that is used is based on Equation 9.11 and consists of a $100(1-\alpha)$ percentile of the χ^2 distribution with six degrees of freedom (in the reference case study, $\chi^2_{2,\alpha} = 12.59$). The numerical values of the T_j^2 statistics are summarized in Table 9.5.

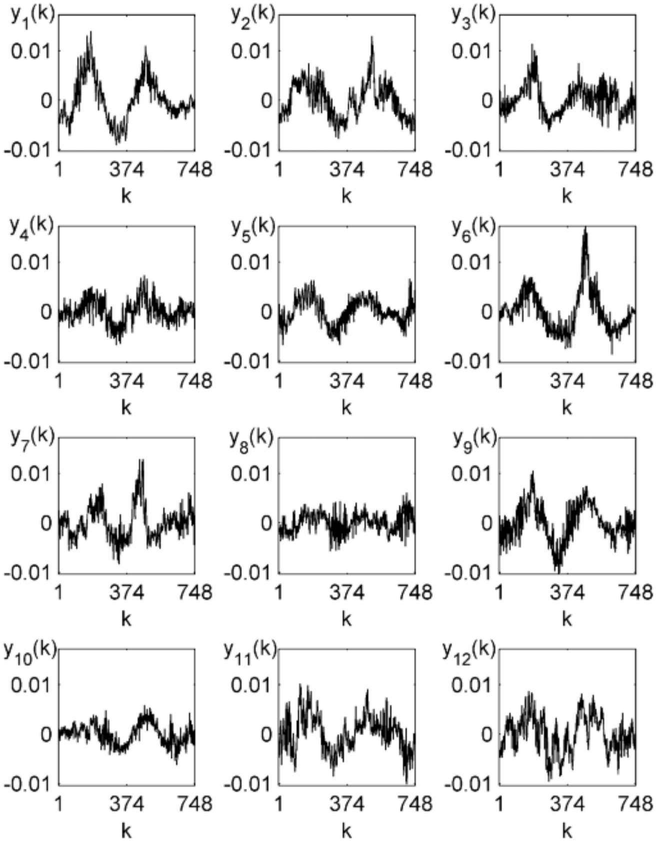


Figure 9.7 Twelve samples of a roundness profile, each of 748 points. Each point represents a functional relationship between an independent variable (*abscissa*) and a dependent one (*ordinate*). The dependent variable is the radial deviation from the nominal radius measured at the angular position. The independent variable is the index of equally spaced observations on each profile

Table 9.4 Spatial autoregressive regression (SARX) model parameters estimated for a group of 12 roundness profiles shown in Figure 9.7

Profile index	$\hat{b}_{1j}$	$\hat{b}_{2j}$	$\hat{b}_{3j}$	$\hat{b}_{4j}$	$\hat{a}_{1j}$	$\hat{a}_{2j}$
1	–0.0672	0.0601	0.0110	–0.0591	0.4070	0.3200
2	–0.0640	0.0418	–0.0016	–0.0010	0.4275	0.3205
3	–0.0327	0.0299	–0.0175	–0.0270	0.2940	0.3235
4	–0.0177	0.0217	0.0109	–0.0299	0.2535	0.2940
5	–0.0245	0.0364	–0.0035	–0.0290	0.3350	0.2875
6	–0.0358	0.0758	0.0369	–0.0386	0.3240	0.3515
7	–0.0074	0.0097	0.0168	–0.0572	0.4130	0.3845
8	–0.0104	0.0021	0.0106	–0.0246	0.3515	0.1370
9	–0.0458	0.0525	0.0146	–0.0476	0.1945	0.2090
10	–0.0137	0.0329	0.0217	–0.0127	0.2380	0.2360
11	–0.0481	0.0428	–0.0115	–0.0205	0.4030	0.3395
12	–0.0351	0.0490	0.0189	–0.0216	0.3950	0.3875

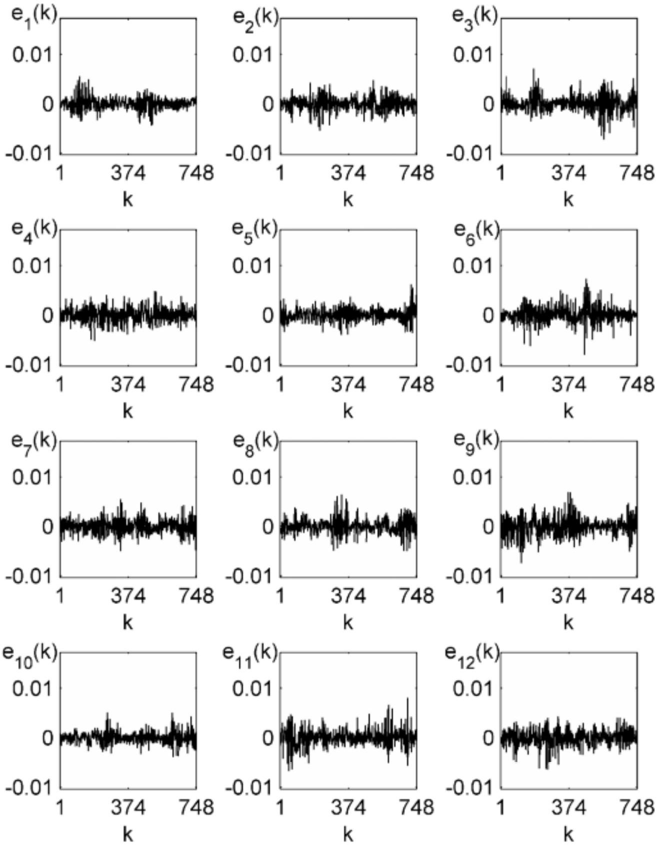


Figure 9.8 Residual errors between observed points depicted in Figure 9.7 and the SARX models for each roundness profile

Table 9.5 T^2 statistics for the 12 roundness profiles shown in Figure 9.7

Profile index	T^2 statistic
1	8.6141
2	10.6301
3	7.4682
4	1.7891
5	4.7382
6	7.6785
7	8.3787
8	10.4605
9	6.3544
10	3.2439
11	7.9012
12	5.9817

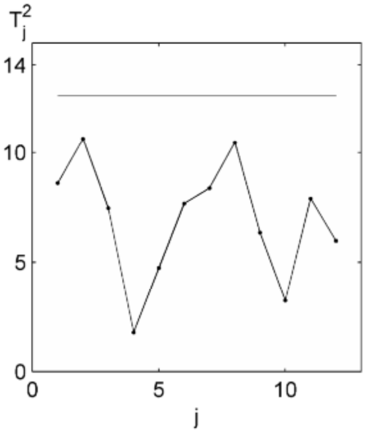


Figure 9.9 Multivariate control chart for the regression-based approach applied to the 12 randomly generated roundness profile data sets shown in Figure 9.7

9.3.2 The PCA-based Model

Ramsay and Silverman (2005) presented an extension of PCA to functional data, *i.e.*, an approach which allows one to find a set of orthonormal functions (also called functional principal components, PCs), so that the original data can be approximated in terms of a linear combination of these basis functions.

In particular, Ramsay and Silverman (2005) showed that, in the case of equally spaced observations, the easiest way to compute the PCs consists in modeling the curve data sampled at regular intervals as a multivariate vector, and performing a traditional PCA on the set of samples collected over different curves.

When the PCA outlines a set of significant PCs to be retained, the coefficients (or loadings) defining these significant PCs can be interpreted as eigenfunctions (also called empirical orthogonal functions). These eigenfunctions do not have a parametric expression and are empirical, since they are obtained from the data at hand. A rough sketch of how PCA works is given in the following.

Assume we organize a sample of n vectors of $p \times 1$ profile data $\mathbf{y}_j$ $(j = 1,2,\ldots n)$ into an $n \times p$ data matrix $\mathbf{Y}$ whose j th row is equal to the transpose of the j th data vector $\mathbf{y}_j$. PCA consists in performing a spectral decomposition of the covariance matrix of $\mathbf{Y}$. The covariance matrix describes the variability of the data observed at each location with respect to the mean value observed at the same location in all the profiles. Therefore, a first step in PCA consists in centering the data by subtracting the average profile.

If $\mathbf{S}$ is the covariance matrix, *i.e.*, $\mathbf{S} = \left[1/(n-1)\right]\sum_{j=1}^{n}\left(\mathbf{y}_j - \overline{\mathbf{y}}\right)\left(\mathbf{y}_j - \overline{\mathbf{y}}\right)'$, where $\overline{\mathbf{y}} = (1/n)\sum_{j=1}^{n}\mathbf{y}_j$ is the sample mean profile, the spectral decomposition consists in finding the $p \times p$ matrices $\mathbf{U}$ and $\mathbf{L}$ that satisfy the following relationship:

$$\mathbf{U}'\mathbf{S}\mathbf{U} = \mathbf{L}\,, \tag{9.15}$$

where $\mathbf{L}$ is a diagonal matrix that contains the eigenvalues of $\mathbf{S}$ (say, l_k), while $\mathbf{U}$ is an orthonormal matrix whose *k*th column $\mathbf{u}_k$ is the *k*th eigenvector of $\mathbf{S}$ (the so-called loadings).

With reference to the j th profile $\mathbf{y}_j$, we denote by $\mathbf{z}_j$ the following vector

$$\mathbf{z}_j = \mathbf{U}'\left(\mathbf{y}_j - \overline{\mathbf{y}}\right) = \left[z_{j1} \ldots z_{jk} \ldots z_{jp}\right]', \tag{9.16}$$

where z_{jk} are the so-called scores. Each profile can then be expressed as a linear combination of loadings $\mathbf{u}_k$, where the weights of the linear combination are the scores z_{jk}:

$$\mathbf{y}_j = \overline{\mathbf{y}} + z_{j1}\mathbf{u}_1 + z_{j2}\mathbf{u}_2 + \cdots + z_{jp}\mathbf{u}_p\,. \tag{9.17}$$

Since the PCs are statistically uncorrelated and each PC has variance equal to the corresponding eigenvalue (l_k), we can rank the PCs according to the associated eigenvalue and decide to retain just the most important ones (*i.e.*, the ones which are associated with larger variances). Different approaches can be used to select the proper set of PCs (Jackson 2003; Jolliffe 2002). For instance, cross-validation can be effectively used to choose the number *m* of significant PCs (Colosimo and Pacella 2007). When a subset *m* of the whole number of *p* PCs is retained $(m < p)$, the original profile can be estimated as

$$\hat{\mathbf{y}}_{j(m)} = \overline{\mathbf{y}} + z_{j1}\mathbf{u}_1 + z_{j2}\mathbf{u}_2 + \cdots + z_{jm}\mathbf{u}_m\,. \tag{9.18}$$

9.3.2.1 Control Limits of PCA-based Control Charts

Similar to the regression-based approach, also in the case of PCA a T^2 control chart can be used for monitoring the vector of the first m retained PCs (MacGregor and Kourti 1995). In this case, the Hotelling statistic is given by (Jackson 2003)

$$T_j^2 = \frac{z_{j1}^2}{l_1} + \frac{z_{j2}^2}{l_2} + \cdots + \frac{z_{jm}^2}{l_m}\,. \tag{9.19}$$

If an unexpected event leads the process to change in a direction orthogonal to that of the first m PCs, the control chart will not be able to issue an alarm. For this reason, another control chart based on the Q statistic (sometimes referred to as the squared prediction error control chart) also has to be used (Jackson 2003).

Given the estimate in Equation 9.18, the Q statistic can be computed as the sum of the squared errors obtained by reconstructing each observation by the first m PCs:

$$Q_j = \left(\mathbf{y}_j - \hat{\mathbf{y}}_{j(m)}\right)' \left(\mathbf{y}_j - \hat{\mathbf{y}}_{j(m)}\right). \tag{9.20}$$

The upper control limit of the T^2 statistics in Equation 9.19 can be computed as (Williams *et al.* 2006)

$$\text{UCL} = \chi^2_{\alpha,m}, \tag{9.21}$$

where $\chi^2_{\alpha,m}$ is the $100(1-\alpha)$ percentile of the χ^2 distribution with m degrees of freedom. With reference to the Q statistic, according to Nomikos and MacGregor (1995), the upper control limit can be computed as

$$\text{UCL} = g\chi^2_{\alpha,h}, \tag{9.22}$$

where g and h can be estimated as $\hat{g} = \hat{\sigma}^2_Q \big/ \left(2\bar{Q}\right)$ and $\hat{h} = 2\bar{Q}^2 \big/ \hat{\sigma}^2_Q$, while $\bar{Q}$ and $\hat{\sigma}^2_Q$ are the sample mean and the sample variance obtained by computing the Q statistics via Equation 9.20 for the set of n profiles.

9.3.2.2 A Numerical Example

With reference to the aforementioned case study of roundness profiles (see Section 9.3.1.2), Colosimo and Pacella (2007) used a cross-validation approach in order to determine the number of significant PCs to be retained. With reference to the roundness profiles of the reference case study, the number of significant PCs is equal to $m = 3$.

Once the PCA has been performed, the retained PCs should be interpreted to gain more insight into the systematic pattern characterizing the machined profiles. To this aim, each eigenfunction $\mathbf{u}_k$ (*i.e.*, the coefficients of each eigenvector, also known as loadings) can be graphically represented as a function of the location. Following this practice, Figure 9.10 reports the diagrams of the first three eigenfunctions ($\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$), which are related to the three retained PCs (Colosimo and Pacella 2007). The first PC, which describes the most important component of variability (13.87%), has a bilobe form. This qualitative observation indicates that the main variability around the mean profile is due to a periodic function characterized by a frequency of two undulations per revolution. The second PC (which accounts for the 10.41% of the total variability) is a mixture of a bilobe and a trilobe contour. This mixture is obtained by combining two periodic functions, namely, harmonics with two and three undulations per revolution. Eventually, the third PC (which accounts for 7.10% of the total variability) has a trilobe contour.

Assume we consider a subgroup of n = 12 roundness profiles, where each profile consists of p = 748 measurements observed at a fixed set of locations. In particular, the 12 samples depicted in Figure 9.7 are considered. A PCA-based control chart can be used to monitor the vector of the first *m* retained PCs. In this case, the plotted statistics are based on the values of the eigenvalues associated with the three retained PCs. With reference to the case study, these values are $l_1 = 0.59 \times 10^{-3}$, $l_2 = 0.44 \times 10^{-3}$, and $l_3 = 0.30 \times 10^{-3}$.

The scores z_{j1}, z_{j2}, and z_{j3} for each profile of index $j = 1, 2, \ldots 12$ are reported in Table 9.6. In particular, the T_j^2 statistics computed for each of the 12 samples is based on Equation 9.19. Since T_j^2 follows a central χ^2 distribution with a number of degrees of freedom equal to the number of retained PCs $(m = 3)$, the upper control limit of this chart that is used is based on Equation 9.21 and consists of a $100(1-\alpha)$ percentile of the χ^2 distribution with three degrees of freedom (in the reference case study, $\chi^2_{3,\alpha} = 7.81$). The numerical values of the T_j^2 statistics are summarized in Table 9.7, while the T^2 control chart is depicted in Figure 9.11.

Table 9.6 Scores associated with the first three retained principal components (PCs) for each profile

Profile index	z_{j1}	z_{j2}	z_{j3}
1	–0.0437	0.0032	–0.0290
2	–0.0264	0.0317	0.0318
3	0.0083	0.0172	0.0087
4	0.0182	–0.0097	0.0063
5	–0.0109	–0.0181	–0.0206
6	0.0031	0.0022	0.0073
7	–0.0412	–0.0375	–0.0062
8	0.0392	–0.0247	–0.0252
9	0.0353	–0.0043	0.0101
10	–0.0237	–0.0058	–0.0121
11	0.0085	–0.0174	0.0234
12	–0.0140	0.0134	0.0072

Table 9.7 T^2 statistics for roundness profiles: case of known in-control parameters (phase II)

Profile index	Case of known in-control parameters
1	6.0988
2	6.8561
3	1.0437
4	0.9142
5	2.3695
6	0.2034
7	6.2359
8	6.1274
9	2.5097
10	1.5229
11	2.6437
12	0.9171

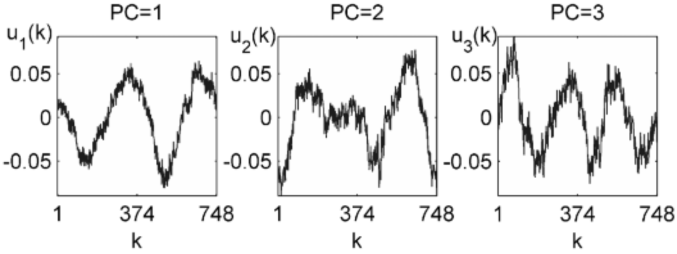


Figure 9.10 The first three eigenfunctions are related to the three retained PCs: **a** The first PC (bilobe contour) describes 13.87% of the variability, **b** the second PC (a mixture of a bilobe and a trilobe contour) accounts for the 10.41% of the total variability, and **c** the third PC (trilobe contour) accounts for 7.10% of the total variability

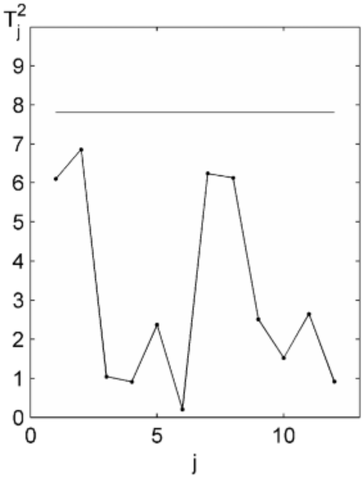


Figure 9.11 PC-analysis-based multivariate control chart assuming known in-control parameters (phase II)

9.4 Conclusions

Data collected by measuring equipment can be modeled as functional data, where the quality outcome (dependent variable) is a function of one or more location variables (independent variables). We presented an approach where all the profiles/surfaces associated with geometric tolerances are either modeled as functional data by using regression models (with spatially correlated errors) or modeled using PCA. Then, all the coefficients related to the selected model of the functional data are monitored via control charting. This procedure allows one to quickly detect an out-of-control signal.

Compared with other approaches for monitoring geometric tolerances (such as the one presented in Chapter 8), the approach based on functional data (or profile) monitoring has the main advantages of reducing the time to detect an out-of-control signal and aiding in the diagnosis of the type of problem behind the unnatural pattern. Colosimo *et al.* (2008) showed that moving from the industrial practice of monitoring synthetic geometric tolerance indicators (see Chapter 8) to approaches based on functional data (such as the ones presented in this chapter) can reduce the time to detect out-of-control signals by 60–70%, as shown in Chapter 11. The application of profile monitoring to quality control of geometric tolerances is a very promising area of future research.

Acknowledgements This research was partially funded by the Italian Ministry of University and Research grant PRIN 2005 – "Characterization of the multiscale geometry of manufactured surfaces for process control and tolerance analysis".

References

Chang TC, Gan FF (2006) Monitoring linearity of measurement gauges. J Stat Comput Simul 76:889–911

Colosimo BM, Pacella M (2007) On the use of principal component analysis to identify systematic patterns in roundness profiles. Qual Reliab Eng Int 23:707–725

Colosimo BM, Pacella M (2010) Control Charts for Statistical Monitoring of Functional Data, Int J Prod Res 48(6):1575–1601

Colosimo BM, Pacella M, Semeraro Q (2008) Statistical process control for geometric specifications: on the monitoring of roundness profiles. J Qual Technol 40:1–18

Colosimo BM, Mammarella F, Petrò S (2010) Quality control of manufactured surfaces. In: Lenz HJ, Wilrich PT (eds) Frontiers of statistical quality control, vol 9. Springer, Vienna

Cressie NAC (1993) Statistics for spatial data. Wiley, New York

Ding Y, Zeng L, Zhou S (2006) Phase I analysis for monitoring nonlinear profiles in manufacturing processes. J Qual Technol 38:199–216

Gupta S, Montgomery DC, Woodall WH (2006) Performance evaluation of two methods for online monitoring of linear calibration profiles. Int J Prod Res 44:1927–1942

Jackson JE (2003) A user's guide to principal components. Wiley, New York

Jensen WA, Birch JB, Woodall WH (2008) Monitoring correlation within linear profiles using mixed models. J Qual Technol 40:167–183

Jin J, Shi J (1999) Feature-preserving data compression of stamping tonnage information using wavelets. Technometrics 41:327–339

Jin J, Shi J (2001) Automatic feature extraction of waveform signals for in-process diagnostic performance improvement. J Intell Manuf 12:257–268

Jolliffe IT (2002) Principal component analysis. 2nd edn. Springer, New York

Kang L, Albin SL (2000) On-line monitoring when the process yields a linear profile. J Qual Technol 32:418–426

Kim K, Mahmoud MA, Woodall WH (2003) On the monitoring of linear profiles. J Qual Technol 35:317–328

LeSage JP (1999). The theory and practice of spatial econometrics. Available via http://www.spatialeconometrics.com/

MacGregor JF, Kourti T (1995) Statistical process control of multivariate processes. Control Eng Pract 3:403–414

Mahmoud MA, Woodall WH (2004) Phase I analysis of linear profiles with calibration applications. Technometrics 46:377–391

Mahmoud MA, Parker PA, Woodall WH, Hawkins DM (2007) A change point method for linear profile data. Qual Reliab Eng Int 23:247–268

Montgomery DC (2000) Introduction to statistical quality control, 4th edn. Wiley, New York

Moroni G, Pacella M (2008) An approach based on process signature modeling for roundness evaluation of manufactured items. J Comput Inf Sci Eng 8:021003

Nomikos P, MacGregor JF (1995) Multivariate SPC charts for monitoring batch processes. Technometrics 37:41–59

Ramsay JO, Silverman BW (2005) Functional data analysis. 2nd edn. Springer, New York

Stover FS, Brill RV (1998) Statistical quality control applied to ion chromatography calibrations. J Chromatogr A 804:37–43

Walker E, Wright SP (2002) Comparing curves using additive models. J Qual Technol 34:118–129

Whittle P (1954) On stationary processes in the plane. Biometrika 41:434–449

Williams JD, Woodall WH, Birch JB, Sullivan JH (2006) Distribution of Hotelling's T2 statistic based on the successive differences estimator. J Qual Technol 38:217–229

Williams JD, Woodall WH, Birch JB (2007) Statistical monitoring of nonlinear product and process quality profiles. Qual Reliab Eng Int 23:925–941

Woodall WH (2007) Current research on profile monitoring. Producao 17:420–425

Woodall WH, Spitzner DJ, Montgomery DC, Gupta S (2004) Using control charts to monitor process and product quality profiles. J Qual Technol 36:309–320

Young TM, Winistorfer PM, Wang S (1999) Multivariate control charts of MDF and OSB vertical density profile attributes. For Prod J 49:79–86

Zhang H, Albin S (2009) Detecting outliers in complex profiles using a 2 control chart method IIE Trans 41:335–345

Zou CL, Zhang YJ, Wang ZJ (2006) A control chart based on a change-point model for monitoring linear profiles. IIE Trans 38:1093–1103

Zou CL, Tsung FG, Wang ZJ (2008) Monitoring profiles based on nonparametric regression methods. Technometrics 50:512–526

Chapter 10
A Model-free Approach for Quality Monitoring of Geometric Tolerances

Massimo Pacella, Quirico Semeraro, Alfredo Anglani

Abstract Profile monitoring can be effectively adopted to detect unnatural behaviors of machining processes, *i.e.*, to signal when the functional relationship used to model the geometric feature monitored changes with time. Most of the literature concerned with profile monitoring deals with the issue of model identification for the functional relationship of interest, as well as with control charting of the model parameters. In this chapter, a different approach is presented for profile monitoring, with a focus on quality monitoring of geometric tolerances. This approach does not require an analytical model for the statistical description of profiles considered, and it does not involve a control charting method. An algorithm which allows a computer to automatically learn from data the relationship to represent profiles in space is described. The proposed algorithm is usually referred to as a neural network and the data set, from which the relationship is learned, consists just of profiles representative of the process in its in-control state. Throughout this chapter, a test case related to roundness profiles obtained by turning and described in Chapter 11 is used as a reference. A verification study on the efficacy of the neural network shows that this approach may outperform the usual control charting method.

M. Pacella
Dipartimento di Ingegneria dell'Innovazione, Università del Salento,
Via per Monteroni, 73100 Lecce, Italy,
e-mail: massimo.pacella@unisalento.it

Q. Semeraro
Dipartimento di Meccanica, Politecnico di Milano,
Via la Masa, 20156 Milan, Italy,
e-mail: quirico.semeraro@polimi.it

A. Anglani
Dipartimento di Ingegneria dell'Innovazione, Università del Salento,
Via per Monteroni, 73100 Lecce, Italy,
e-mail: alfredo.anglani@unisalento.it

10.1 Introduction

When the quality of a manufactured product is related to a geometric tolerance, the process should be represented by a variable which is functionally dependent on one or more explanatory variables (*functional data*). Hence, we observe a set of values over a range which, when plotted, takes the shape of a curve or profile. The process should be considered in control if this functional relationship is stable with time. When the process moves into an out-of-control state, the observed profiles should show the signature of this change and a proper approach should be used to signal unusual patterns.

The issue of monitoring profiles has been defined as being the most promising area of research in statistical process control. Approaches for profile monitoring share a common structure (Woodall *et al.* 2004), which consists in (1) identifying a suitable parametric model of the functional relationship monitored and (2) estimating the model parameters from in-control data; (3) designing a multivariate control chart of the estimated parameters and a univariate control chart of the variability of the residuals. The studies presented in the literature can be classified with reference to the modeling method considered: mainly linear regression (Kang and Albin 2000; Kim *et al.* 2003; Mahmoud and Woodall 2004; Gupta *et al.* 2006) or approaches for multivariate data reduction such as principal component analysis (Jones and Rice 1992; Walker and Wrigth 2004; Colosimo and Pacella 2007).

However, the identification of a suitable model of profile data may become a cumbersome activity, thus representing an obstacle to the use of profile monitoring in real applications. For example, when the quality characteristic of interest is a roundness profile, a possibility consists in modeling the radial deviations by fitting a Fourier series, *i.e.*, using sinusoidal functions at several frequencies (called harmonics), which play the role of explanatory variables in space. The issue of model fitting using harmonics was investigated in Chapter 9. Irrespective of the number of harmonics included in the model, the residuals obtained are affected by autocorrelation. Therefore, a more involved approach was exploited, in which the harmonic regression was combined with a spatial autoregressive model of the residual errors.

Techniques for multivariate data reduction, in particular those based on principal component analysis, may be useful in these cases because they have been shown to be effective in dealing with profiles without requiring a specific model type selection, *i.e.*, a suitable set of explanatory variables in space as well as a proper autoregressive model of the residuals. However, principal component analysis may sometimes fail in identifying some structured pattern in profile data. Furthermore, the method can require an extra effort both to manage the abstract nature of principal components and to design a suitable control chart for the residuals (Colosimo and Pacella 2010).

These modeling issues behind the approaches for profile monitoring give rise to the need to develop a complementary method which can be easily implemented by quality practitioners. This is the aim in the present chapter, in which a different approach is discussed for profile monitoring. This approach does not require an analytical model for the statistical description of the profiles considered, and it does not involve control charting. In particular, an algorithm which allows computers to automatically learn from data the relationship to represent the profiles in space, is described. This algorithm is commonly referred to as a neural network and the data set from which the relationship is automatically learned consists just of profiles representative of the process in its natural, or in-control, state. Furthermore, the neural network can produce a signal when an input profile does not fit, according to a specific criterion, to the learned relationship to represent the profiles in space.

The approach presented in this chapter can be exploited for profile monitoring when a geometric tolerance is the quality characteristic of interest. The neural network is defined as a model-free approach because it can be exploited when an analytical model for the statistical description of the profiles considered is not available. Furthermore, learning from in-control data and profile monitoring can be automatically implemented with a computer. The practitioner just needs to set one input parameter of the computer algorithm, depending on the performance in terms of the false-alarm rate required for the actual application. The specific method proposed in the present chapter is based on the adaptive resonance theory (ART) neural network trained by an unsupervised approach. The proposed approach will be mainly applicable to quality characteristics which are related to a two-dimensional profile (*e.g.*, roughness, waviness, roundness, straightness).

An approach based on a ART neural network was first presented by Pacella and Semeraro (2007b) as an adaptation of their previous work, related to univariate quality characteristics (Pacella *et al.* 2004a, b), to the case of geometric tolerances on manufactured items, with a focus on monitoring roundness. Indeed, profile monitoring based on the use of neural networks is a promising area of research since a neural network, when compared with other profile monitoring approaches, has the main advantage of being easily implemented in practice. Furthermore, recent achievements in the field of neural network theory related to the development of new models able to handle functional data (Rossi and Conan-Guez 2005; Rossi *et al.* 2005) open interesting scenarios for future research on innovative approaches for automatic profile monitoring.

The remainder of this chapter is organized as follows. Section 10.2 gives a brief overview of machine learning and neural network theory. The expert reader can skip this section and go on to the next one, which provides a review of the literature on the general topic of neural networks for quality monitoring. A step-by-step description of the model-free approach based on the ART neural network is provided in Section 10.4. A verification study on the efficacy of the neural network approach for profile monitoring is discussed in Section 10.5. The last section provides the conclusions and some final remarks. An appendix is also included, which provides some specific details on the neural network algorithm.

10.2 An Introduction to Machine Learning

The increasing availability of affordable point-based measurement systems, capable of acquiring large amounts of data as point coordinates in reasonable times and with high accuracy, makes it increasingly possible to develop tools for quality monitoring of geometric tolerances capable of operating on richer information content and with more sophisticated approaches. Nowadays, the possibility of managing large data sets with specific *data analysis tools* paves the way to the development of innovative approaches in the field of quality monitoring.

Data measured as point coordinates on a machined profile are supposed to contain some useful information one wants to *learn* about the actual manufacturing process. The construction of a model for observed data points, which is often referred to as the *learning procedure*, encodes the structures, or *patterns*, in the data. "Pattern" is here a generic term intended to express any type of rule or dependency structure present in the data. These patterns represent the information about the system that one is looking for. In actual applications, *e.g.*, profile monitoring in manufacturing, data are often noisy measurements and the dependency expressed by patterns is not deterministic.

Pattern recognition is the basis of *learning theory*. However, since the observed data comprise a finite number of samples, there is a fundamental limitation regarding the information that can be learned from such data. The main difficulty of learning is that observed data give a partial and often noisy view of the system. Therefore, it is impossible to estimate the missing information with perfect certainty using only the observed data.

The aim of learning is to identify inherent patterns, *i.e.*, patterns that are exhibited by subsequent data collected from the system and not only by the observed data. A learning procedure that is able to identify inherent patterns generalizes well. The generalization property makes it possible to infer with some confidence the missing information of a partially observed state of the system. In contrast, if the learning procedure returns patterns that are present in the observed data but absent from other data collected from this system, it is said to be overfitting. When overfitting occurs, inference of missing information might be completely erroneous.

Designing *automatic* learning procedures, *i.e.*, algorithms that can be implemented by computers, for tackling practical data analysis tasks is often viewed as the subject of the *machine learning* research field. Usually, in this research field it is not sufficient to have good estimators from learning theory if their computational cost is excessive. Excessive computational costs might be due to the size of the data set, but working with highly flexible models can also lead to excessive costs. Furthermore, the fitted model should be able to be exploited for classification and/or prediction tasks. Finally, the usefulness of the learning procedure should be measured by a proper evaluation, which can be assessed by using a quantitative criterion (*e.g.*, in a classification task, by the proportion of correct classifications) or, when data analysis enters the preliminary stages of a study where it is not possible to use a quantitative measure, by a qualitative criterion (*e.g.*, how the learning procedure can help in obtaining understanding about the system).

There are essentially two types of learning procedure, referred to as *supervised* and *unsupervised* tasks, which are briefly discussed in the following subsection. For a more complete overview of learning theory, there are many standard references, such as the books by Vapnik (2000) and Bishop (2006). Also, the first part of the book by Bishop (1995) presents the subject of pattern recognition.

10.2.1 Supervised and Unsupervised Learning

The basic idea of *supervised learning* is that data samples are composed of an input and a target part. The target is usually a scalar value, while the input is often described by a vector. The goal of the supervised learning is to derive from the data set the dependency of targets on inputs, so that the model is capable of returning target predictions for new inputs.

When one has some knowledge of the structures that should be present in the data, a supervised learning procedure can be adopted. However, in practical applications there is no such prior knowledge of the rules in the system, either because it is too complex to characterize these rules or because the system is significantly stochastic. This is typically the case of manufacturing processes, where observed data are noisy, dependencies are highly stochastic, and there is no simple physical rule to represent them.

When one has little knowledge about the patterns present in the data, looking for *clusters* in the observed data is a good starting point. The second type of learning task is the unsupervised one, a.k.a. *clustering*. The goal of clustering is to identify inherent separations in the data. One could view clustering as a classification problem without label information. Data belonging to a cluster should be close to one another, while data from different clusters should be far apart. While in the case of supervised learning adjusting a model requires the definition and the minimization of a loss function, for unsupervised learning the loss function is a measure of the similarity of samples within each cluster. Different natural clustering would be found just by changing the similarity measure. Setting the number of clusters is also an important issue as it might depend on the similarity measure.

A task related to clustering is *quantization* (Graf and Luschgy 2000). The aim of quantization is to encode a large set of input vectors by finding a smaller set of representative prototypes (or templates). Each prototype provides a good approximation to the original data in a cluster by summarizing the specific characteristics of that cluster. The principle is to use a discrete set of prototypes to approach continuous stochastic vectors based on the minimization of an error cost function. That means the prototypes can represent the original distribution with the least error. The basic motivation of vector quantization is dimensionality reduction or data compression. Actually, the term "quantization" originates in the theory of signal processing. In this context, "quantization" means a process of discretizing signals.

10.2.2 Neural Networks

Neural networks are computer algorithms, typically thought of as black boxes, used in actual applications to learn specific knowledge, to adapt it to new situations, and to provide reliable classifications and approximations of data (Haykin 1998). A neural network algorithm implements a specific *learning procedure* and, hence, can be exploited to identify the fundamental functional relationship or pattern in data sets.

The adjective "*neural*" is due to the fact that these algorithms simulate in a very simplified form the ability of brain neurons to process information. The principle is to combine in a network simple processing functions, which are called "neurons" or "nodes", linked by weighted connections. The function of the synapse, the structure responsible for storing information in the brain, is modeled by a modifiable weight, which is associated with each connection between two neurons. Within each neuron all the weighted input signals are summed up and a signal is then produced as an output. In particular, the output signal is computed as the response of a *link function* (such as the hyperbolic tangent function) on the summation of the weighted input signals of that neuron. Neurons set in parallel form a layer of the network, while the output signal of a neuron is fed to the neurons in the subsequent layer as an input signal. A neural network is composed of successive layers, namely, the input layer, one or more hidden layers, and the output layer. In general, all neurons in a layer have the same link function and are fully connected to the neurons in adjacent layers. As an example, Figure 10.1a graphically depicts a general architecture for a three-layer neural network. The input layer contains four nodes used as input data for four data points. The output layer consists of two nodes. The single hidden layer of the neural network contains five nodes. Figure 10.1b depicts a model of a neuron with four input connections, where a specific weight is associated with each connection.

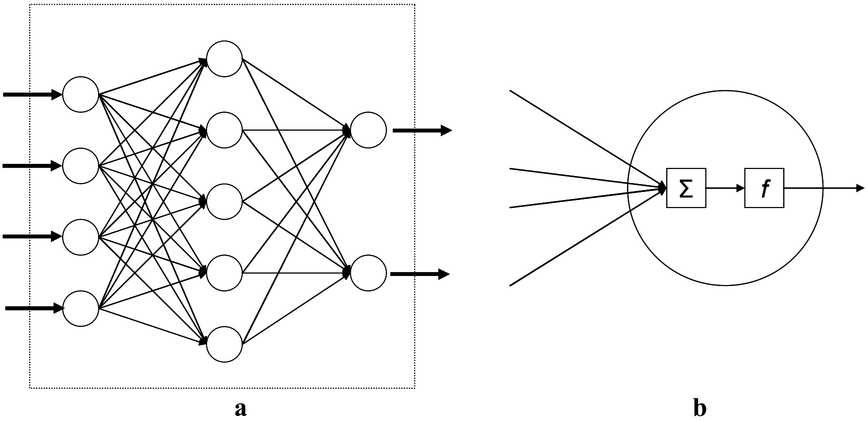


Figure 10.1 **a** The structure of a three-layer neural network, and **b** a model of a neuron with four input connections

In the following subsections, two neural networks are discussed, namely, the multilayer perceptron (MLP) and the ART. The MLP represents the common model used for supervised learning tasks, while the ART is related to unsupervised tasks.

10.2.3 Supervised Learning: the MLP Model

MLP is a flexible neural network model configured by setting the number of layers, the number of neurons in each layer, and the types of link functions.

The *training stage* of MLP corresponds to the optimization of the different layers of weights in order to accomplish a specific supervised learning task.

The success of MLP is basically due to the approximation property of this model. Indeed, theoretical results (Hornik *et al.* 1989) demonstrate that a MLP with one hidden layer of nonlinear processing functions (followed by a linear one) is capable of approximating "*any measurable function to any desired degree of accuracy*". According to the results of Hornik *et al.* (1989), "*... any lack of success in applications must arise from inadequate learning, insufficient numbers of hidden units or the lack of a deterministic relationship between input and target*".

The main drawback with MLP is that this model can easily become "too flexible" and hence it can be difficult to avoid overfitting of observed data (the network fits the observed data very well but generalizes very poorly). The overfitting problem (a.k.a. overtraining) is more likely to occur in MLPs than in other models owing to the large parameter set to be estimated. Indeed, the most difficult problem is how to develop a network of appropriate size for capturing the underlying patterns in the data. Although MLP theory suggests that more hidden nodes typically lead to improved accuracy in approximating a functional relationship, they also cause the problem of overfitting. Therefore, determination of how many hidden layers and hidden neurons to use is often chosen through experimentation or by trial and error.

Different weight elimination and node pruning methods have been proposed in the literature for building the optimal architecture of MLP (Reed 1993; Roy *et al.* 1993; Wang *et al.* 1994; Murata *et al.* 1994; Cottrell *et al.* 1995; Schittenkopf *et al.* 1997), but none of these methods can guarantee the best solution for every possible situation. The basic idea with these methods is to find a parsimonious model that fits the data well. Generally, a parsimonious model not only gives an adequate representation of the data, but also has the more important generalization capability.

Another way to tackle the overfitting problem is to divide the set of observed data into three subsets, namely, *training*, *validation*, and *testing* data. The training and validation parts are used for model building, and testing is used for evaluation of the model. In particular, the training set is used for computing and updating iteratively the network weights. During training, the error on the validation set is monitored, and this set of data is not used for updating the network weights. The validation error normally decreases during the initial phase of training, as does the training set error. However, when the network begins to overfit the training data, the error on the validation set begins to rise, while the training set error continues to decrease. When the validation error increases for a specified number of iterations, the training is stopped (*early stopping*). The network with the best performance on the validation set is then evaluated on the testing data set.

As previously mentioned, the overfitting problem is more likely to occur in a MLP model since it presents a large parameter set to be estimated. If the number of parameters in a network is much smaller than the total number of samples in the training set, then there is little or no chance of overfitting. Furthermore, overfitting does not apply to some neural network paradigms, such as the unsupervised ones (*e.g.*, the ART model) because they are not trained using an iterative process.

Recently, MLP models have been adapted to functional data. In particular, a few studies in which the extension of MLP to functional inputs was proposed from a theoretical point of view are reported in the literature. One of these approaches was first proposed by Stinchcombe (1999). Subsequently, Rossi and Conan-Guez (2005) proposed a functional MLP, although this model suffers from the need for a specialized implementation and from long training times. Rossi *et al.* (2005) proposed another functional MLP based on projection operators. This method has some advantages over a specialized implementation of MLP, especially because the projections can be implemented as a preprocessing step that transforms functions into adapted vector representations. The vectors obtained like this are then processed by a standard MLP model. In general, the functional MLP models proposed in the literature are interesting from a theoretical point of view. From all of these studies it appears that functional MLPs are a valuable tool for data analysis when a functional representation of input variables is possible.

10.2.4 Unsupervised Learning: the ART Model

ART is an algorithm able to cluster input vectors which resemble each other according to the stored prototypes. ART can adaptively create a new cluster corresponding to an input if this specific pattern is not similar to any existing prototype.

In a physical system when a small vibration of a proper frequency produces a large amplitude vibration, it is defined as resonance. Indeed, ART gets its name from the fact that information, *i.e.*, the output of neurons, reverberates back and forth between two layers, namely, F1 (the comparison layer) and F2 (the recognition layer), which are fully connected by weights. On the one hand, the comparison layer (F1) acts as a feature detector that receives external input; on the other hand, the recognition layer (F2) acts as a category classifier that receives internal patterns (a schematic representation of the ART model is depicted in Figure 10.2).

The application of a single input vector leads to a set of activity that the neural network develops in the so-called *resonant state* and which produces different top-down templates (prototypes) from layer F2 to layer F1. Each *template* is associated with one of the *cluster* nodes in layer F2.

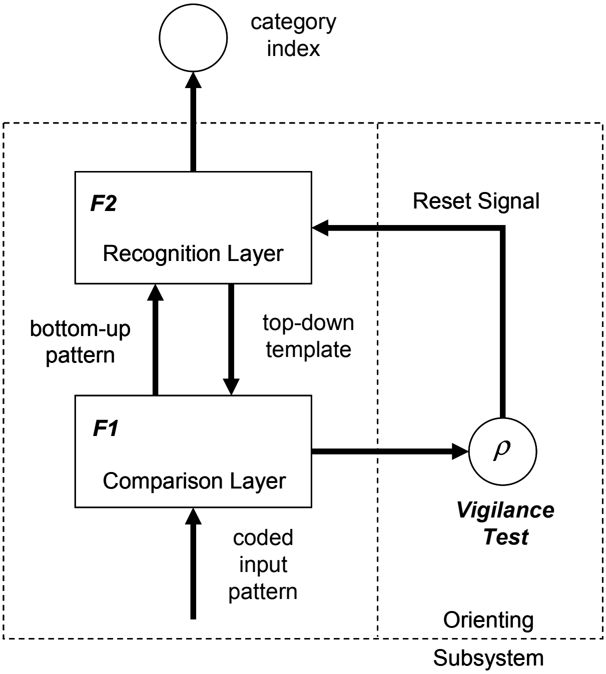


Figure 10.2 Adaptive resonance theory architecture

The orienting subsystem of the ART model is responsible for generating a reset signal to the recognition layer when the bottom-up input pattern and the top-down template mismatch according to a vigilance criterion. This signal, if sent, will cause either a different cluster to be selected or, if no other cluster is available, the end of the resonance state. During training, the formerly coded template associated with the cluster node that represents the best match to the current bottom-up input will be modified to include the input features. If there are no clusters that match to the current bottom-up input, a new one is initialized with the incoming pattern. This is called the *vigilance test*, and is incorporated in the orienting subsystem of the neural network. The ART model allows control of the vigilance test, *i.e.*, the degree of similarity of patterns placed in the same cluster. The similarity depends on the vigilance parameter ρ, where $\rho \in [0,1]$. If ρ is small, the result is inclined to a coarse categorization. On the other hand, if ρ is chosen to be close to 1, many finely divided clusters are formed.

The ART model operates in a plastic mode (*i.e.*, a continuous and cumulative training mode) as long as new patterns are presented to it. This type of neural network was firstly introduced to solve the stability/plasticity problem, *i.e.*, to provide a method by which a neural network can incrementally learn new patterns without forgetting old knowledge. During training, the ART neural network categorizes input patterns of data into clusters with similar features, and when it is confronted by a new input, it produces a response that indicates which cluster the pattern belongs to (if any). Detecting whether an input vector resembles the natural categories formed during training is the function of the *matching algorithm*.

Among different ART models, the fuzzy ART is considered henceforth. It inherits the design features of other ART models and incorporates computations from fuzzy set theory by which it can cluster analog patterns. While a detailed description of fuzzy ART can be found in the papers of Huang *et al.* (1995), Georgiopoulos *et al.* (1996, 1999), and Anagnostopoulos and Georgiopouolos (2002), some analytical details of the matching algorithm along with the step-by-step implementation of training are given in the Appendix.

10.3 Neural Networks for Quality Monitoring

With the movement toward computer-integrated manufacturing, the automation of quality monitoring is considered essential for practical applications. Indeed, profile monitoring based on the use of neural networks is a promising new area of research since a neural network, when compared with other profile monitoring approaches, has the main advantage of being easily implemented by a computer program for the automation of quality monitoring.

Applications of neural networks for quality monitoring presented in the literature have been mainly limited to univariate quality characteristics. The approaches proposed in the literature can be classified into two categories: control chart pattern recognition and unnatural process behavior detection (Zorriassantine and Tannock 1998). These approaches are briefly reviewed in the following two subsections as the use of a neural network for the case of geometric tolerances can be obtained as an extension and adaptation of these methods.

10.3.1 Control Chart Pattern Recognition

Control chart pattern recognition provides a mechanism for identifying different types of predefined patterns in the series of process quality measurements plotted on a univariate control chart. The recognized patterns then serve as the primary information for identifying the causes of unnatural process behavior.

Hwarng and Hubele (1993a, b) carried out extensive studies on pattern recognition by training a MLP in order to detect some basic abnormal patterns on a univariate control chart (*e.g.*, shift, trend, cycle, mixture patterns). Guh and Tannock (1999) investigated the feasibility of MLP to identify concurrent patterns (where more than one pattern exists together, which may be associated with different assignable causes). The authors used a four-layer MLP model with an input layer of 16 neurons (used to input from 16 consecutive sample data points), an output layer of four neurons, and two hidden layers each of 13 neurons. They found that once the number of hidden neurons exceeded 13, the performance was not enhanced and the total training time was increased. They also observed that there is no established theoretical method to determine the optimal configuration of a MLP model, thus most of the design parameters must be determined empirically.

Guh and Hsieh (1999) presented a control system composed of several interconnected MLPs both to recognize the unnatural patterns and to estimate their parameters. Guh (2005) proposed a hybrid-learning-based model integrating MLPs and decision trees as an effective identification system of patterns in the series of process quality measurements. This approach showed high performance in detecting and recognizing some unnatural patterns on the control charts.

Generally, the size of a MLP neural network, *i.e.*, the number of model parameters, increases as the number of input vector elements increases. In the literature, data-window sizes from 16 to 64 observations have been used. Even with 16 data-window observations, the need to train hundreds of weights to classify unnatural patterns is a normal requirement.

In most of the approaches proposed in the literature, coding schemes were applied on the quality measurements after standardization. In the coding process, the measured variable range was divided into *N* zones (where the width of each zone was prespecified), each returning an integer code. The objective of the coding process was to reduce the effect of the noise in the input data before the data were presented to the neural network while retaining the main features in the data. The choice of the coding zone width was critical for MLP classification performance. Gradations that are too small might not be able to detect the important features in the data owing to the effect of random noises. On the other hand, if the gradations are too large, the true process variations might be lost (Cheng 1997; Guh and Tannock 1999).

10.3.2 Unnatural Process Behavior Detection

In the other category, unnatural process behavior detection, Pugh (1991) reported the first application of a MLP model for mean shift detection in a manufacturing process. The author also compared the performance of the implemented neural network with that of Shewhart's control chart. Smith (1994) implemented a four-layer MLP to signal shifts in means or variance in X-bar and *R* control charts. Chang and Ho (1999) used a neural network to discover shifts in variances in two steps. The first part is a neural network which decides whether the pattern is in or out of control. The second part provides a coded value for the shift magnitude. Similarly, Ho and Chang (1999) proposed a MLP model for monitoring process mean and variance shifts simultaneously and classifying the types of shifts. The performance of the MLP model in detecting changes in the process mean was found to be superior to that of a combined Shewhart–cumulative sum (CUSUM) control scheme (Cheng 1995). Cheng and Cheng (2001) proposed a MLP model to monitor exponential mean shifts. Pacella and Semeraro (2007a) proposed a modified MLP model which employs feedback connections between layers to discover mean shifts in the case of serially correlated data.

Al-Ghanim (1997) proposed an unsupervised neural-based system capable of signaling any unnatural change (not just a shift of the mean) in the behavior of a manufacturing process. In particular, the binary implementation of the ART model was trained on a set of natural data in order to cluster them into groups with similar features. After training, the neural network can provide an indication that a change in process outputs has occurred when the series of process data does not fit to any of the learned categories. However, the author found that his pioneering method did not have the same degree of sensitivity as other neural networks (*e.g.*, a MLP). This drawback can be mainly ascribed to the binary coding of the ART algorithm as it is a less flexible way of using process data than a method based on graded continuous number encoding.

Subsequent research extended Al-Ghanim's method and presented outperforming ART-based approaches for unnatural behavior detection (Pacella *et al.* 2004a, b; Pacella and Semeraro 2005). In particular, simplified fuzzy ART algorithms, which do not require binary coding of input data, were presented. In Pacella *et al.* (2004a) the neural network was trained using a series of process natural output data. Pacella *et al.* (2004b) demonstrated that the training set can even be limited to a single vector whose components are equal to the process nominal value. In the posttraining phase, fuzzy ART compares input vectors with learned clusters and produces a signal if the current input does not fit to any of the natural prototypes.

This approach can achieve similar performance in signaling a sustained change of process mean as that of a CUSUM control chart, but at the same time it is also capable of detecting a wide set of potential unnatural changes that cannot be addressed by a sole CUSUM chart. Indeed, for transient or dynamic changes of the process mean, fuzzy ART can outperform charting techniques such as a Shewhart control chart with a set of run rules and sensitizing rules. Since fuzzy ART can model different control strategies simultaneously, it can be exploited as a unique tool for signaling a generic modification in the state of the process.

Furthermore, fuzzy ART responses to an input stimulus can be easily explained (Pacella and Semeraro 2005), in contrast to other neural networks such as the MLP model, where typically it is more difficult to realize why an input produces a specific output. On one hand, this is not a problem for many applications in which the emphasis is on prediction or classification rather than on model building or model understanding. On the other hand, the method of choosing the values of neural network parameters is not well implemented as it is based on an experimental process where different values are used and evaluated. In actual applications, this leaves the user to empirically develop, for the process control case at hand, the relationship between the performance of the neural network and its parameters. This can be very time-consuming.

10.4 A Neural Network Approach for Profile Monitoring

An approach based on a fuzzy ART neural network for profile monitoring was presented by Pacella and Semeraro (2007b) as an adaptation of their previous work related to univariate quality characteristics (Pacella *et al.* 2004a, b) to the case of geometric tolerances (roundness profiles). The emphasis in their work was on signaling process changes from the underlying in-control model. In other words, the fuzzy ART network was not intended to provide a classification of the pattern detected during the operating phase, but to signal any unnatural behaviors in profile data.

The main advantage of this approach is that the in-control model is autonomously derived by the neural network, without requiring any further intervention from the analyst.

The issues of using a fuzzy ART neural network for profile monitoring of geometric tolerances are discussed in later subsections. The case study described in Chapter 11 is considered as a reference. It consists of roundness profiles of $p = 748$ entries. Each profile is scaled by subtracting the least-squares estimate of the radius and centered on the least-squares estimate of the center. Figure 10.3 shows both the polar diagram (Figure 10.3a) and the Cartesian diagram (Figure 10.3b) of the data set used for training the neural network algorithm.

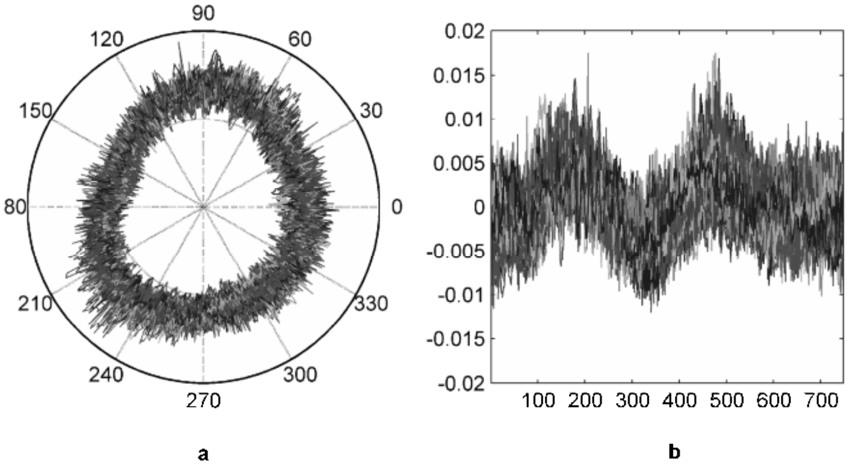


Figure 10.3 Experimental data after the scaling steps and the alignment phase – 100 roundness profiles of 748 data each: **a** polar diagram**,** and **b** the corresponding Cartesian diagram

10.4.1 Input and Preprocessing Stage

For each profile, a vector of p measurements is observed. In the reference case study, each measurement represents the deviation (either positive or negative) of the sampled point from the least-squares estimate of the radius.

Each roundness vector should be considered as the input of a fuzzy ART neural network with p nodes in the input layer. However, since fuzzy ART can only accept values ranging between 0 and 1, a preprocessing stage must be implemented. This stage takes as the input the p-dimensional data vector of deviations, finds the minimal value, and subtracts this value in each component. In this way, a data vector, with a minimum value equal to zero, is eventually obtained.

No additional preprocessing is required at this point and hence the resulting p-dimensional vector can be presented to the neural network. Indeed, a millimeter scale is assumed for the radius deviations: the maximum value of the resulting data vector (the difference between the maximum and minimum radius deviations of the original roundness profile) is always considered to be not greater than 1 mm for common turning processes.

10.4.2 Training

Training of the neural network works as follows. Given a list of profiles observed while the process is in its in-control state, we want the fuzzy ART to cluster these patterns into a suitable number of categories. The vigilance parameter ρ controls the required degree of similarity among input profiles. This parameter has to be chosen at the beginning of the training phase on the basis of the number of clusters, as well as on the fineness of each single cluster, we want to obtain.

In principle, ρ should be chosen in order to maintain the false-alarm rate about equal to a predefined value (Pacella *et al.* 2004a, b; Pacella and Semeraro 2005). This serves to provide an unbiased comparison of the approach with other techniques when the process drifts to unnatural states. To this aim the set of in-control profiles is split into two subsets. We refer to these two subsets as the *training* and *testing* lists, respectively (note that as overfitting does not apply to fuzzy ART, validation data are not required). First, we train the neural network on the training list by using a given vigilance. Then, we check the performance of the trained network, in terms of false alarms produced on profiles of the testing list, using the same vigilance value. Training and testing are then repeated in order to test the effect of different settings of the vigilance parameter.

Generally, the vigilance factor may have a dual effect on the false-alarm rate. On the one hand, a greater value of vigilance may cause the false-alarm rate to increase as long as the growth of the constructed clusters (committed nodes) is small. On the other hand, excessively generated clusters at a high vigilance factor can increase the chances of matching an input pattern with one of the categories, thus causing the false-alarm rate to decrease (Al-Ghanim 1997).

From our experimental study, we observed that when a single natural cluster is formed during training, the false-alarm rate shows an increasing trend as the vigilance factor is increased. This monotonic relationship between the false-alarm rate and the vigilance parameter simplifies the selection of a proper value to be used in order to obtain a predefined false-alarm rate. Therefore, the fuzzy ART neural network is trained on the data set of in-control profiles in order to have no more than one cluster (one committed node).

A seven-step method for implementing the fuzzy ART neural network for monitoring p-dimensional profiles is summarized in the following subsection.

10.4.3 The Method for Implementing the Neural Network

1. Consider the fuzzy ART network with p nodes in the input layer (and hence with $2p$ nodes in field F1). Initialize the iteration index $i = 1$ and the vigilance step $s = 0.0001$. Consider a training list of m natural profiles (in our study $m = 100$).
2. Train the neural network on the given training list using a vigilance parameter of $\rho_i = 1 - i \cdot s$.
3. Repeat step 2, by setting $i = i + 1$, until a single category (committed node) in layer F2 is obtained. Set ρ_u equal to the maximum value inducing one category. From this step on, only values in $\left[0, \rho_u\right]$ are considered for the vigilance. It should be noted that when one category is formed, there exists no competition among alternative committed nodes in layer F2 and the choice parameter of the network has no influence (Georgiopoulos *et al.* 1996, 1999).
4. Reinitialize the iteration index $i = 0$ and use vigilance step $s = 0.00001$. Consider a testing list of M natural profiles (in our study $M = 100$).
5. Train the neural network on the training list of m natural profiles by using a vigilance parameter of $\rho_i = \rho_u - i \cdot s$.
6. Disengage learning and check the performance on the M profiles in the testing list. The goal is to evaluate the number of vectors in the testing list that mismatch to the category formed during training in step 5. Set a equal to this number of false alarms.
7. The estimated false-alarm rate is $\hat{\alpha} = a/M$ (with an approximate standard error of $\sqrt{\hat{\alpha}(1-\hat{\alpha})/M}$). Repeat steps 5 and 6 by setting $i = i + 1$ until $\hat{\alpha}$ is equal to an acceptable rate. Set ρ equal to the vigilance parameter eventually found.

Two observations are in order.

1. A set of *m* in-control training profiles is used for training, and other, different *M* in-control profiles are used for tuning of the vigilance parameter. These two lists of profiles are obtained from the process while it is in its natural or in-control state. No additional model identification or fitting of profile data is required by the analyst.
2. The iteration of steps 5–7 consists in training the network over the *m* profiles and testing the network using the *M* in-control profiles. Any iteration is performed at a specific vigilance level, that is, training and testing are performed using the same value of the vigilance.

10.5 A Verification Study

In this section, a verification study on the efficacy of the proposed neural network approach is described. To this aim, computer simulation is used.

As observed in Chapter 9, roundness profiles can be described by a harmonic model combined with random residual errors. In this chapter, a simplified model is used for simulation, in which a common autoregressive model of order 2 – AR(2) – is used for the residuals. Although AR(2) is not an accurate model for describing the spatial correlation among adjacent measurements on physical profiles, it can be adequate for a simplified modeling (Moroni and Pacella 2008).

Let $y_j(k)$ represent the radial deviation from the nominal radius measured at the angular position $\theta_k = (k-1)\frac{2\pi}{p}$, where $j = 1,2,\ldots$ is the index of the profile. The simulation model is

$$y_j(k) = \sqrt{\frac{2}{p}}\left[c_{1j}\cos(2\theta_k) + c_{2j}\sin(2\theta_k) + c_{3j}\cos(3\theta_k) + c_{4j}\sin(3\theta_k)\right] + \upsilon_j(k),$$

$$\upsilon_j(k) = \frac{1}{1 - c_{5j}B - c_{6j}B^2}\varepsilon_j(k), \tag{10.1}$$

where *B* is the backshift operator $\left[Bx(k) = x(k-1)\right]$ and ε_t is white noise. The vector of six model parameters $\mathbf{c}_j = \left[c_{1j} \quad c_{2j} \quad c_{3j} \quad c_{4j} \quad c_{5j} \quad c_{6j}\right]$ in Equation 10.1 changes from profile to profile according to a six-variate normal distribution of mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ $\left[\mathbf{c}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})\right]$. Values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are summarized Table 10.1 (Moroni and Pacella 2008). An empirical distribution function of residuals (estimated from actual roundness profiles) was used to simulate instances of 748 measurement errors $\varepsilon_j(k)$ in Equation 10.1 (Colosimo *et al.* 2008).

Table 10.1 Values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the simulation model in Equation 10.1

$$\boldsymbol{\mu}' = \begin{bmatrix} -0.0341 & 0.0313 & 0.0080 & -0.0322 & -0.3714 & -0.2723 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.0004 & -0.0002 & 0.0001 & 0.0000 & 0.0001 & 0.0003 \\ -0.0002 & 0.0004 & 0.0001 & 0.0000 & 0.0001 & -0.0002 \\ 0.0001 & 0.0001 & 0.0002 & 0.0000 & 0.0001 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0003 & 0.0003 & 0.0003 \\ 0.0001 & 0.0001 & 0.0001 & 0.0003 & 0.0072 & 0.0012 \\ 0.0003 & -0.0002 & 0.0000 & 0.0003 & 0.0012 & 0.0036 \end{bmatrix}$$

In order to verify the efficacy of the approach, the ability to detect unnatural patterns during phase II is estimated. To this aim, occurrences of assignable causes are simulated by two spindle-motion errors as follows (Cho and Tu 2001).

Bilobe out of control, which is simulated by increasing the amplitude of the second harmonic in the baseline model:

$$y_j(k) + \sqrt{\frac{2}{p}}\delta_2\left[c_{1j}\cos(2\theta_k) + c_{2j}\sin(2\theta_k)\right], \tag{10.2}$$

where c_{1j} and c_{2j} are the coefficients in the baseline model related to the second harmonic, used to simulate the in-control profile of index j. δ_2 is the increasing factor for the amplitude of the second harmonic. Values of $\delta_2 = 0.25, 0.5, 0.75, 1$ are considered to simulate different severities of this out-of-control condition.

Trilobe out of control, which can be simulated by increasing the amplitude of the third harmonic with respect to the baseline model:

$$y_j(k) + \sqrt{\frac{2}{p}}\delta_3\left[c_{3j}\cos(3\theta_k) + c_{4j}\sin(3\theta_k)\right], \tag{10.3}$$

where c_{3j} and c_{4j} are the coefficients of the third harmonic in the baseline model, for the in-control profile of index j. δ_3 is the increasing factor for the amplitude of the third harmonic. Values of $\delta_3 = 0.25, 0.5, 0.75, 1$ are considered.

10.5.1 Implementation of the Fuzzy ART Neural Network

The fuzzy ART neural network, implemented in MATLAB®, consisted of 748 nodes in the input layer (each node corresponds to one specific location of profile

Table 10.2 Experimental results of the fuzzy adaptive resonance theory (ART) neural network: **a** the number of categories formed during training (steps 2 and 3), and **b** the false-alarm rate during tuning (steps 5–7). The value of the vigilance parameter eventually used is highlighted in *bold*

a

i	ρ_i	Categories
1	1.0000	100
2	0.9999	100
…	…	…
21	0.9980	80
22	0.9979	78
…	…	…
41	0.9960	37
…	…	…
61	0.9940	23
…	…	…
81	0.9920	13
…	…	…
101	0.9900	7
…	…	…
111	0.9890	4
…	…	…
121	0.9880	3
…	…	…
140	0.9861	2
141	0.9860	2
142	0.9859	1

b

i	ρ_i	Alarms
0	0.98590	22
1	0.98589	16
2	0.98588	15
3	0.98587	12
4	0.98586	9
5	0.98585	7
6	0.98584	7
7	0.98583	6
8	**0.98582**	**5**
9	0.98581	4
10	0.98580	4
11	0.98579	4
12	0.98578	4
13	0.98577	4
14	0.98576	4
15	0.98575	3
16	0.98574	3
17	0.98573	3
18	0.98572	2
19	0.98571	2
20	0.98570	1

data), 1,496 nodes in field F1 (equal to double the number of input nodes) and a single node in field F2.

The training list consisted of the $m = 100$ roundness profiles of the real case study. By applying steps 2 and 3 of the procedure described in Section 10.4.3, we found the range of vigilance values which allows clustering of training data into one category. From our experimental results, we observed that fuzzy ART classified the training set into one category when the value of the vigilance factor is $\rho \in [0, 0.9859]$.

A testing list of M in-control profiles was then considered. In this work, the testing list was simulated using the in-control model in Equation 10.1. In spite of this, in actual applications the testing list could be obtained directly from the process while it is in its in-control state. Indeed, no additional parametric modeling of profile data is required for implementation of the proposed neural network approach.

We decided to consider $M = 100$ in-control profiles also for the testing list. From our experimental results we observed that such a dimension is adequate to calibrate the vigilance parameter when the reference false-alarm rate is approximately 5% (or greater than 5%). However, the lower the false alarm rate is, the higher the dimension M should be in order to reduce the uncertainty associated with the estimate of the false-alarm rate produced by the neural network.

By applying steps 5–7 of the procedure described in Section 10.4.3, the vigilance value of $\rho = 0.98582$ was eventually identified in order to have a false-alarm rate of approximately 5%.

Table 10.2 details some of the experimental results obtained during the implementation of the fuzzy ART neural network. In particular, Table 10.2a reports the effect of the vigilance parameter on the number of categories formed during training (steps 2 and 3). It can be observed that as the vigilance parameter decreases, the number of categories formed during training decreases too. Table 10.2b reports the effect of the vigilance parameter on the number of false alarms observed on the testing list (steps 5–7). In this case, as the vigilance parameter decreases, the number of false alarms decreases too.

10.5.2 Run Length Performance

In order to verify the efficacy of the neural network approach for profile monitoring, we compare its performance with the performance of the individuals control charts of the out-of-roundness (OOR) values computed by the least-squares and minimum-zone algorithms (see Chapter 8). A separate control chart was implemented for each of these two methods.

The performance index is the average run length (ARL) required by the competing methods to signal out-of-control states in phase II. The run length is defined as the number of samples taken until an out-of-control signal is issued.

For each new profile generated, the least-squares and minimum-zone OOR values were computed. Each of the individuals control charts released an alarm when the corresponding statistic exceeded the control limits. Each new profile generated was also presented as an input to the fuzzy ART neural network, formerly trained on the set of 100 roundness profiles of the reference case study (vigilance parameter 0.98582). The neural network released an alarm when the current input mismatched to the natural category formed during training according to the algorithm described in the Appendix.

Profiles were simulated until each of the three competing approaches (*i.e.*, least squares, minimum zone, and ART) issued an alarm. For each approach, we stored the run length (number of samples until the first alarm). This procedure was performed 1,000 times. Hence, a sample of 1,000 run lengths for each case was used to compute both the ARL value and the corresponding 95% confidence interval (based on the *t*-based statistic with 999 degrees of freedom).

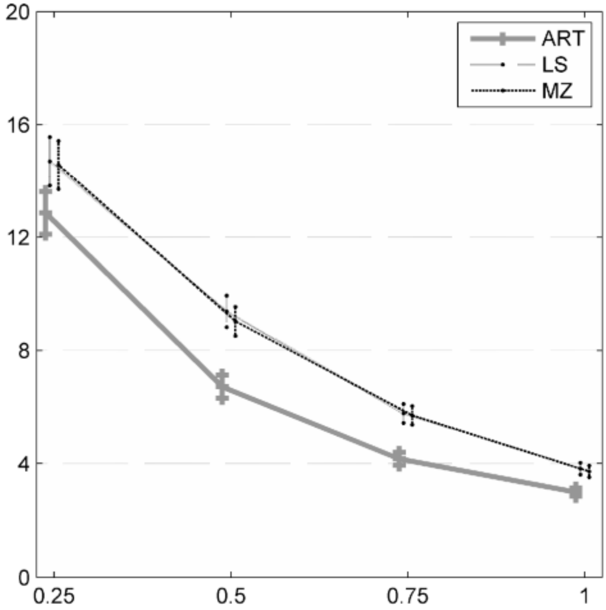


Figure 10.4 Bilobe. Confidence interval (95%, 1,000 replications) on the average run length (*ARL*) (*ordinate*) versus the magnitude of the out-of-control effect (*abscissa*). *LS* least squares, *MZ* minimum zone

Table 10.3 reports the confidence interval for the ARLs obtained by simulating in-control roundness data. It can be observed that the competing methods have the same performance when no out-of-control conditions affect profiles data (with a 5% false-alarm rate the expected in-control ARL is $1/0.05 = 20$). Hence, the performance in phase II is related to the ability to detect out-of-control profiles, given that all the approaches are designed to achieve the same false-alarm probability of about equal 5%.

Figures 10.4 and 10.5 graphically depict the simulation results for the two out-of-control models. In particular, Figure 10.4 refers to bilobe out-of-control errors, while Figure 10.5 refers to trilobe out-of-control errors. (In both figures, the data have been jittered on the *x*-axis to improve the readability). In order to select the best approach in each case study (where each case study is characterized by a spe-

Table 10.3 ARL and confidence interval (95%, 1,000 replications) of competing methods with reference to the baseline model

Method	ARL	95% confidence interval
ART	20.45	[19.17, 21.72]
LS	20.50	[19.25, 21.76]
MZ	20.50	[19.35, 21.87]

LS least squares, *MZ* minimum zone

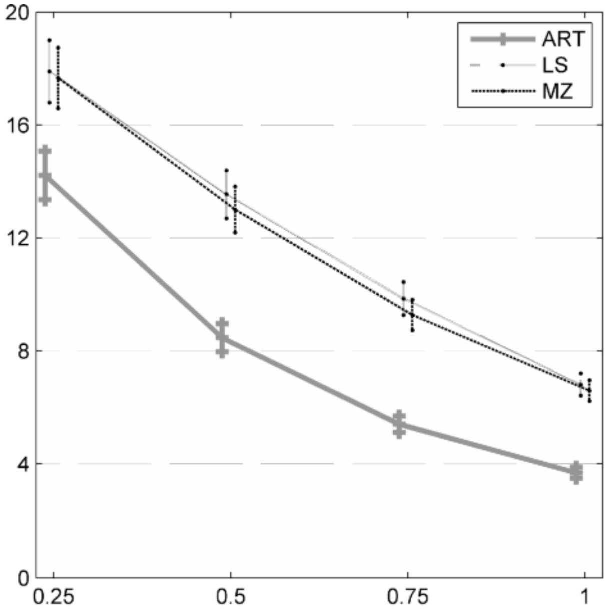


Figure 10.5 Trilobe. Confidence interval (95%, 1,000 replications) on the ARL (*ordinate*) versus the magnitude of the out-of-control effect (*abscissa*)

cific kind of out-of-control and a specific severity of the out-of-control condition), we need to determine the method corresponding to the lowest value of the ARL.

From the results summarized in Figures 10.4 and 10.5, we can conclude that the efficacy of the neural network approach in profile monitoring is proven. Indeed, the performance obtained by using the model-free approach based on a neural network model (ART) is superior to that produced by both the least-squares and the minimum-zone control charts in detecting the two types of motion error that have been simulated, despite the level of the severity of the change under study. Given this successful verification of its efficacy, as well as the inner simplicity in implementing the approach, the use of the proposed fuzzy ART neural network in practice can be justified in actual applications.

10.6 Conclusions

In this chapter, the implementation of a neural network for phase II of profile monitoring was presented. The approach allows a computer to learn from data the relationship to represent the in-control profiles in space. The aim is to develop a new method which can be easily implemented by practitioners in actual applications for automating profile monitoring.

We have shown that fuzzy ART can be effectively used to this aim. The approach is quite simple to automate using a computer through the seven-step procedure presented in Section 10.4.3 and the fuzzy ART training algorithm (in the Appendix). In this chapter, we used 100 roundness profiles to train the neural network, and another 100 roundness profiles to tune the vigilance parameter, which allows a 5% false-alarm rate. The proposed approach is quite general and can be easily extended to deal with different two-dimensional geometric specifications.

With reference to the ability to detect out-of-control states in phase II, a verification study showed that the proposed neural network approach is more effective than the usual control charting methods based on the OOR values in detecting two types of out-of-control conditions (modeled as a systematic form deviation of profile data due to errors of the spindle motion).

When compared with the methods of multivariate charting of fitted parameter vectors (*e.g.*, those discussed in Chapter 9), the neural network approach has a basic advantage, namely, it can be exploited in those applications where an analytical model for the statistical description of profiles considered is not available. Furthermore, a simulation study (not reported here) showed that the neural network has good robustness (*i.e.*, performance is invariant) for moderate contamination of the training data set by out-of-control profiles (about 5%).

As for the ability in signaling out-of-control states in phase II, the neural network may not produce an outperforming result when compared with the model-based approaches. As expected, the model-based techniques may be more effective than the model-free approach described in this chapter in signaling changes in the functional data (for specific production scenarios). Furthermore, an additional disadvantage of the algorithm presented in this chapter is that such an approach is limited to phase II of profile monitoring only, while, in general, the control chart approach can be used for both phase I and phase II.

Despite these drawbacks, profile monitoring based on the use of neural networks is a very promising area of research in the computer-integrated manufacturing field. Indeed, the automation of quality monitoring is becoming more and more important in practical applications, owing to the availability of affordable point-based measurement systems capable of acquiring large amounts of data as point coordinates in reasonable times and with high accuracy. The possibility of managing large data sets, automatically acquired from the process, using automatic online procedures implemented by computers paves the way to the development of innovative approaches in the field of quality monitoring.

In order to select a specific method in a given production scenario, a comparison of the performance of the neural network approach with the performances of the methods presented in the previous chapters is given in Chapter 11.

Appendix

Fuzzy ART operates over all of the committed nodes along with a single uncommitted node. Each committed node (of index j) has a vector $\mathbf{w}_j = \left(w_{j1}\ w_{j2}\ \ldots\ w_{j2p} \right)$ of adaptive weights (which represents the coded template). The number of committed nodes n $\left(j = 1, 2, \ldots, n \right)$ is arbitrary, while the dimension of vector $\mathbf{w}_j$ is $2p$.

Let $\mathbf{x}$ be a p-dimensional input vector $\mathbf{x} = \left(x_1\ \ldots\ x_k\ \ldots\ x_p \right)$, where each component x_t ranges in $\left[0,1\right]$. The matching algorithm of fuzzy ART is as follows.

Step 1: Initialization

During training, initialize the number of committed nodes to $n = 0$. Then, set a choice parameter $\alpha \in \left[0, \infty\right]$ (a small value is used in this work: $\alpha = 10^{-6}$), and a vigilance parameter $\rho \in \left[0,1\right]$.

Step 2: Complement coding

Expand each new input $\mathbf{x}$ into the $2p$-dimensional vector $\mathbf{x}_c$ defined as follows:

$$\mathbf{x}_c = \left(x_1\ \ldots\ x_p\ 1 - x_1\ \ldots 1 - x_p \right). \tag{10.4}$$

Step 3: Category choice

For each committed node of index $j = 1, 2, \ldots, n$, compute the bottom-up input T_j as follows (note that T_j is a scalar):

$$T_j\left(\mathbf{x}\right) = \frac{\left|\mathbf{x}_c \wedge \mathbf{w}_j\right|}{\alpha + \left|\mathbf{w}_j\right|}, \tag{10.5}$$

where operator $\wedge$ gives the vector $\mathbf{u} \wedge \mathbf{v} = \left(\min\left\{u_1, v_1\right\} \cdots \min\left\{u_t, v_t\right\} \cdots \right)$, while operator $\left| \ \right|$ gives the scalar $\left|\mathbf{u}\right| = \sum_t \left|u_t\right|$.

Choose the committed node in layer F2 that receives the maximum bottom-up input. Assume this node has index j^* : $T_{j^*} = \max\left\{T_j : T_j \geq 0, j = 1, \ldots, n\right\}$. If more than one T_{j^*} is maximal, choose the category with the smallest j index.

Two cases can be distinguished:

1. If there are no categories for classifying the current input, release an alarm. During training, select the uncommitted node by setting $n = n+1$, $j^* = n$, and $w_{j^*} = \mathbf{x}_c$. Introduce a new uncommitted node in layer F2. Go to the beginning of step 2.
2. Otherwise, go to step 4.

Step 4: Resonance or reset

Check to see whether node j^* satisfies the following vigilance criterion:

$$\frac{\left|\mathbf{x}_c \wedge w_{j^*}\right|}{\left|\mathbf{x}_c\right|} \geq \rho . \tag{10.6}$$

Two cases can be distinguished:

1. If the vigilance criterion is satisfied, the current input is classified in the category of index j^* (no alarm is released). During training, update the weight vector $w_{j^*}^{(new)} = \mathbf{x}_c \wedge w_{j^*}^{(old)}$ (fast learning). Go to the beginning of step 2.
2. Otherwise, exclude node j^* by setting the choice function $T_{j^*} = -1$ for the duration of the input presentations to prevent its persistent selection during the search. Go to the beginning of step 3.

Acknowledgements This research was partially funded by the Italian Ministry of University and Research grant PRIN 2005 – "Characterization of the multiscale geometry of manufactured surfaces for process control and tolerance analysis".

References

Al-Ghanim A (1997) An unsupervised learning neural algorithm for identifying process behavior on control charts and a comparison with supervised learning approaches. Comput Ind Eng 32:627–639

Anagnostopoulos GC, Georgiopouolos M (2002) Category regions as a new geometrical concepts in fuzzy-ART and fuzzy-ARTMAP. Neural Netw 15:1205–1221

Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, New York

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York

Chang SI, Ho ES (1999) A two-stage neural network approach for process variance change detection and classification. Int J Prod Res 37:1581–1599

Cheng CS (1995) A multi-layer neural network model for detecting changes in the process mean. Comput Ind Eng 28:51–61

Cheng CS (1997) A neural network approach for the analysis of control chart patterns. Int J Prod Res 35:667–697

Cheng CS, Cheng SS (2001) A neural network-based procedure for the monitoring of exponential mean. Comput Ind Eng 40:309–321

Cho NW, Tu JF (2001) Roundness modeling of machined parts for tolerance analysis. Precis Eng 25:35–47

Colosimo BM, Pacella M (2007) On the use of principal component analysis to identify systematic patterns in roundness profiles. Qual Reliab Eng Int 23:707–725

Colosimo BM, Pacella M (2010) Control Charts for Statistical Monitoring of Functional Data, Int J Prod Res 48(6):1575–1601

Colosimo BM, Pacella M, Semeraro Q (2008) Statistical process control for geometric specifications: on the monitoring of roundness profiles. J Qual Technol 40:1–18

Cottrell M, Girard B, Girard Y, Mangeas M, Muller C (1995) Neural modeling for time series: a statistical stepwise method for weight elimination. IEEE Trans Neural Netw. 6:1355–1364

Georgiopoulos M, Fernlund H, Bebis G, Heileman GL (1996) Order of search in fuzzy ART and fuzzy ARTMAP: effect of the choice parameter. Neural Netw 9:1541–1559

Georgiopoulos M, Dagher I, Heileman GL, Bebis G (1999) Properties of learning of a fuzzy ART variant. Neural Netw 12:837–850

Graf S, Luschgy H (2000) Foundations of quantization for probability distributions. Lecture notes in mathematics, vol 1730. Springer, Berlin

Guh RS (2005) A hybrid learning-based model for on-line detection and analysis of control chart patterns. Comput Ind Eng 49:35–62

Guh RS, Hsieh YC (1999) A neural network based model for abnormal pattern recognition of control charts. Compu. Ind Eng. 36:97–108

Guh RS, Tannock JDT (1999) Recognition of control chart concurrent patterns using a neural network approach. Int J Prod Res 37:1743–1765

Gupta S, Montgomery DC, Woodall WH (2006) Performance evaluation of two methods for online monitoring of linear calibration profiles. Int J Prod Res 44:1927–1942

Haykin S (1998) Neural networks. A comprehensive foundation, 2nd edn. Macmillan, New York

Ho ES, Chang SI (1999) An integrated neural network approach for simultaneous monitoring of process mean and variance shifts-a comparative study. Int J Prod Res 37:1881–1901

Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. Neural Netw 2:359–366

Huang J, Georgiopoulos M, Heileman GL (1995) Fuzzy ART proprieties. Neural Netw 8:203–213

Hwarng HB, Hubele NF (1993a) Back-propagation pattern recognizers for X-bar control charts: methodology and performance. Comput Ind Eng 24:219–235

Hwarng HB, Hubele NF (1993b) X-bar control chart pattern identification through efficient off-line neural network training. IIE Trans 25:27–40

Jones MC, Rice JA (1992) Displaying the important features of large collections of similar curves. Am Stat 46:140–145

Kang L, Albin SL (2000) On line monitoring when the process yields a linear profile. J Qual Technol 32:418–426

Kim K, Mahmoud MA, Woodall WH (2003) On the monitoring of linear profiles. J Qual Technol 35:317–328

Mahmoud MA, Woodall WH (2004) Phase I analysis of linear profiles with calibration applications. J Qual Technol 46:380–391

Moroni G, Pacella M (2008) An approach based on process signature modeling for roundness evaluation of manufactured items. J Comput Inf Sci Eng 8:021003

Murata N, Yoshizawa S, Amari S (1994) Network information criterion-determining the number of hidden units for an artificial neural network model. IEEE Trans Neural Netw 5:865–872

Pacella M, Semeraro Q (2005) Understanding ART-based neural algorithms as statistical tools for manufacturing process quality control. Eng Appl Artif Intell 18:645–662

Pacella M, Semeraro Q (2007a) Using recurrent neural networks to detect changes in autocorrelated processes for quality monitoring. Comput Ind Eng 52:502–520

Pacella M, Semeraro Q (2007b) An unsupervised neural network approach for on-line monitoring of machined profiles. Paper presented at the 7th annual conference of ENBIS (European Network for Business and Industrial Statistic), Dortmund, Germany, 24–26 September

Pacella M, Semeraro Q, Anglani A (2004a) Manufacturing quality control by means of a fuzzy ART network trained on natural process data. Eng Appl Artif Intell 17:83–96

Pacella M, Semeraro Q, Anglani A (2004b) Adaptive resonance theory-based neural algorithms for manufacturing process quality control. Int J Prod Res 40:4581–4607

Pugh GA (1991) A comparison of neural networks to SPC charts. Comput Ind Eng 21:253–255

Reed R (1993) Pruning algorithms – a survey. IEEE Trans Neural Netw 4:740–747

Rossi F, Conan-Guez B (2005) Functional multi-layer perceptron: a nonlinear tool for functional data analysis. Neural Netw 18:45–60

Rossi F, Delannay N, Conan-Guez B, Verleysen M (2005) Representation of functional data in neural networks. Neurocomputing 64:183–210

Roy A, Kim LS, Mukhopadhyay S (1993) A polynomial time algorithm for the construction and training of a class of multilayer perceptrons. Neural Netw 6:535–545

Schittenkopf C, Deco G, Brauer W (1997) Two strategies to avoid overfitting in feedforward networks. Neural Netw 10:505–516

Smith AE (1994) X-bar and R control chart interpretation using neural computing. Int J Prod Res 32:309–320

Stinchcombe MB (1999) Neural network approximation of continuous functionals and continuous functions on compactifications. Neural Netw 12:467–477

Vapnik VN (2000) The nature of statistical learning theory, 2nd edn. Springer, New York

Walker E, Wrigth SP (2002) Comparing curves using additive models. J Qual Technol 34:118–129

Wang Z, Di Massimo C, Tham MT, Morris AJ (1994) A procedure for determining the topology of multilayer feedforward neural networks. Neural Netw 7:291–300

Woodall WH, Spitzner DJ, Montgomery DC, Gupta S (2004) Using control charts to monitor process and product quality profiles. J Qual Technol 36:309–320

Zorriassantine F, Tannock JDT (1998) A review of neural networks for statistical process control. J Intell Manuf. 9:209–224

Chapter 11
Quality Monitoring of Geometric Tolerances: a Comparison Study

Bianca Maria Colosimo and Massimo Pacella

Abstract While in the previous chapters different approaches for quality monitoring of geometric tolerances were discussed, in the present one a comparison study is provided in which all of these approaches are considered. The aim is to allow practitioners to select a specific method in a given situation. A manufacturing reference case study is first detailed, namely, profiles measured on machined items subject to geometric specification (roundness). Then, two simulation scenarios are considered for the comparison study, where each scenario is obtained by perturbing the real case study. Competing approaches are ranked in each production scenario considering as a performance index the quickness in detecting out-of-control shapes.

11.1 Introduction

When the quality of a manufactured product is critically related to a shape rather than to a dimension, the main problem is how quality monitoring of functional data can be implemented. Functional data (Ramsay and Silverman 2005) refer to information summarized in the form of profiles where each data point is the observed response at a given location (*e.g.*, a spatial or temporal location).

B.M. Colosimo
Dipartimento di Meccanica, Politecnico di Milano, Via la Masa 1,
20156 Milan, Italy,
e-mail: biancamaria.colosimo@polimi.it

M. Pacella
Dipartimento di Ingegneria dell'Innovazione, Università del Salento,
Via per Monteroni, 73100 Lecce, Italy,
e-mail: massimo.pacella@unisalento.it

An example of functional data discussed in the literature is the set of points measured on a machined profile which is subject to geometric tolerances. In this case, each point collected on the profile is related to a specific spatial location. Colosimo *et al.* (2008) focused on the spatial correlation which often characterizes adjacent points of machined profiles. The authors showed that a suitable model to represent the signature left by a turning process on roundness profiles is a spatial autoregressive regression (SARX). In this analytical model, sinusoidal functions are used as regressors, while the correlation structure of the residuals is properly modeled in space. In order to signal any deviation from the in-control behavior, the parameters of this model are monitored by multivariate control charting, while the estimate of residual variance is monitored by univariate control charting.

Colosimo and Pacella (2007) also explored the advantages of using principal component analysis (PCA) for geometric specification modeling. The choice of using PCA instead of regression can be motivated by its inner suitability in dealing with complex profiles, without requiring the selection of a specific type of model. Similarly to the regression-based approach, the projection coefficients used to describe the systematic way in which profiles vary around the mean profile are used to design a multivariate control chart. The estimate of the residual variance is monitored with a univariate control chart (Colosimo and Pacella 2010).

Despite the specific modeling issues behind these two analytical/parametric approaches described in Chapter 9, which have been presented in the literature to specifically deal with functional data, practitioners may be tempted to skip their adoption in order to use simpler methods. The selection of a proper type of regressor or the use of spatial correlation structures can become cumbersome activities. A data reduction technique such as PCA may sometimes fail to identify some structured patterns in profile data (Colosimo and Pacella 2007). Moreover, PCA requires an extra effort in managing the abstract nature of principal components. These drawbacks may represent an obstacle to the introduction of both kinds of analytical/parametric techniques in actual applications for modeling and monitoring profiles.

A different approach consists in using an algorithm that allows computers to automatically learn from data the signature left by the machining process, namely, an adaptive resonance theory neural network (see Chapter 10). The data set from which the relationship is automatically learned by the algorithm consists of profiles representative of the natural process (in-control profiles). After training, the algorithm produces a signal when an input profile does not fit, according to a specific criterion, to the learned prototype of the manufacturing signature. A neural network does not require an analytical model for the statistical description of the profiles considered. The practitioner just needs to collect profiles representative of the process in its natural state and to set an input parameter of the algorithm, depending on the desired performance in terms of false alarms.

An even simpler approach for monitoring functional data is the location control chart described in Chapter 8. This approach consists in applying a common control chart to data observed at each given location. In practice, the location control chart consists in designing a control region around the ideal or mean shape observed on a set of machined items. An alarm is issued when at least one point in the whole set of data observed in a profile exceeds the control limits. The rationale behind this approach is that if the observed shape is in control, the data observed at that specific location should stay within that interval with a given probability.

Similar to a neural network, the location control chart does not require an analytical model for the statistical description of the profiles considered (model-free approach). Differently from a neural network, the location control chart has the main advantage of visually representing the control region around the ideal or mean shape of profiles. This can help the practitioner to identify the potential causes of alarms, while a neural network algorithm is a black box, which makes it more difficult to realize why an input profile produces or does not produce an alarm.

Among the competing approaches, the individuals control chart of the geometric error characterizing the profiles is also considered. This approach, described in Chapter 8, consists in summarizing all the information contained in the functional data in just one synthetic variable. This variable measures the geometric form error as the (maximum) distance between the actual profile and the ideal geometry. Then the usual control charting methods can be applied to the estimated geometric errors for quality monitoring. With reference to roundness, the geometric error is summarized in a value called out of roundness (OOR). In this chapter, the least-squares algorithm is used to estimate the OOR value for a roundness profile.

Both the location control chart and the individuals control chart of the geometric error represent approaches at a lower level of complexity, because they only require the use of common control charting for profile monitoring. Thanks to their inner simplicity, these approaches may be widely preferred by practitioners and, in fact, they can be considered representative of industrial practice. On the other hand, at a higher complexity level, approaches based on combining an analytical model of functional data with multivariate and univariate control charting represent innovative methods which have been designed with the specific objective to deal with functional data. At an intermediate level of complexity, a neural network, which represents a general-purpose algorithm for a computer to automatically learn a relationship from data, is also considered.

In order to allow practitioners to select a specific approach for monitoring functional data in a given situation, this chapter provides a numerical comparison, based on simulation, of (1) the location control chart, (2) the individuals control chart of the geometric error characterizing the profiles, (3) the regression-based approach, (4) the PCA-based approach, and (5) the adaptive resonance theory neural network.

The comparison study is based on two simulation scenarios which are obtained by perturbing a case study of machined profiles subject to geometric specification (roundness profiles obtained by turning, where each curve is characterized by 748 evenly distributed measurements). Each measurement on a profile is the radial distance at a specific angular position. The perturbed scenarios are designed to represent two different although realistic productive situations.

The performances of the competing approaches are measured as the ability to detect unusual patterns in the functional data during the operating phase of the control chart (known as phase II). Basically, the objective in this chapter is to investigate situations where each approach should be preferred to the others, thus to provide some guidelines for implementing profile monitoring in actual applications.

The remainder of this chapter is organized as follows. In Section 11.2, the reference case study of roundness profiles obtained by turning is discussed in detail. In Section 11.3, the simulation models for each production scenario obtained from the reference case study are presented. In Section 11.4, the out-of-control models considered in the study are discussed. In Section 11.5, the approaches for profile monitoring are compared with reference to phase II of profile monitoring. Section 11.6 presents the conclusions and some final remarks.

11.2 The Reference Case Study

A turning process was used to machine 100 items starting from C20 carbon steel cylinders, which were supplied in rolled bars (original diameter 30 mm) and were machined to a nominal diameter of 26 mm (using a cutting speed of 163 m/min, a feed of 0.2 mm/rev, and two cutting steps of 1-mm depth each). Figure 11.1 shows a cylindrical item before machining and one after machining by lathe turning.

Each cylindrical surface obtained by lathe turning was measured by using a coordinate measuring machine (CMM). Sampling was performed by continuous scanning of 748 generatrices on each turned specimen and extracting from each cylindrical surface ten roundness profiles (at different distances from the spindle).

A statistical analysis of the data obtained showed that the roundness profile changes as the distance from the spindle changes. The technological reason behind this behavior is due to a different inflection of the workpiece while moving far from the spindle. The case study considered in this chapter refers to the profile that is the most distant from the spindle (represented in Figure 11.2). This profile is extracted from each of the 100 items machined.

Figure 11.3 shows an item measured by continuous scanning of its surface on a CMM. Each roundness profile sampled on an item consisted of 748 equally spaced measurements of the radius. The original measurements sampled using the CMM were scaled by subtracting the least-squares estimate of the radius and centered at the least-squares estimate of the center.

A roundness profile can be described by a polar representation where a set of points, evenly distributed on the machined surface, represent the deviation of the actual measurement from the nominal radius at different angle locations (Cho and Tu 2001). Note that such deviations can be either positive or negative. A polar representation of the whole experimental data set is shown in Figure 11.4. The corresponding Cartesian diagram, in which the independent variable is the location index on the part and the dependent one is the deviation from the nominal radius, is depicted in Figure 11.5.

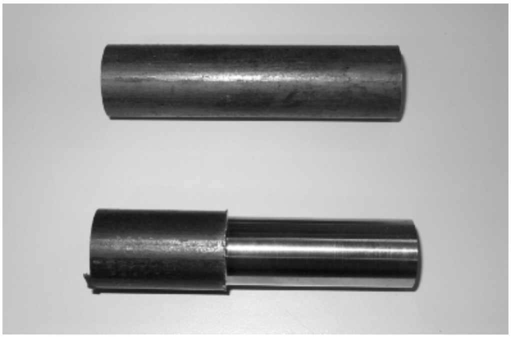


Figure 11.1 Two cylindrical items of the reference case study: a rough piece and a final piece after machining by lathe turning

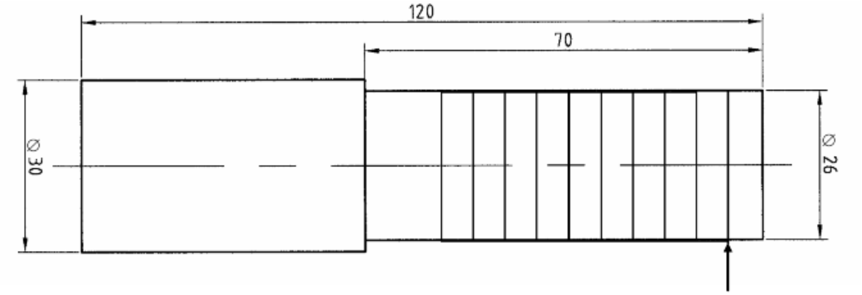


Figure 11.2 The component machined by lathe turning (the *arrow* shows the roundness profile under study)

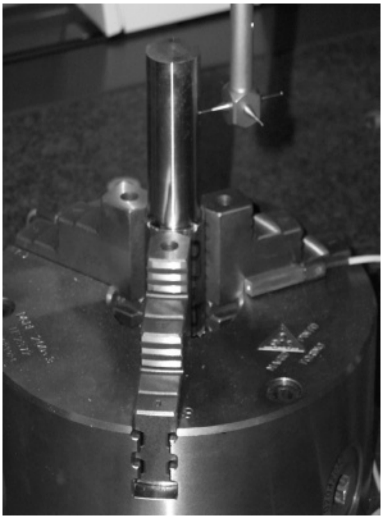


Figure 11.3 An item of the reference case study measured by using a coordinate measuring machine where sampling was performed by continuous scanning of 748 generatrices

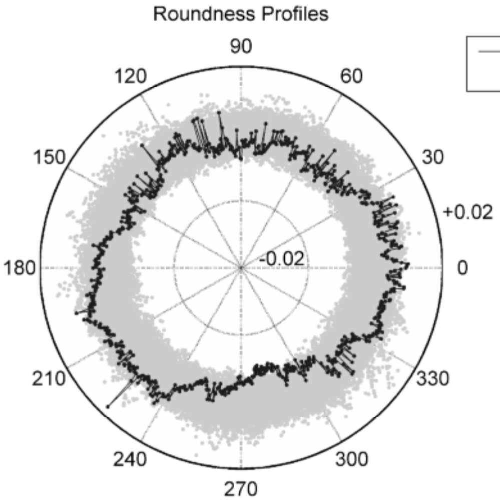


Figure 11.4 Polar diagram of experimental data of real roundness profiles. One hundred roundness profiles of 748 points each. One of the 100 profiles is depicted as a *bold line*

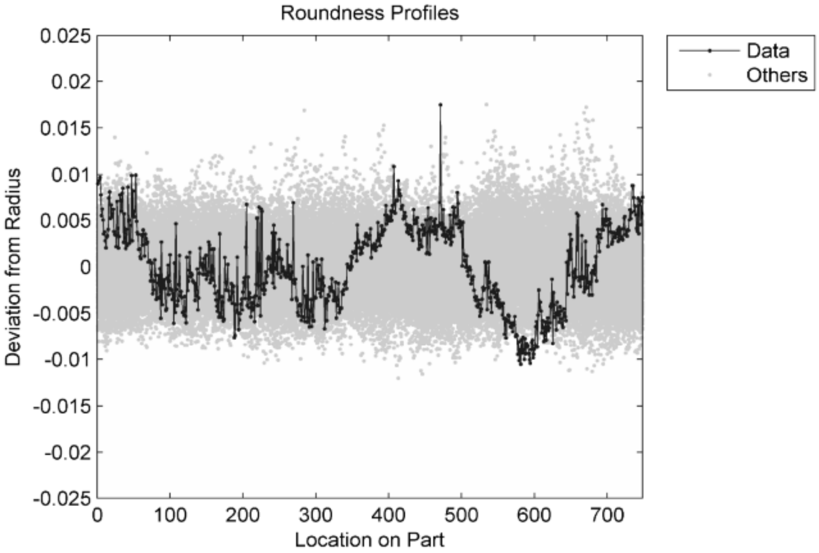


Figure 11.5 Cartesian diagram of experimental data of real roundness profiles. One hundred roundness profiles of 748 points each. One of the 100 profiles is depicted as a *bold line*

From Figures 11.4 and 11.5 it seems that no systematic pattern characterizes the roundness profiles of the reference case study. This appearance hides a common problem of shape analysis, which consists in feature registration or alignment. In fact, the profiles shown in Figures 11.4 and 11.5 are actually misaligned because of the random contact angle of the turning process. Therefore, a further step is

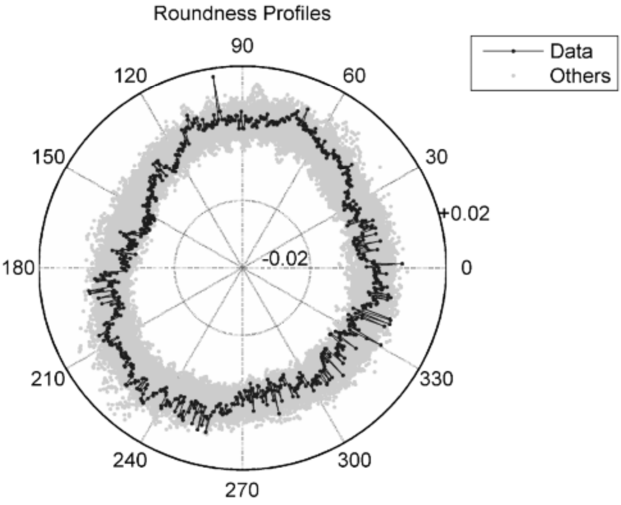


Figure 11.6 Aligned experimental data of real roundness profiles: polar diagram. One of the 100 profiles is depicted as a *bold line*

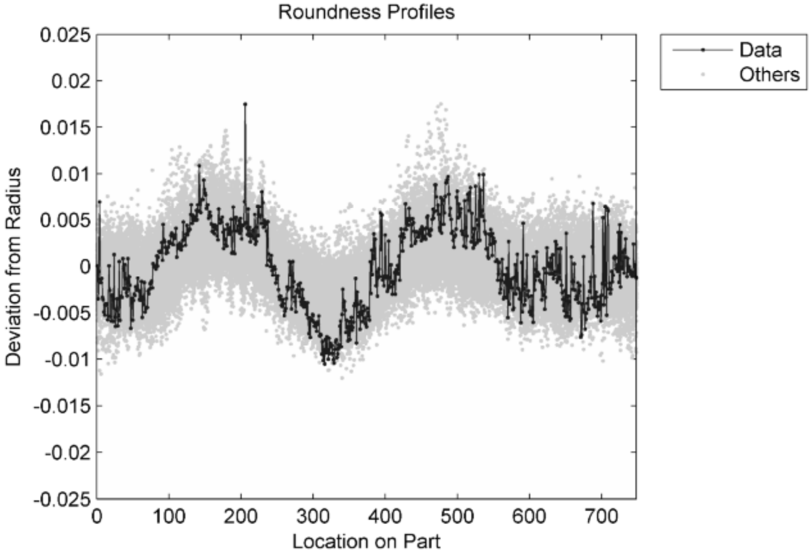


Figure 11.7 Aligned experimental data of real roundness profiles: Cartesian diagram. One of the 100 profiles is depicted as a *bold line*

needed in order to register all the sampled profiles by minimizing the phase delay caused by the random contact angle (Colosimo and Pacella 2007).

Figures 11.6 and 11.7 show, respectively, the polar and the Cartesian diagrams of the aligned roundness profiles after a proper registration procedure has been implemented on the data. From a visual inspection, it can be easily observed that the roundness profiles share a common shape (pattern), *i.e.*, the turning process leaves a specific signature on the machined components (Colosimo *et al.* 2008).

The following subsection gives the technical details on the registration procedure implemented on the roundness profile of the reference case study. The reader may skip these details without loss of continuity.

11.2.1 The Registration of Profiles

Variation among profile data can be considered as being composed of two components: location (or phase) variability and amplitude variability. A usual preliminary step in profile data analysis is the registration (or alignment) of profiles, which is performed with the aim of removing, or minimizing, phase variation so that the analyses can focus on amplitude effects only.

Let us start with the simple case in which two profiles have to be registered. Without loss of generality, the first profile can be used as a reference and the second one can be registered with respect to this reference profile. When profiles refer to closed curves, as happens for the roundness profile, registration can be performed via circular shifts. A circular shift consists in moving one position ahead all the data observed on the profile (where "circular" refers to the fact that the last observation becomes the first one after the shift has been performed). Since each profile is observed on a fixed set of equally spaced p locations, the number of possible circular shifts is equal to the number of points measured on the profile minus one $(p-1)$. Registration consists in determining the best number of circular shifts that have to be performed on the current profile in order to minimize the phase delay between the current and the reference profiles.

As an example, Figure 11.8a shows the reference profile and the current profile. Both roundness profiles have a similar shape (bilobe), while they differ in both amplitude and phase. The correlation between the two profiles is –0.715.

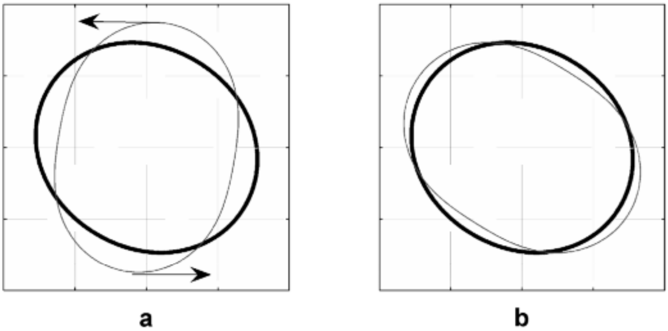


Figure 11.8 Polar diagrams of two ideal bilobe profiles (reference profile *bold line*): **a** misaligned profiles, and **b** aligned profiles

By performing all the possible circular shifts of the current profile, we find that the best number of circular shifts is the one corresponding to the highest correlation between the current profile and the reference profile. In this simple example, since there is no additional noise, the highest correlation 1; the corresponding alignment between the reference profile and the current profile is shown in Figure 11.8b.

As shown in this example, correlation between profiles can act as a performance criterion in the registration procedure. Note that covariance (correlation) refers to data observed for different profiles (cross-correlation).

Given two profiles, the correlation between the ith and the jth profile is given by $r_{ij} = \frac{s_{ij}}{s_i s_j}$, where $s_{ij} = \frac{1}{p-1}\sum_{k=1}^{p}[y_i(k) - \bar{y}_i][y_j(k) - \bar{y}_j]$ is the covariance between the i th and the j th profile, $s_j^2 = \frac{1}{p-1}\sum_{k=1}^{p}\left[y_j(k) - \bar{y}_j\right]^2$ is the variance associate with the ith profile, and $\bar{y}_j = \frac{1}{p}\sum_{k=1}^{p} y_j(k)$ is the mean value of each profile (which is equal to 0 if the least-squares estimator of the radius is subtracted from each profile as the first step). Therefore, in order to align the ith profile with respect to the jth one, all the possible circular shifts of the ith profile are performed and the corresponding correlation indexes r_{ij} are computed. The best circular shift is the one corresponding to the maximum value of the correlation index.

This procedure allows one to register a set of profiles with respect to a given one. In fact, if the jth profile is used as a reference, all the remaining $n - 1$ profiles can be registered with respect to this reference profile. The result of this registration procedure is summarized by an $n \times n$ symmetric correlation matrix $\mathbf{R}_{(j)}$, where the subindex j is used to denote the profile which was used as a reference in this registration step.

This procedure can be repeated by considering in turn each of the n profiles as the reference one. Eventually, n possible registration configurations are available, each characterized by a specific $\mathbf{R}_{(j)}$ ($j = 1,...,n$). In order to select the best registration among this set of n possible alternatives, denote with $\lambda_{1(j)}, \lambda_{2(j)}, \ldots, \lambda_{n(j)}$ the eigenvalues of $\mathbf{R}_{(j)}$. Since the correlation matrix is symmetric, these eigenvalues are real, nonnegative, and their sum is equal to the trace of the correlation matrix (*i.e.*, the sum of the diagonal element $\sum_{i=1}^{n} \lambda_{i(j)} = n$). Consider two extreme cases. When there is no cross-correlation among the profiles (worst case), the correlation matrix $\mathbf{R}_{(j)}$ is an identity matrix and $\lambda_{1(j)} = \lambda_{2(j)} = \ldots = \lambda_{n(j)} = 1$. On the other hand, when ideally there is the maximum cross-correlation among the profiles (best case), the correlation matrix is an "all-one" matrix with rank equal to 1. In this case, all the eigenvalues but the first one are equal to zero, since the rank of a symmetric matrix is equal to the number of eigenvalues that are greater than zero. Furthermore, since the sum of the eigenvalue is always equal to the trace of the matrix, the first eigenvalue is equal to the trace, *i.e.*, $\lambda_{1(j)} = n$ and $\lambda_{2(j)} = \ldots = \lambda_{n(j)} = 0$. Consider now the summation of the squared eigenvalues $\sum_{i=1}^{n} (\lambda_i)^2$. In the worst case it is equal to *n*, while in the best case it is equal to n^2. In any other case, since the eigenvalues are nonnegative, the sum of squared eigenvalues ranges between n and n^2, *i.e.*, $n \le \sum_{i=1}^{n} \left(\lambda_{i(j)}\right)^2 \le n^2$. Therefore, the best configuration is the one which maximizes the sum of the squared eigenvalues.

The registration procedure is eventually summarized in the following steps:

1. Set $j = 1$ as a reference sample. For each of the remaining $i = 1, \ldots, n$ profiles, where $i \ne j$, perform all the possible $p - 1$ circular shifts and select the one that maximizes r_{ij}. Denote by $\mathbf{R}_{(j)}$ the correlation matrix obtained after aligning all the samples with the reference one.
2. Set $j = j + 1$ and repeat step 1 until all the profiles have been considered as a reference (*i.e.*, until $j \le n$).
3. Among the n configurations obtained, choose the optimal one, say, j^*, which satisfies the following criterion:

$$j^* = \arg\max_{j=1\ldots n} \sum_{i=1}^{n} \left(\lambda_{i(j)}\right)^2, \tag{11.1}$$

where $\lambda_{1(j)}, \lambda_{2(j)}, \ldots, \lambda_{n(j)}$ are the eigenvalues of the correlation matrix $\mathbf{R}_{(j)}$ characterizing the *j* th configuration.

11.3 Production Scenarios

Chapter 9 details a SARX model of the data in the reference case study previously described. In this model, the dependent variable $y_j(k)$ represents the radial deviation from the nominal radius measured at the angular positions $\theta_k = (k-1)\frac{2\pi}{p}$, where $j = 1, 2, \ldots$ is the index of the profile and $k = 1, 2, \ldots p$ is the index of equally spaced observations on each profile.

The SARX model is composed of two parts, namely, the large-scale component and the small-scale component. The large-scale component is modeled as a combination of two sinusoidal functions of the angular position θ_k [called harmonics, *i.e.*, $\cos(h\theta_k)$ and $\sin(h\theta_k)$, where $h = 2,3$]. These two sinusoidal functions are exploited in order to model, respectively, the ovality and the triangularity of roundness profiles. The small-scale component of the SARX model describes the correlation structure as a generic spatial autoregressive of order 2.

Since four parameters are used for the two harmonics and two more parameters are needed to describe the correlation structure, each profile of index j is associated with a vector of $d = 4+2$ parameters $\mathbf{c}_j = \left[c_{1j} \; c_{2j} \; c_{3j} \; c_{4j} \; c_{5j} \; c_{6j} \right]$. The first four coefficients $c_{1j}, c_{2j}, c_{3j}, c_{4j}$ are used to represent the parameters of the harmonics (c_{1j}, c_{2j} refer to the second harmonic, while c_{3j}, c_{4j} refer to the third one). Coefficients c_{5j} and c_{6j} refer to the two parameters of the spatial autoregressive model.

By combining instances of the d-length parameter vector $\mathbf{c}_j$ with instances of the p-length vector $\boldsymbol{\varepsilon}_j$ (of independently and normally distributed errors with zero mean and common variance), one can use the SARX model to simulate on a computer realistic roundness profiles (Colosimo and Pacella 2010). In this chapter, the focus is on two different production scenarios.

On the one hand, the first scenario mimics the case study. In this production scenario, the d-length parameter vector $\mathbf{c}_j$ changes from profile to profile according to a d-variate normal distribution $\mathbf{c}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where N represents a multivariate normal (Gaussian) distribution of mean $\boldsymbol{\mu}$ (d-length vector) and covariance matrix $\boldsymbol{\Sigma}$ ($d \times d$ symmetric matrix). The actual values of the mean vector $\boldsymbol{\mu}$ and of the covariance matrix $\boldsymbol{\Sigma}$ exploited for the simulation of this scenario are summarized in Table 11.1.

On the other hand, the second scenario is obtained by perturbing the SARX model of the case study with reference to the variability that characterizes the d-length parameter vector $\mathbf{c}_j$. In particular, a null matrix is considered as a covariance matrix ($\boldsymbol{\Sigma} = \mathbf{0}$) for the d-variate normal distribution of the model parameters. Hence, a fixed vector of parameters is exploited for each instance ($\mathbf{c}_j = \boldsymbol{\mu}$).

Table 11.1 Parameters characterizing the mean $\boldsymbol{\mu}$ and the variance $\boldsymbol{\Sigma}$ of the distribution of coefficients $\mathbf{c}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for the actual roundness data (Colosimo *et al.* 2008)

$$\boldsymbol{\mu}' = \left[-0.0341 \quad 0.0313 \quad 0.0080 \quad -0.0322 \quad 0.3021 \quad 0.2819 \right]$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 4.0646 & -2.0200 & 0.6540 & 0.2652 & -0.8844 & -2.4101 \\ -2.0200 & 3.8961 & 1.4851 & 0.0614 & -1.2123 & 1.9568 \\ 0.6540 & 1.4851 & 2.2346 & -0.1074 & -1.1844 & 0.5958 \\ 0.2652 & 0.0614 & -0.1074 & 3.1214 & -1.4993 & -3.7224 \\ -0.8844 & -1.2123 & -1.1844 & -1.4993 & 38.0199 & 15.8999 \\ -2.4101 & 1.9568 & 0.5958 & -3.7224 & 15.8999 & 43.2491 \end{bmatrix} \times 10^{-4}$$

The first production scenario, where model parameters $\mathbf{c}_j$ can change from profile to profile, is referred to as *random-effect* model, and the second production scenario, in which parameters $\mathbf{c}_j$ do not change, is referred to as *fixed-effect* model. Indeed, a fixed-effect model is usually assumed in traditional approaches for profile monitoring (Woodall *et al.* 2004). In manufacturing, assuming that the input material is very stable and homogeneous, profile-to-profile variability of the small-scale component (spatial structure) can probably be neglected. Furthermore, assuming a process is more stable and/or more controlled, also the profile-to-profile variability in the large-scale component (harmonic structure) tends to vanish. In other words, while the first scenario with random effects corresponds to a common machining process, in which natural causes of variability affect the variability in both the parameters and the residuals of the model, the second scenario, with fixed effects, can be considered representative of a process that is stabler and/or more controlled in which natural causes can affect only the residuals of the model.

11.4 Out-of-Control Models

The performance of the competing approaches is measured as the ability to detect unnatural patterns in the functional data during the operating phase of process monitoring (also known as phase II).

In order to evaluate the performance of the competing approaches in phase II of process monitoring, the occurrences of assignable causes are simulated by a total of three out-of-control conditions. These out-of-control conditions are simulated by spindle-motion errors (Cho and Tu 2001), which are modeled by introducing a spurious harmonic in the baseline model of the roundness profile data. Each condition is then characterized by a parameter directly proportional to the severity of the out-of-control condition introduced in the baseline model. In particular, denoting by $y_j(k)$ the measurement of index $k = 1, 2, \ldots p$ on the profile of index $j = 1, 2, \ldots$, the out-of-control conditions are simulated according to the following models.

Half-frequency spindle-motion error, which can be due to wear on one ball bearing spindle or to whirling in a hydrodynamic bearing. This out-of-control condition can be modeled as follows:

$$y_j(k) + \sqrt{\frac{2}{p}}\delta_1 \sin\left(\frac{1}{2}\theta_k\right), \tag{11.2}$$

where the second term represents the out-of-control state arising in phase II. δ_1 is the size of the shift. Values of $\delta_1 = 0.1, 0.15, 0.2, 0.25$ are considered to model different severities of this out-of-control condition.

Bilobe out of control, which can be caused by an improper setup of the workpiece or by an increased bilobe motion error already affecting the spindle lathe. This out-of-control condition can be simulated by increasing the amplitude of the second harmonic in the baseline model:

$$y_j(k)+\sqrt{\frac{2}{p}}\delta_2\left[c_{1j}\cos(2\theta_k)+c_{2j}\sin(2\theta_k)\right], \tag{11.3}$$

where c_{1j} and c_{2j} are the coefficients in the baseline model related to the second harmonic used to simulate the in-control profile of index j. δ_2 is the increasing factor for the amplitude of the second harmonic. Values of $\delta_2 = 0.1, 0.2, 0.3, 0.4$ are considered to simulate different severities of this out-of-control condition.

Trilobe out of control, which can be due to an increase in the trilobe motion error already existing in the baseline model of the spindle or to an excessive force imposed by the clamping fixture. Similarly to the previous case, this third unnatural condition can be simulated by incrementing the amplitude of the third harmonic with respect to the baseline model:

$$y_j(k)+\sqrt{\frac{2}{p}}\delta_3\left[c_{3j}\cos(3\theta_k)+c_{4j}\sin(3\theta_k)\right], \tag{11.4}$$

where c_{3j} and c_{4j} are the coefficients related to the third harmonic in the baseline model, used to simulate the in-control profile of index j. δ_3 is the increasing factor for the amplitude of the third harmonic. Values of $\delta_3 = 0.1, 0.2, 0.3, 0.4$ are also considered in this case.

Note that the spurious harmonic can influence just one frequency (as for the bilobe and trilobe, which influence the amplitude of the second and third harmonics, respectively), or more frequencies at once (as in the case of the half frequency, which also produces shifts in the amplitude of harmonics greater than the third).

Data obtained under these phase II models are also scaled (by subtracting the least-squares estimation of the radius) and centered (on the least-squares estimation of the center) before applying the profile monitoring method. In fact, we are assuming that centering and scaling are standard steps applied to data when the focus is on the out-of-roundness value only (Cho and Tu 2001).

11.5 Performance Comparison in Phase II of Process Monitoring

The objective in phase II is to quickly detect any change in the process from its in-control state. Hence, the monitoring approaches are compared in terms of the average run length (ARL), where the run length is defined as the number of samples taken until an out-of-control signal is issued.

Performance comparison is based on the ideal assumption that the in-control parameters for each competing method are known. Indeed, computer simulation is used in our work to obtain a large dataset of in-control profiles in order to estimate as closely as possible the parameters of each method. For each of the two production scenarios under study, all simulations were conducted by first tuning each competing approach in order to achieve the same in-control ARL value of about 100. Hence, the performances in phase II are related to the ability to detect out-of-

Table 11.2 Phase II simulation results for the production scenario with the random-effect model. Actual average run lengths (ARLs) and corresponding standard deviations in *parentheses* (1,000 trials)

		LOC CC		OOR CC		REG CC		PCA CC		ART NN	
Half frequency	0.1	73.26	(2.18)	98.15	(3.12)	93.18	(2.83)	80.04	(2.56)	66.50	(2.06)
	0.15	50.53	(1.57)	97.89	(3.06)	78.68	(2.63)	61.07	(1.90)	48.64	(1.59)
	0.2	35.61	(1.16)	85.00	(2.71)	68.49	(2.16)	44.43	(1.37)	35.05	(1.05)
	0.25	22.49	(0.72)	63.51	(2.01)	48.50	(1.56)	29.05	(0.89)	25.90	(0.82)
Bilobe	0.1	64.08	(1.92)	72.39	(2.32)	76.41	(2.42)	64.29	(1.89)	54.57	(1.82)
	0.2	36.07	(1.16)	44.66	(1.42)	47.48	(1.60)	38.03	(1.22)	29.50	(0.92)
	0.3	24.01	(0.77)	25.69	(0.81)	29.76	(0.92)	21.21	(0.67)	18.65	(0.56)
	0.4	14.98	(0.45)	18.40	(0.57)	16.93	(0.53)	12.61	(0.38)	11.66	(0.35)
Trilobe	0.1	70.44	(2.16)	88.29	(2.80)	72.51	(2.27)	70.08	(2.19)	58.85	(1.89)
	0.2	43.66	(1.44)	67.87	(2.18)	47.72	(1.50)	37.55	(1.17)	37.16	(1.17)
	0.3	30.33	(0.94)	46.82	(1.49)	27.57	(0.85)	21.70	(0.67)	23.73	(0.77)
	0.4	19.43	(0.60)	33.81	(1.04)	17.39	(0.54)	12.60	(0.40)	15.41	(0.46)

LOC CC location control chart, *OOR CC* out of roundness control chart, *REG CC* regression-based approach control chart, *PCA CC* principal-component-analysis-based control chart, *ART NN* adaptive resonance theory neural network

Table 11.3 Phase II simulation results for the production scenario with the fixed-effect model. Actual ARLs and corresponding standard deviations in *parentheses* (1,000 trials)

		LOC CC		OOR CC		REG CC		PCA CC		ART NN	
Half-frequency	0.1	33.31	(1.00)	42.47	(1.36)	1.27	(0.02)	6.00	(0.17)	40.48	(1.28)
	0.15	12.26	(0.38)	19.71	(0.59)	1.00	(0.00)	1.51	(0.03)	19.53	(0.64)
	0.2	4.81	(0.13)	11.21	(0.33)	1.00	(0.00)	1.03	(0.01)	10.55	(0.34)
	0.25	2.37	(0.06)	5.83	(0.17)	1.00	(0.00)	1.00	(0.00)	5.34	(0.15)
Bilobe	0.1	72.11	(2.13)	58.40	(1.86)	6.75	(0.19)	45.89	(1.40)	49.60	(1.56)
	0.2	37.70	(1.21)	21.72	(0.69)	1.20	(0.02)	7.02	(0.20)	19.87	(0.63)
	0.3	21.47	(0.63)	8.87	(0.27)	1.00	(0.00)	1.71	(0.04)	8.30	(0.23)
	0.4	10.63	(0.32)	4.22	(0.12)	1.00	(0.00)	1.05	(0.01)	3.81	(0.10)
Trilobe	0.1	86.85	(2.72)	82.43	(2.71)	16.70	(0.50)	64.30	(1.94)	59.63	(1.84)
	0.2	60.26	(1.92)	38.99	(1.24)	2.44	(0.06)	20.18	(0.57)	31.84	(1.01)
	0.3	33.59	(1.04)	20.37	(0.67)	1.11	(0.01)	5.20	(0.15)	16.86	(0.51)
	0.4	21.56	(0.66)	10.42	(0.32)	1.01	(0.00)	1.87	(0.04)	9.82	(0.30)

control profiles, given that all the approaches are designed to achieve the same false-alarm probability of approximately 1%. To this aim, we also evaluated the performance of the competing approaches in the case in which no out-of-control condition was present in phase II, just to check the correctness of tuning for each control chart in each simulated scenario.

Tables 11.2 and 11.3 summarize the simulation results for the two production scenarios under study. In particular, Table 11.2 refers to the case of the random-effect model, while Table 11.3 refers to the case of the fixed-effect model. Each table reports the ARLs estimated by computing a set of 1,000 run lengths, given new profiles simulated according to a specific out-of-control model.

Since in actual industrial applications the analyst is not expected to know *a priori* what kind of out-of-control condition will affect the production process and how severe it will be, we consider a measure of the overall performance for each of the five competing approaches in each production scenario. To do this, we consider the mean ARL values for each competing approach in signaling a generic out-of-control condition of any severity for that production scenario. We assume that all out-of-control conditions previously considered are equally probable and that the analyst knows the model for the monitored functional data (this is plausible when a retrospective phase of control charting has been accomplished).

Figures 11.1 and 11.2 graphically depict the 95% Bonferroni confidence intervals of the overall ARLs presented by the three competing approaches in each production scenario considered in our study. A discussion on the performance comparisons is given in the following two subsections for the production scenarios with the random-effect model and the fixed-effect model, respectively.

11.5.1 Production Scenario with the Random-Effect Model

From the results reported in Table 11.2, graphically summarized in Figure 11.9, it can be observed that the control chart of the geometric error has the lowest power of detection when compared with the competing methods.

Similarly, the regression-based approach has a small power of detection in signaling out-of-control conditions influencing one or more harmonics. This may be mainly ascribed to the variability in the regression parameters which characterizes this production scenario with random effects for the baseline model. It should be noted that for the out-of-control conditions considered in our study (half-frequency, bilobe, and trilobe), the majority of alarms released by the regression-based approach are produced by the multivariate control chart of the vector of fitted parameters. The extra variability in the regression parameters, which naturally characterizes the vector of fitted parameters in this scenario, causes a lower detection power of the regression-based control charts, in particular of the multivariate control chart.

It can be noted that the PCA-based approach, in many cases, outperforms the regression-based approach. With reference to scenarios of random effects, the

PCA-based approach consists of a multivariate control chart based on the first four retained principal components and of a univariate control chart for monitoring the residuals.

Furthermore, from Table 11.2 it can also be noted that the location control chart has performance comparable to that observed for the PCA-based approach. In a few cases, especially in the case of the half-frequency out-of-control conditions, surprisingly the location control chart outperforms the PCA-based approach. This result shows that the simple location control chart can be considered a valuable alternative to parametric methods for profile monitoring, at least in a production scenario with random effects.

Nevertheless, in the case of random effects, the comparison study shows that the neural network approach should always be preferred to signal almost all the out-of-control conditions, with the only exception of half-frequency where the performance is equal to that of the location control chart. This result shows that the neural network is able to model the manufacturing signature and to signal correctly an out-of-control condition even if extra variability naturally characterizes the parameters of the model.

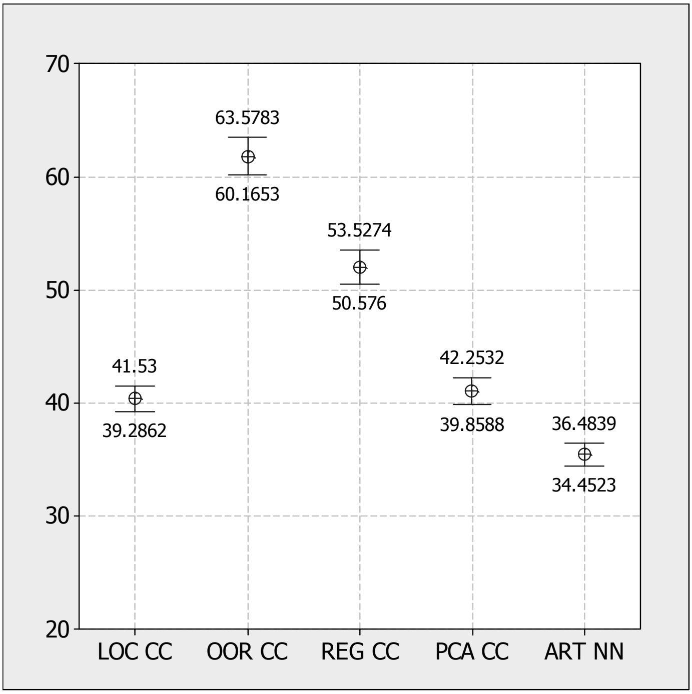


Figure 11.9 The 95% Bonferroni confidence intervals of the overall average run length (ARL) for the competing approaches for the production scenario with the random-effect model. *LOC CC* location control chart, *OOR CC* out of roundness control chart, *REG CC* regression-based approach control chart, *PCA CC* principal-component-analysis-based control chart, *ART NN* adaptive resonance theory neural network

11.5.2 Production Scenario with the Fixed-Effect Model

By comparing the results summarized in Figure 11.10 with those in Figure 11.9, one can observe that, with reference to the fixed-effect model as a reference production scenario, *i.e.*, a process stabler and/or more controlled than in the previous case, each of the competing approaches has better performance in signaling any kind of out-of-control condition. This can be simply explained by observing that no extra variability naturally affects the in-control profiles.

Among the competing approaches, the regression-based method improves dramatically its performance in signaling any kind of out-of-control condition. Indeed, the regression-based approach outperforms all of the other competing methods, even if in a few cases, especially when a high severity of the out-of-control condition is considered, the regression-based and the PCA-based approaches may have comparable performance.

Note that for the out-of-control conditions considered in our study (half-frequency, bilobe, and trilobe), the majority of alarms released by the regression-based approach are produced by the multivariate control chart of the vector of fitted parameters, while the PCA-based approach consists of a univariate control

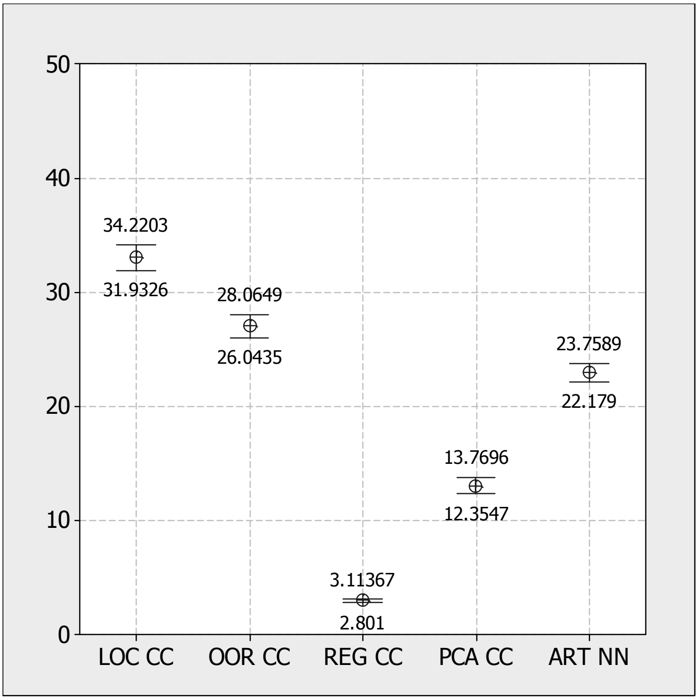


Figure 11.10 The 95% Bonferroni confidence intervals of the overall ARL for the competing approaches for the production scenario with the fixed-effect model

chart only (a Q control chart). When PCA is performed in the case of a fixed-effect model, no significant principal components are identified (Colosimo and Pacella 2007) as the PCA is performed after data centering and this first step consists in subtracting the mean pattern (described by the fixed-effect model) from each profile datum. PCA is thus performed just on the error terms and hence no significant principal component is correctly reported. Also, note that in this case the Q statistic is given by the sum of the squared difference between data observed at each location and the average profile at that location.

From the results in Table 11.3, it is also worth noting that the neural network has a good performance in signaling the out-of-control conditions, even if its performance in this scenario is not even close to comparable with the performances of the parametric approaches (regression-based and PCA-based). However, while the neural network approach has a better performance than the location control chart in signaling shifts in the second and third harmonics, the location control chart outperforms both the neural network and the OOR control chart in the case of the half-frequency out-of-control condition. For this specific out-of-control condition, the OOR control chart and the neural network have a similar performance.

Nevertheless, from Figure 11.10 it can be noted that the location control chart has the lowest overall power of detection when compared with the competing methods. As expected, also the OOR control chart does not have better performance than the model-based control charts (regression-based and PCA-based) and the neural network.

11.5.3 Overall Performance Measure

From the results previously discussed, no specific approach appears to be preferred among the five competing ones. Indeed, while the regression-based approach should be used to signal out-of-control profiles when a production scenario with fixed effects in the baseline model is considered, either the PCA-based approach or the neural network should be preferred in the case of random effects.

As mentioned already, both of the production scenarios are representative of actual industrial applications. Hence, both production scenarios, with the random-effect model and the fixed-effect model, should be considered as equally probable. For this reason, we also estimate a measure of the overall performance for each of the five competing approaches in a generic production scenario, assuming that the two kinds of production scenario previously considered are equally probable in actual applications. This overall performance is obtained as the mean ARL values, for each competing approach, in signaling a generic out-of-control condition of any severity for any type of production scenario (random-effect model and fixed-effect model). As previously done, we assume that all out-of-control conditions are equally probable, as the analyst is not expected to know *a priori* what kind of out-of-control condition will affect the production process and how severe it will be.

Figure 11.11 graphically depicts the 95% Bonferroni's confidence intervals of the overall ARLs of the five competing approaches in signaling a generic out-of-control condition of any severity for any type of production scenario.

It can be observed that the regression-based and the PCA-based approaches should be preferred as these approaches have better performance. It is also worth noting that the neural network also has an overall performance substantially similar to the overall performances of the model-based approaches, even though it is slightly worse than the overall performances of the regression-based and PCA-based approaches. Therefore, it is fair to conclude that the neural network approach can be considered a valuable option for profile monitoring.

As expected, the location control chart does not have overall performance better than the overall performances which characterize the regression-based approach, the PCA-based approach, and the neural network. Nevertheless, the location control chart outperforms the OOR control chart and, hence, when a method of very low complexity is needed for an actual application, the location control chart should always be preferred to the control chart of the geometric error.

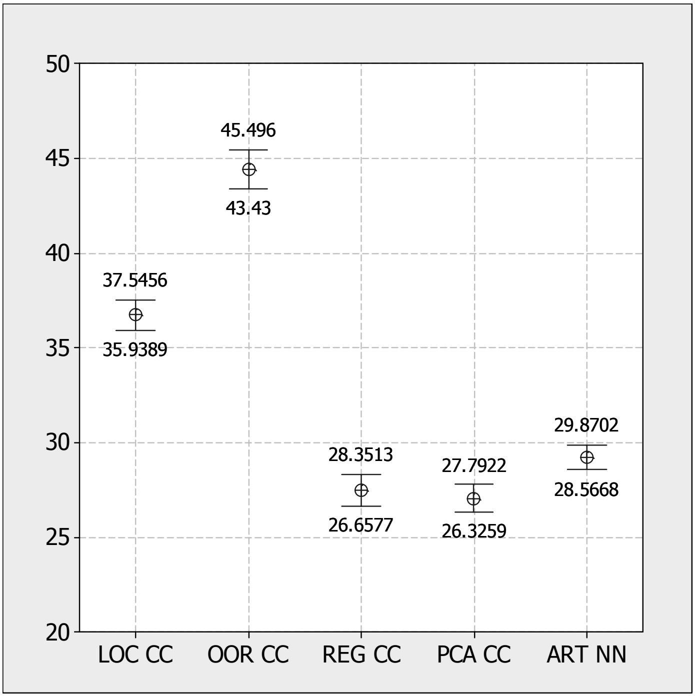


Figure 11.11 The 95% Bonferroni confidence intervals of the overall ARL for the competing approaches (both random effects and fixed effects as equally probable production scenarios)

11.6 Conclusions

This chapter presented a comparison study between different approaches for profile monitoring. Both analytical models and nonanalytical methods presented in Chapters 8–10 were considered.

With the former kind of approach, the in-control shape of the profiles is summarized by a parametric model, while profile monitoring is based on monitoring the parameters of this model. The control charts are based on the estimated parameters of the model from successive profile data observed over time (regression-based control charts and PCA-based control charts were considered in this work).

On the other hand, with the latter kind of approach, one can monitor one or more measures that reflect the discrepancies between observed profiles and a baseline profile established using historical data. In this chapter, both a location control chart and a neural network were considered as representative of nonparametric approaches for profile monitoring. With use of using a location control chart, a control region is implemented around a mean curve, where the borders of this control region are computed by the common Shewhart approach. In the case of a neural network, the baseline profile model is automatically established by means of a computer algorithm. This method allows computers to automatically learn from historical data the relationship to represent the profiles in space. The data set, from which the relationship is learned by the neural network, consists of in-control profiles only. After training, the neural network produces a signal when an input profile does not fit, according to a specific criterion, to the learned prototype.

Finally, a common control chart of the estimated geometric errors was also included in the comparison study, as this kind of approach is still the most representative of industrial practice.

By comparing the overall performance in phase II, which refers to the ability to signal a generic out-of-control condition of any severity for any type of production scenario (with a random-effect and a fixed-effect model), we can conclude that the extra effort required by the regression-based and PCA-based approaches is worthwhile. In fact, both the regression-based and the PCA-based approaches are more effective in signaling a generic change in the functional data. However, while the PCA-based approach shows superior robustness to a change of the productive scenario, the performance of the regression-based approach may be reduced dramatically in signaling out-of-control profiles in the case of a production scenario with a random-effect model.

Furthermore, though the neural network may be less effective for specific out-of-control conditions than the analytical/parametric approaches, the performances observed by using such a method are either comparable to or superior to those produced by the regression-based and PCA-based approaches in several cases. Therefore, it is fair to conclude that the neural network approach can be considered a valuable option for profile monitoring, especially when a model of the profiles is not available.

As expected, the individuals control chart of the geometric errors is not suitable for profile monitoring and hence this approach is not recommended for actual applications. In fact, when a method of very low complexity is needed for actual applications, either the location control chart or the neural network should always be preferred.

Acknowledgements This research was partially funded by the Italian Ministry of University and Research grant PRIN 2005 – "Characterization of the multiscale geometry of manufactured surfaces for process control and tolerance analysis".

References

Cho NW, Tu JF (2001) Roundness modeling of machined parts for tolerance analysis. Precis Eng 25:35–47

Colosimo BM, Pacella M (2007) On the use of principal component analysis to identify systematic patterns in roundness profiles. Qual Reliab Eng Int 23:707–725

Colosimo BM, Pacella M (2010) Control Charts for Statistical Monitoring of Functional Data, Int J Prod Res 48(6):1575–1601

Colosimo BM, Pacella M, Semeraro Q (2008) Statistical process control for geometric specifications: on the monitoring of roundness profiles. J Qual Technol 40:1–18

Ramsay JO, Silverman BW (2005) Functional data analysis. 2nd edn. Springer, New York

Woodall WH, Spitzner DJ, Montgomery DC, Gupta S (2004) Using control charts to monitor process and product quality profiles. J Qual Technol 36:309–320

Index