Doctoral theses at NTNU, 2020:259

Petr Chelishchev

Sample Strategies, Data Processing and Algorithms for Coordinate Measurements

Doctoral thesis

Petr Chelishchev

ISBN 978-82-326-4866-5 (printed ver.)
ISBN 978-82-326-4867-2 (electronic ver.)
ISSN 1503-8181

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Engineering
Department of Mechanical and Industrial
Engineering

Doctoral theses at NTNU, 2020:259

NTNU
Norwegian University of
Science and Technology

NTNU

Petr Chelishchev

Sample Strategies, Data Processing and Algorithms for Coordinate Measurements

Thesis for the Degree of Philosophiae Doctor

Trondheim, September 2020

Norwegian University of Science and Technology
Faculty of Engineering
Department of Mechanical and Industrial Engineering

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Engineering
Department of Mechanical and Industrial Engineering

ISBN 978-82-326-4866-5 (printed ver.)
ISBN 978-82-326-4867-2 (electronic ver.)
ISSN 1503-8181

Doctoral theses at NTNU, 2020:259

Printed by NTNU Grafisk senter

To Oleg and Irina

Preface

The work presented in this PhD thesis was carried out at the Department of Mechanical and Industrial Engineering of the Faculty of Engineering, the Norwegian University of Science and Technology (NTNU) into the five-year period from August 2015 to May 2020. The work was supervised by Professor Knut Sørby from the Department of Mechanical and Industrial Engineering (NTNU) and by co-supervisor, Doctor Vegard Brøtan from the SINTEF Manufacturing. The funding of this work was giving by the Research Council of Norway (NRF).

The PhD thesis is a part of a research project – Avansert Toleransesetting og Måleteknikk (ATOM). The project was established by professor Knut Sørby in collaboration with TechnipFMC and Kristiansands Skruefabrik og Mekanisk Verksted (KSMV).

This thesis is dedicated to the solution of the actual problems in production industry and product inspection. These problems have been selected together with my supervisor, based on my previous work experience in design (Rapp Bomek AS), production (RossNor Marine Ltd.), measurement control (Conoptica AS), in addition with our project meetings and discussions with other colleges from TechnipFMC and KSMV.

In order to solve the selected problems, different aspects of science were involved such as the parametric and non-parametric Statistics, Computation Geometry, Artificial Intelligence, advanced Manufacturing and Processing, Geometrical Dimensioning and Tolerancing (GD&T) based on the international standards (ISO) related to a newly revised Geometrical Product Specifications (GPS) family. During this work, it also became necessary to utilize software like PC-DMIS for the Coordinate Measuring Machine (CMM) and Spatial Analyzer (SA) for the laser tracker, and programming tools like MATLAB and R for development of algorithms and statistical simulations. The results of their applications in the GD&T inspection field are presented in this thesis.

I am deeply grateful to my supervisor, Professor Knut Sørby for his contributions and continues scientific supervising with tactful guiding of my ideas into the right track and for encouraging me along this five-year period. I am thankful to my co-supervisor Vegard Brøtan for sharing with me his practical experience in coordinate measurements. I appreciate to Kurt Martinsen, GD&T/GPS specialist engineer from TechnipFMC and former General Director of KSMV, Øystein Holte for their motivations and contributions to ATOM project and sharing with us the company experience in GD&T inspection issues. My special thanks to associated professor Alexander Popov from Baltic State Technical University for collaboration. I also wish to express my gratitude to professor Vadim Privalov from Institute of Physics, Nanotechnology and Telecommunications of Polytechnic University. My very special thanks to my parents for their patience and understanding.

The last but not least, I wish to thank the Norwegian Research Foundation for the financial support of this PhD work.

Summary

The PhD thesis proposes new approaches and algorithms related to the field of Geometrical Dimensioning and Tolerancing (GD&T) inspection based on ISO standards on Geometrical Product Specifications (GPS). The main purpose of this work is to improve reliability and accuracy of measurement strategies in GD&T inspection (tolerance verification), which plays an important role in manufacturing industry. The results of this work provide contributions for further development and standardization of measurement strategies and procedures with optimized sample strategies to provide a desired level of the measurement uncertainty.

The objective of the PhD project is to identify key factors that influence the measurement uncertainty and confidence level, try to clarify effects of these influences to improve the current state of measurement strategies in GD&T inspection. Generally, the measurement uncertainty is affected by many factors, e.g. choice of measuring equipment, calibration, control of environment conditions, workpiece orientation and clamping. In measurements with Coordinate Measurement Machines (CMMs), addition factors like stylus system qualification, choice of probe configuration, probe deflection, traveling speed, approach vector, coordinate system alignment, and datum system definition will influence on the measurement uncertainty.

Most of the measurements involved in this thesis were performed in a Leitz PMM-C-600 CMM, with an analogue measuring probe. All factors determining the uncertainty of CMM measurements can be divided into four main categories: *equipment*, *environment*, *workpiece*, and *operator influence*. The last category is mostly associated with the measurement strategy and procedure performed by the operator, which is of particular importance for the uncertainty in CMM measurements. Thereby, the following objectives were defined for this PhD project:

- development and investigation of sample strategies with *optimal sample size*;
- investigation of *outlier detection methods*;

- development of *algorithms* for calculation of *substitute* (reference) *elements*.

The choice of sample strategies and algorithms should be based on the tolerance requirements and the expected workpiece geometry deviations. The number of measurement points that are necessary to discover the decisive area of the workpiece surface for reliable geometrical verification is an important research question. In spite of many efforts of previous research, it remains an unsolved problem. Uniform guidelines based on international standards, which could determine criteria for optimal choice of the sample size has not been established so far.

Taking into account all aspects mentioned above, the main contributions of this PhD research are the following:

- An approach to define a confidence level for statistical tolerance intervals of various types of distributions and different sample sizes of measured variables based on CMM measurements of circular profiles after turning and milling machine operations (associated with ISO16269-6)
- Classification of data outliers and an investigation of outliers detecting procedures according to ISO16269-4
- An approach for optimizing the sample size for two-point diameter verification according to ISO14405-1
- An approach based on Artificial Neural Networks (ANN) to evaluate the maximum estimated error of the form deviation related to ISO1101
- New methods for defining substitute elements in estimation of the minimal distance between planes of cuboid objects by using the Minimum Volume Bounding Box (MVBB) principle

The contributions are presented in six papers, where five papers have been presented at conferences and conference proceedings, and one paper have been submitted for publication in a scientific journal. The contributions are addressed to quality engineers, metrology specialists and other researchers in metrology and GD&T inspection field.

Contents

Part II Papers

Part I

Main report

1 Introduction

1.1 Background

The desired geometry of a product is defined by the design specification, while the actual quality can be limited by technical capabilities of suppliers. In this case, we are considering the quality in terms of geometrical characteristic of the manufactured product. The technical capabilities of the suppliers include many aspects. Primarily, it concerns of availability of modern facilities i.e. machine tools, measuring equipment and employee qualifications. However, the availability of modern equipment itself does not guarantee the best quality and reliability of the final product. The competency level in the measurement technique is varying from one supplier to another. Therefore, a uniform measurement procedure for quality product control is demanded. The final decision about the product quality is taken in conformity with Geometrical Dimensioning and Tolerancing (GD&T) requirements [1]. In Europe, the GD&T requirements are based on Geometrical Product Specifications (GPS) [2] of ISO standards. As a part of the work presented in this thesis, a number of aspects towards a uniform measurement procedure is taken into account.

The GD&T inspection of critical components in subsea technology area is one of the best examples [3] when the right decision is especially important for assemblies and installations which are often carried out in a deep-water condition. Any geometry or dimension deviations of the critical parts can bring either to system malfunction or to uncompleted installation. Such failures may cost to suppliers and operating companies significant financial losses, what actually happens in practice time after time. Similar problems also exist in other industrial sectors but may be with less significant consequences. The reasons for product failures are not always easy to identify. Each particular case might relate either to the Product Design or to the GD&T inspection issues [4]. In many cases, the issues related to the Product Design can be eliminated by use of Computer-Aided Tolerancing (CAT) [5-8] before any drawings have been sent to production. However, this thesis is dedicated to an investigation of possible reasons and solutions for eliminating the GD&T inspection issues caused by measurement errors.

After all achievements in the accuracy improvements of equipment for coordinate measurements, there is still a number of factors related to measurement strategies, that may lead to the differences in the result which might be even larger than the value of the actual geometrical and dimension deviations themselves. The factors related to the measurement strategies (e.g. alignment, sample strategy, evaluation methods, outlier detection, data filtering etc.) are strongly dependent on the operator qualification and

company operation procedures. Consequently, measurement results of the same workpiece according to the same design specifications (production drawings) may vary dramatically from one operator to another. The main reason for such situation is that some of the mentioned factors related to the measurement strategy are neither complete standardised nor applied as default settings by a metrology software.

The goal of this research to determine and estimate an optimal value for those measurement factors and parameters, which should be standardized and set as default software settings in order to avoid an extra measurement uncertainty due to the operator influence. The existing algorithms (i.e. evaluation methods) of computation geometry for substitute elements have been further developed in this work.

Modern manufacturing and automation technology faces a number of challenges such as requirements of high precision and accuracy. Coordinate Measuring Machines (CMMs) play an important role in the part inspection and quality control. CMMs are universal and widely employed automated measuring systems in industry [1, 9]. Most of the research presented in this thesis is related to measurement with CMMs [10-13]. The CMM probe may utilize either contact (mechanical) or non-contact (optical) measuring principle [14]. Therefore, CMMs are very flexible and can be used for calibrations and measurements of complex components. It may include but not limited to such applications as GD&T inspection, process-capability studies (SPC), measuring of prototypes, calibration of gages and reference test-pieces etc. [9]. However, CMMs have not only benefits, but also some drawbacks. Initially, CMM is a post control system, and parts need to be taken from the CNC machine and delivered to a metrology laboratory. The other drawback is a limitation of overall dimensions of the component that can be physically measured by CMM.

There is one common principle, which is applied for all measurement systems including CMMs. Measurements are valid for accreditation with quality assurance systems only if all estimated measurement uncertainties are traceable to the meter-unit [9, 15]. After an innovation of lasers, a new stage in the interferometry and a new definition of the meter became possible [16-19].

1.2 Research challenges and questions

In spite of great efforts in developing a large number of ISO standards related to Geometrical Product Specification (GPS), not all aspects concerning measurement strategy have been covered [1]. Often the manufacturers due to lack of competence in GPS inspection have subjective interpretations of technical drawing requirements and analyses of the measured data. The proper procedures and measurement technique are not always applied. As a result, identical measuring task may have different approaches and not always comparable results. Thus, the purpose of this work is to investigate measurement accuracy of different measurement strategies and the supporting algorithms for integrating them into the measurement software and procedures. The challenges and questions related to the identified above problems are considered in the next sections.

1.2.1 Influence of the sample strategy on GD&T inspection

In spite of many scientific and technical proposals dedicated to the sample size problem, it remains unsolved. There is not a standardized guide regarding to the sample strategy provided so far. The detail state-of-the-art is given in section 2.1 or more extended version can be also found in [20]. The choice of the sample strategy strongly depends on the tolerance types and workpiece conditions. The sample size is a trade-off between the measurement uncertainty and time consumption. The inspection based on a finite sample size allows representing a fail part as a good one. These conditions often push suppliers to apply too small sample sizes. As result, a product delivered to a customer has inspection results approved, but the product does not correspond to the design specification indicated on the drawings. Then the following questions can be formulated:

- What is the minimum number of the measuring points that is necessary to determine the dimensional and geometrical deviations in order to make a reliable decision?
- What are the main influence factors (e.g. tolerance type, nominal size, the actual workpiece shape, measurement equipment, etc.)?
- How to estimate the influence and interrelation of these factors?
- Is it possible to define a sample size for the specific task based on the expected geometrical deviation of the workpiece?

1.2.2 Data outliers

According to ISO16269-4 [21], before any computation methods are performed on measured data, it must be checked for outliers. The presence of outliers in the measured data may affect dramatically on the final result of data filtration, data statistics, geometry computation methods, especially those which are based on extreme values such as contacting methods maximum inscribed, minimum circumscribed, Minimal Volume Bounding Box (MVBB), which will be clarified later.

However, outliers are not always incorrect measurements. The presence of outliers may indicate whether the manufacturing, measuring or data processing failures. Therefore, not only the presence of outliers is important but an investigation of their origins as well. To distinguish wrong measurements from the correct data can be a challenging task. Such task may be sophisticated even for an experience operator. Then the following questions can be formulated:

- What are outliers?
- What are possible reasons for outliers?
- How to distinguish a bad measurement from the workpiece deviations?
- Which outlier detection methods can be selected for the metrology tasks?
- How to estimate the efficiency of the selected methods?

1.2.3 Computation of the substitute elements

In order to estimate the form deviation, the actual feature derived from the measured points must be associated with an ideal feature (reference substitute element, e.g.

circle, cuboid, plane, line etc.) computed by an associated method. The reference substitute elements can be defined based on the three commonly used associated methods [1, 9]:

1. Minimum Zone (MZ) / Chebyshev method computes the minimum distance between two substitute elements bounding the actual feature (e.g. two circles, two parallel lines or planes etc.)
2. Contacting methods compute the bounding substitute elements such as maximum inscribed (MI) or minimum circumscribed (MC)
3. Least squares (LS) / Gauss method computes the minimum sum of the squares of local deviations of the actual feature from the substitute element

All three methods based on the same actual feature provide different result and only the LS method solution is always unique. Meanwhile, for the assessment of geometrical form deviations, MZ method is the most relevant method according to ISO 1101 and provides the least possible form deviation value compared to the other methods. Still the MZ method is computationally expensive and not always unique, thus the LS method often used by commercial software as the default method unless otherwise specified.

Circumscribed Bounding Box approach based on minimum volume principle has been presented in this work (Chapter 4). Three associated methods were demonstrated based on different level of the approximation. As it will be shown further, the evaluation results are varying from one method to another. Then the following questions can be formulated:

- How to select a proper associated method?
- How to keep the reproducibility of the evaluated results (make it comparable)?
- What is an estimated difference between three known MVBB methods?

1.3 Research objectives and approaches

The main objective of this work is to *improve* the measurement uncertainty and as a result the reliability of making decisions about product acceptance. Then the *influence factors* that affect the measurement uncertainty must be *identified*. The next step is *clarifying* the degree of the effect caused by the factors.

1.3.1 Objectives

The following research objectives were identified:

1. Investigation of the influence of the sample size for geometry assessment in Coordinate Measuring Machines;
2. Detection of outliers in CMM measurements;
3. Development of evaluation methods for workpieces with cuboid (rectangular parallelepiped) form.

1.3.2 Approaches

There are three main phases of this work that can be emphasized: a development of the research project plan, a development of the scientific articles and a development of the presented thesis. A more specific list for the last two phases is given below:

1. Study of literature related to the measuring systems. Discussion of the industrial measurements systems is given in the paper 5 [22].
2. Different analytical approaches were exploited to clarify the degree of influence of the sample size. Some approaches are based on non-parametric, parametric [10] and order statistics [11]. Another alternative based on Artificial Neuron Networks (ANN) [12] was applied to consider some specific details involved into the measuring procedures in the real practise.
3. In order to provide a comprehensive investigation both graphical and analytical approaches have been engaged in the outlier detection procedure [13, 21].
4. Estimation of the minimum distances between opposite faces of cuboid object were performed with the Minimum Volume Bounding Box (MVBB) methods [23]. These methods are well known in the Computation Geometry but their technique could be further approved. The technique applied in these methods is supported by the Analytical Geometry, and especially an apparatus of the Linear Algebra.

A more detail information about the objective factors and the applied approaches will be provided in the Theoretical Background and the Main Results (Chapter 2-4) of this thesis.

1.4 Limitations

The NTNU measuring laboratory is equipped by Coordinate Measurement Machine (CMM) Leitz PMM-C-600. All measurements used in the developed algorithms and simulations, were performed with CMM. The measured data provided by other alternative measuring systems have not been involved for simulations presented in this report.

Most of the approaches are developed for two-dimension problems e.g. the sample size problem is considered for radial sections only. The problem regarding number of sections (and their locations) has not been considered in this work. Very few sources were dedicated to the problem related to the *radial section method*, *generatrix method*, *extreme position method* etc. exploited for tolerance inspection [1, 24]. It seems that existing methods are not well studied so that this problem remains opened, and ISO standard does not provide a complete guidance regarding these measurement strategies [25].

All simulations related to the circular profiles employed Least Squares (LS) method, which always provides the unique solution. Other evaluation methods such as Minimum Zone (MZ, Chebyshev), contacting methods (MI, MC) may provide more than one solution, and they have not been involved in this report.

Generally, the filtration technique is applied on measurement data to separate roughness and waviness from geometrical deviation and for attenuation of measure-

ment uncertainty. The filtration methods are well defined and standardized in ISO 16610 series. When filtration is required, then the minimum number of measurement points and the tip ball diameter of the probe are defined by the filter characteristics. According to the Nyquist theorem, at least seven points per undulation must be sampled [26]. In general, the measuring repeatability of Leitz PMM-C-600 is less than one micrometer. The stylus tip ball diameters of 5 mm and 10 mm were used for measurements presented in this work. Thereby the roughness components were significantly filtered out by the tip ball. No additional digital filtration was applied and hence, it has not been considered in this report.

1.5 Interconnections between Papers and Objectives

The main results of this PhD report are presented in the form of contributions for the objectives formulated in section 1.3.1. Altogether, it is a collection of six scientific papers. The interconnections between the objectives and the papers are illustrated in **Fig. 1**. As shown by the block diagram, the objective "Sample Size" has been divided into three sub-objectives. Chapters 2–4 describe the paper contributions and some additional observations of corresponding research.

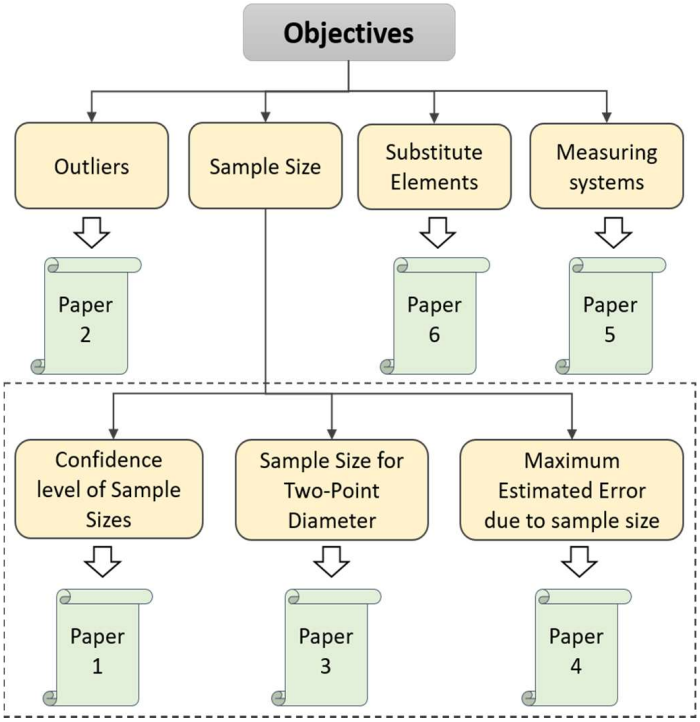


Fig. 1. The block diagram of interconnections between objectives and papers

2 Optimization of sample size

2.1 Overview of the sample strategy for discrete point measurements

Most of the work presented in the author's papers [10-13] are based on CMM applications. Coordinate measurements with CMM is a collecting of measuring points on the surface of a workpiece. The minimum number of sampled points depends on the geometry of a substitute element that must be calculated using mathematical algorithms. However, the mathematical minimum number of points is insufficient to estimate the geometry of a real workpiece, due to form deviation. Since there are always form errors, more measuring points are required. Though an official guide based on the international standard has not been submitted yet, some recommendations for the mathematical and practical minimum number of points are provided by British Standard BS-7172 [24] and [1, 9]. Still, the recommended minimum number of points may lead to underestimation of the actual form deviation. The evaluation of estimation error due to the finite sample size is the objective of numerous studies.

Mesay et al. [27] have classified the manufacturing process error into systematic and random components. In another paper, Qimi, Mesay et al. [28] have estimated the systematic roundness errors by use of Fourier analyses. In a case study on machined parts with a diameter of 20 mm, they found that at least 22 measured points were necessary to achieve a consistent value of the roundness error.

Other authors have investigated the measurement uncertainty due to the sample size based on approximation of an aperiodic deterministic profile with Fourier series [29, 30]. There were some more simple approaches based on the normal distribution [31], chi-square distribution [32, 33], and fuzzy logic [34]. A more comprehensive study can be found in [20].

Moschos et. al [35] suggested a Bayesian regularized artificial neural network (BRANN) model trained with relatively small sample size to predict a variability of large data sample. The intention of the method was to improve the final uncertainty by decreasing the uncertainty in the coordinates of each point. Other authors determine an optimal inspection sample size based on measurement errors approximated by ANN for various machine processes and nominal sizes [36]. A few studies have been conducted in applying ANN for reverse engineering for automatic inspection with a CMM [37, 38]. A comprehensive study has been presented by Sładek [39] on simulation methods based on ANN and Monte Carlo-methods for estimation of uncertainty of virtual coordinate measurements. The developed Virtual Neuro CMM was compared with other existing models such as MegaKal and Virtual CUT. The MegaKal model has been developed in the National Metrology Institute of Germany (Physikalisch Technische Bundesanstalt, PTB). The Cracow University of Technology in the Laboratory of Coordinate Metrology has developed the Virtual CUT model.

As long as the nature of the workpiece surface errors is a result of a large number of factors with not fully understood correlation interfered with measurement uncertainty in various environment conditions, it leads to the development of a more general solution with virtual workpiece profile. This virtual profile provides opportunity for simulation of measurements to enable a choice of an optimal sample strategy. Such approach was presented in Paper 4 and it is described in section 2.6.2 of this thesis as well.

The optimal choice of discrete points also depends on utilized evaluation methods and tolerance types [40, 41]. Since, the reliability and quality of CMM sample assessment depends on the density and the location of the measured points [25], the inspection is often a compromise between time consumption, cost, and the measuring uncertainty.

2.2 Kernel density estimation

In real practice, the data distribution is often unknown and/or may contain outliers. Therefore, the statistical tolerance intervals should be estimated without the assumptions about a specific statistical distribution. The non-parametric statistic can be considered as a helpful option in such cases. The main idea of the non-parametric statistics is to avoid assumptions about a probability function. The distribution is estimated directly from a data.

One of the most known graphical method of non-parametric statistic is the histogram technique. The data is estimated in the form of rectangles with their bases of equal bandwidth b:$[r_{min} + (k-1)b, r_{min} + kb], (k = 1,2, \dots m)$, where r_{min} is the smallest random variable in the data sample. The rectangles are centered in the midpoints of each interval with their heights corresponding to relative frequencies [42]. Then the estimator of the probability density function (pdf) can be expressed by

$$\hat{f}(r) = \frac{1}{b} \frac{Number\ of\ observations\ r_i\ in\ the\ particular\ rectangle}{Number\ of\ all\ observations\ in\ the\ data\ sample} \tag{1}$$

This is an example of the simplest non-parametric estimator used for a *relative frequency histogram* or *a probability histogram*. A choice of the rectangle bandwidth b has a significant effect on the form of the histogram. An example of a probability histogram is shown in **Fig. 2**. The main disadvantage of the relative frequency histogram is discreteness.

In order to avoid a lack of continuity, another alternative of non-parametric estimation of pdf [43] can be used e.g. a Rosenblatt-Parzen type kernel estimator of $f(r)$ at a given point r:

$$\hat{f}(r) = (Nb)^{-1} \sum_{i=1}^{N} K\left(\frac{r - r_i}{b}\right), \qquad -\infty < r < \infty \ . \tag{2}$$

Where N is a data sample size, and K is a standardized weighting function with the bandwidth $b = 1$ (smoothing parameter). The weighting function is called the *kernel*. The kernel defines the form and properties of the weighting function. There is a num-

ber of commonly used kernels such as Rectangular, Triangular, Gaussian and Epanechnikov function etc. If a function has properties of a symmetric pdf:

$$\int K(t)dt = 1, \qquad \int tK(t)dt = 0, \qquad \int t^2 K(t)dt = C, \tag{3}$$

where $t = \frac{r-r_i}{b}$ and C is a non-zero finite constant it can be employed as a kernel. Also the estimator $\hat{f}(r)$ is expected to have the properties of pdf :

$$\begin{cases} f(r) \geq 0, \; for\, all\, r \in \mathbb{R} \\ \int_{-\infty}^{\infty} f(r)dr = 1 \\ P(r-b < R < r+b) = \int_{r-b}^{r+b} f(r)dr\,. \end{cases} \tag{4}$$

The proper choice of K and b is a subject for optimization [44, 45]. With larger smoothing parameter b the fluctuations of $\hat{f}(r)$ is reduced. The optimal value of b depends on many factors e.g. type of kernel function, shape of unknown pdf $f(r)$, the data sample size N, etc.. The accuracy of the kernel density estimator may be estimated as a mean square error (MSE), a mean integrated squared error (MISE) and an asymptotic mean integrated squared error (AMISE). The optimal kernel function with respect to MISE is the Epanechnikov weighting function [46]:

$$K(t) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}t^2\right), & for\ |t| < \sqrt{5}, \\ 0, & elsewhere. \end{cases} \tag{5}$$

An example of a pdf data estimation by kernel is illustrated in **Fig. 2**.

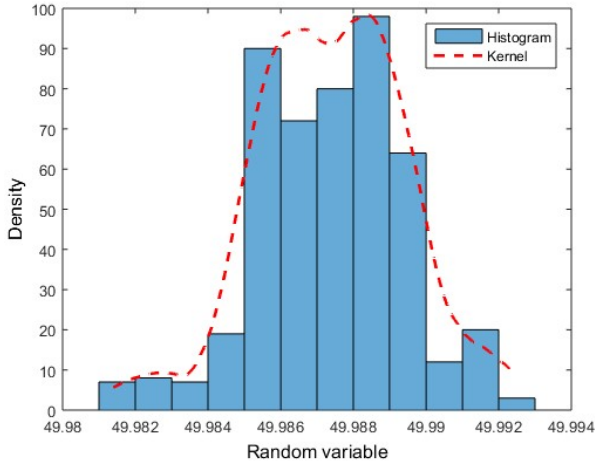


Fig. 2. An example of data estimation with Histogram and Epanechnikov kernel estimator with default bandwidth based on the sample size of 480 observations

The main advantage of kernel estimation is that we get the continuous function $\hat{f}(r)$ as an output of the method. This continuous function can be further used for development of simulation algorithms. Some of the papers included in this thesis (Paper 1, Paper 2, Paper 3 [11, 13, 10]) involves the non-parametric technique based on the kernel density estimators (sections 2.6 and 3.2). The estimators of pdf used in these papers utilize the Epanechnikov kernel.

2.3 Analytical approach based on Statistical Hypothesis

In this section, we are also considering the parametric approach based on the hypothesis test. In order to analyze the statistical inference, the entire population must be examined. However, such approach may be either very expensive or not even available. An alternative to this is to use statistical hypothesis. The applications based on statistical hypothesis has a wide demand among engineering and scientific practice. The statistical hypothesis technique is used for some outlier detection procedures [13, 47, 48], which is described in Chapter 3. In addition, an example of an alternative analytical approach is to be demonstrated here. An optimization of the sample size for the two-point diameter (Paper 3) is based on this principle [10].

According to the definition in [42], *a statistical hypothesis is an assertion or conjecture concerning one or more populations*. Instead of measuring the entire population, we can employ random samples taken from the population to collect a necessary evidence to reject or not reject the stated hypothesis. The choice of the *test statistic* and the *statement of the null hypothesis* is very important, and it is a crucial part of the reliability of the hypothesis test. In order to provide a strongly supported *alternative hypothesis*, one should accept the alternative hypothesis in a form of rejection of a null hypothesis. As an example, if we would like to provide a strong evidence that a new developed engine has lower gasoline consumption, the tested null hypothesis should be of the form “there is no reduction of the gasoline consumption of the new

engine". As a result, the acceptance of the engine improvement (alternative hypothesis) is achieved by the rejection of the null hypothesis.

A sample of data must be collected from the population in order to provide sufficient evidence to reject the null hypothesis. Two possible wrong decisions may be taken in the hypothesis test. One wrong decision is the rejection of the null hypothesis when, in fact, it is true. This is called an *error type I*. The probability of this error is often designated α, which is also called as *level of significance*. The other error is committed by acceptance (non-rejection) of the null hypothesis, when, in fact, the alternative hypothesis is true. That is called an *error type II* [42]. The probability of this error is designated β. The probability of committing error type I or error type II depends on the sample size. The sample size can be determined with reasonable probability values of both type errors.

To keep this simple, let us consider the situation where measuring data δ_i follows the normal distribution $N(\mu_1, \sigma_1)$ with the mean value μ_1 and the variance σ_1^2. We will test if the mean value μ_1 is larger than a reference mean μ_0 of $N(\mu_0, \sigma_0)$, i.e. $\mu_1 > \mu_0$. Such hypothesis test is called a *one-tailed test* (upper-tailed test in the underlying situation). Thereby the tested hypotheses can be formulated in this way: the null hypothesis is $H_0: \mu_1 = \mu_0$, hence the alternative hypothesis is $H_1: \mu_1 > \mu_0$. We use the sample mean $\bar{\delta} = \frac{1}{n}\sum_{i=1}^{n} \delta_i$ as the estimator of the population mean μ_1 and as the test statistic. Now, we can determine the lower bound value δ_k for a critical region:

$$\bar{\delta}_k = u_{1-\alpha} \cdot \sqrt{\sigma_0^2 / n} + \mu_0 \,, \tag{6}$$

where $u_{1-\alpha}$ is the quantile of levels $1-\alpha$ for distribution $N(0,1)$ and n is the sample size of the random measurements δ_i. In our example, we consider the level of significance $\alpha = 0.05$ i.e. $u_{0.95} = 1.645$.

Then according to the terms of the errors of type I and II, formulated above, we can write the following:

$$\begin{cases} \mathrm{P}\left[\bar{\delta} > \bar{\delta}_k \,/\, H_0\right] = 1 - \mathrm{F}\left(\dfrac{\bar{\delta}_k - \mu_0}{\sqrt{\sigma_0^2 / n}}\right) = \alpha, \\ \mathrm{P}\left[\bar{\delta} \le \bar{\delta}_k \,/\, H_1\right] = \mathrm{F}\left(\dfrac{\bar{\delta}_k - \mu_1}{\sqrt{\sigma_1^2 / n}}\right) = \beta \,. \end{cases} \tag{7}$$

The expressions in (7) can be converted into the form:

$$\begin{cases} \dfrac{\overline{\delta}_k - \mu_0}{\sqrt{\sigma_0^2 / n}} = u_{1-\alpha} \\ \dfrac{\overline{\delta}_k - \mu_1}{\sqrt{\sigma_1^2 / n}} = u_{\beta}. \end{cases} \tag{8}$$

Then, the upper equation of (8) can be converted into (6) and substituted into the lower equation of (8) so that all this can be expressed in terms of the required number of observations (an optimal sample size):

$$n = \left(\frac{u_\beta \sigma_1 - \sigma_0 u_{1-\alpha}}{\mu_0 - \mu_1} \right)^2, \tag{9}$$

where u_β is the quantile of levels β for distribution $N(0,1)$ related to the error type II. Thus, the optimal value of the sample size can be determined for a reasonable balance between α and β values. The relation of the sample size n and the mean difference $\Delta\mu = \mu_0 - \mu_1$ was computed in the *R programming environment*, and the results are shown in **Fig. 3**. The quantile $u_\beta = -1.645$ corresponding to the level $\beta = 0.05$ was applied for the calculation. Three data sets A, B and C with corresponding standard deviations $\sigma_1^A = 1.1\ \mu m$, $\sigma_1^B = 0.9\ \mu m$, and $\sigma_1^C = 1.4\ \mu m$ are considered.

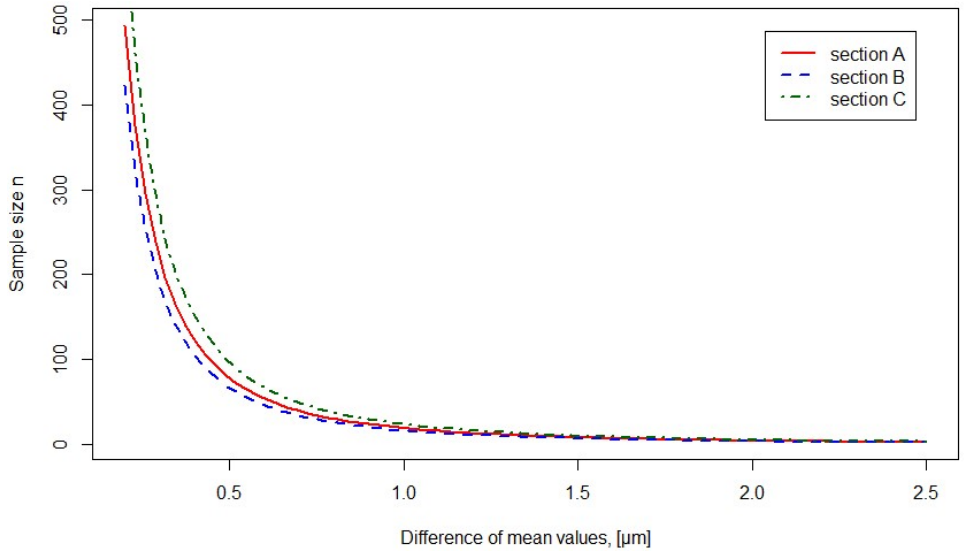


Fig. 3. The sample size vs mean difference for cross-sections A, B and C

According to **Fig. 3** and (9), the sample size n can be reduced, if the distance between two means $\Delta\mu$ is increased e.g. for the mean difference $\Delta\mu = \sigma_0 = 1.6$ µm, the determined sample sizes are $n_A = 8$, $n_B = 8$, and $n_C = 10$ measurements respectively. For detection of smaller $\Delta\mu$, the sample size increases dramatically. However, in prac-

tise more parts have deviations with larger $\Delta\mu$, thus fewer measuring points might be required for their inspection.

2.4 Distribution-free model

In practice, a sample of measured variables $r_1, r_2, \ldots r_n$ often follows an unknown distribution and hence, the use of parametric statistics can be misleading. Then a statistical tolerance interval can be determined with the sample order statistics $r_{(1)} \leq r_{(2)} \leq r_{(n)}$ of a data sample of n independent observations [49]. Statistical tolerance intervals limited by the smallest and the largest sample order statistics independent on the sample distribution are called *distribution-free statistical intervals* or *non-parametric statistical intervals*. ISO 16269-6 [50] provides the procedures for the determination of required sample size based on the order statistics for the desired population proportion p with the desired confidence level $1-\alpha$. According to ISO 16269-6, "*a statistical tolerance interval* is an estimated interval, based on a sample, which can be asserted with *confidence level* $1-\alpha$ (e.g. 0.95), to contain at least a specified *proportion* p (e.g. 0.95) of the items r_k in the population. The limits of a statistical tolerance interval are called *statistical tolerance limits*."

According to Wilks [51, 52] and ISO16269-6 (in case of the continuous function), the interval with 100(1-α) % confidence that at least 100p % of the population lies between the v^{th} smallest observation (i.e., order statistic $r_{(v)}$) and the w^{th} largest observation (i.e., order statistic $r_{(n-w+1)}$) of the sample (see **Fig. 4**), is determined by solving the cumulative binomial distribution function for the smallest sample size n_{min} as follows:

$$\sum_{k=0}^{v+w-1} \binom{n}{k} p^{n-k} (1-p)^k \leq \alpha, \quad (0 < p < 1), (0 < \alpha < 1), \tag{10}$$

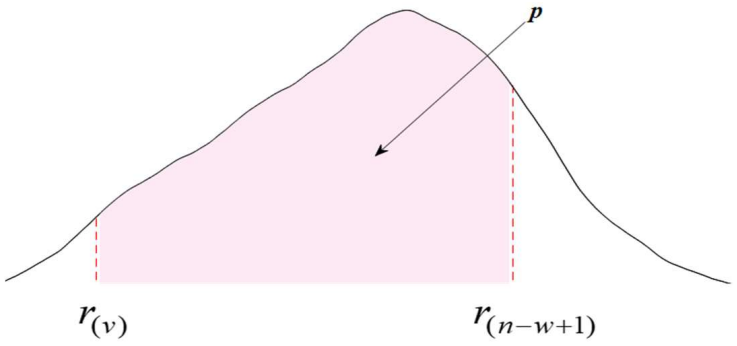


Fig. 4. The population content p constrained between the v^{th} smallest observation and the w^{th} largest observation of the sample of an unknown distribution

where $w \geq 0, v \geq 0, v + w \geq 1$, and $\binom{n}{k}$ is the binomial coefficient. The binomial coefficient is calculated as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}, \quad for\ k = 0,1,2,...n \ . \tag{11}$$

The form deviation of a workpiece feature (e.g. roundness) is assessed as the difference between two extreme measured values (e.g. $r_{max} - r_{min}$). The form deviation of the workpiece might be associated with *the two-sided distribution-free statistical tolerance interval*. Then, the interval with 100(1-α) % confidence that at least 100p % of the surface profile lies between the smallest observation (r_{min}) and the largest observation (r_{max}) of the sample, is determined by the following expression, which is derived from (10):

$$n_{min} \cdot p^{n_{min}-1} - (n_{min} - 1) \cdot p^{n_{min}} \leq \alpha \tag{12}$$

where $v + w = 2$ and n_{min} is the minimum sample size. The *robust estimations* of the minimal sample sizes based on the distribution-free model (12) for two-sided nonparametric statistical tolerance interval are tabulated in **Table 1**. The number of observations is the natural numbers $n_{min} \in \mathbb{N}$, thus all negative and complex solutions are excluded. The values are rounded up to the nearest integer.

The Paper 1 is based on the presented approach and the contribution results are given in section 2.6.1.

Table 1. The minimum sample size n_{min} for the two-sided nonparametric tolerance limits

Confidence level, $100(1-\alpha)\%$	**Proportion of population, *p***				
	0.500	**0.750**	**0.900**	**0.950**	**0.990**
50	3	7	17	**34**	168
75	5	10	27	53	269
90	7	15	38	77	388
95	8	18	46	**93**	**473**
99	11	24	64	130	662

2.5 Artificial Neural Network Approach

The state-of-the-art in Artificial Neural Network (ANN) is based on our understanding of biological neurons' function [53]. One of the most important advantage of

ANN is that it can imitate the behaviour of an unknown relation between an input and output data. In case of lack of knowledge about an analytical function, the ANN can provide relatively precise solution based on the limited experimental data, which is called *training set*. It is important to notice that the solution of an ANN is not a unique solution, but one that satisfied the minimal error requirements.

An artificial network is composed of differently connected *artificial neurons*, which are named as *processing elements* (PE). The fundamental principle of the PE is shown in **Fig. 5**.

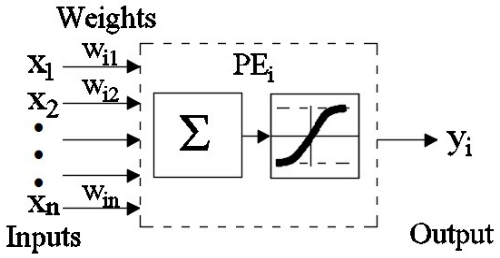


Fig. 5. A single Processing Element (PE) of ANN

The PE executes a number of functional operations with a given input such as *multiplication* by weight scalar w_{ij}, *substitute* of threshold (bias) scalar s_i, *summation* of all results a_i from other inputs x_j, and finally, *transformation* of results by the activation function $f_i(a_i)$ [54]. The final output signal can be mathematically expressed as

$$y_i = f_i\left(\sum_{j=1}^{n} w_{ij}x_j - s_i\right) = f_i(a_i), \tag{13}$$

where the summation a_i is given by

$$a_i = \sum_{j=1}^{n} w_{ij}x_j - s_i \tag{14}$$

There is a number of activation factions $f_i(a_i)$ available i.e. a step function, a linear function, a log-sigmoid function etc. In this work, we exploit the tan-sigmoid activation function (*tansig*), which is commonly used for calculating a layer's output to achieve faster and better stability of network processing [55]. The tan-sigmoid activation function can be written in the following form:

$$y_i = f_i(a_i) = \frac{2}{1+e^{-2a_i}} - 1 \tag{15}$$

The activation function (15) is defined in the interval [-1; 1].

In fact, the PE can have more than one output and it is often the case. The PEs (artificial neurons) are connected into input, hidden and output layers creating the artificial neural network. Multilayer ANN can include many hidden layers but to reduce a computation time many commercial systems usually do not exceed two hidden layers. In general, the basic procedure for design of a neural network of an arbitrary architecture may include a number of standard steps [54, 56]:

1. Preparing and pre-processing (normalizing) training data [57]
2. Creating a network structure
3. Configuring the network
4. Initialization of weights and biases
5. Training, validation and testing of the network

In our study, we utilize a Supervised Back-Propagation (BP) Artificial Neural Network. The network type of Supervised Back-Propagation is related to the learning strategy of ANN [54, 58, 59]. ANNs based on the BP strategy are popular in many applications, including the industrial sectors. An example of a multilayer feedforward PB ANN with a single hidden layer is illustrated in **Fig. 6**.

The BP learning includes both a *forward* and *backward* phase. The forward phase is a calculation of the actual outputs (network response). The backward phase is a adjustment of the weights in order to reduce the deviation between the actual output and the desired target until the deviation achieve an acceptable value. The processing operations of network between layers are described in the simplified form (without biases) below. Then the network input for the hidden layer is the following [54]:

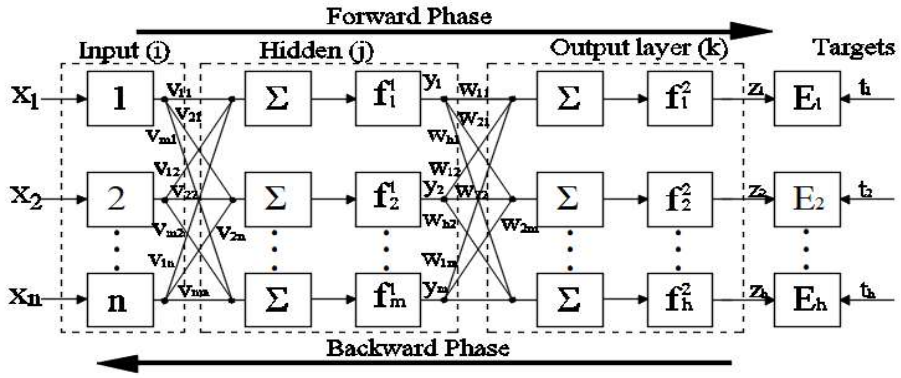


Fig. 6. Multilayered *n-m-h* feedforward PB ANN with a single hidden layer

$$I_j = \sum_{i=1}^{n} v_{ji} x_i, \quad (i = 1,\ 2,...n;\ j = 1, 2,...m)\ , \tag{16}$$

where v_{ij} is randomly generated weight scalars in the interval [-1; 1] corresponding to the *tansig* activation function (15) between a unit *i* of the input layer and a unit *j* of the hidden layer. Then, the network input for the output layer is the following:

$$H_k = \sum_{j=1}^{m} w_{kj} y_j, \quad (j = 1, 2,...m,\ k = 1, 2,...h)\,, \tag{17}$$

where w_{kj} is randomly generated weight scalars in the interval [-1; 1] corresponding to the linear activation function (*purelin*) between a unit *j* of the hidden layer and a unit *k* of the output layer. The output of the unit *j* of the hidden layer can also be expressed through the corresponding activation function:

$$y_j = f(I_j), \quad (j = 1, 2,...m)\,. \tag{18}$$

Analogically, the output of the unit *k* of the output layer can be expressed by:

$$z_k = f(H_k), \quad (k = 1, 2,...h). \tag{19}$$

Eventually, the *forward phase* of calculation of the actual network output can be expressed by substituting (17) into (19) with substituting y_j by (18) and substituting I_j by (16):

$$z_k = f(H_k) = f\left(\sum_{j=1}^{m} w_{kj} y_j\right) = f\left(\sum_{j=1}^{m} w_{kj} f(I_j)\right) = f\left(\sum_{j=1}^{m} w_{kj} f\left(\sum_{i=1}^{n} v_{ji} x_i\right)\right)\,. \tag{20}$$

The difference between the *output* z_k and the *target* t_k of the feedforward network is typically calculated as the *mean squared error* (MSE):

$$\varepsilon_{mse} = \frac{(z_k - t_k)^2}{2}\ , \tag{21}$$

and a total MSE over all network output nodes (the final output):

$$E = \frac{1}{m}\sum_{k=1}^{m} e_k^2\ , \tag{22}$$

where $e_k = z_k - t_k$.

The backpropagation learning based on Levenberg-Marquardt training algorithm is used in this work for the *backward phase* (adjusting of the weights v_{ij}, w_{kj}) [60]. The

new modification was designed to improve the convergence speed of network training without calculating the Hessian matrix. The Levenberg-Marquardt algorithm is a modification of Newton's (Gauss-Newton) method [61]:

$$\Delta w = -[\nabla^2 E(w)]^{-1} \nabla E(w) . \tag{23}$$

If function $E(w)$ has a form of the performance function given by (22), then the following approximation can be applied:

$$\nabla E(w) = J^T(w) e(w) \ , \tag{24}$$

and

$$H(w) = \nabla^2 E(w) = J^T(w) J(w) , \tag{25}$$

where H is the Hessian matrix, J is the Jacobian matrix of first-order partial derivatives of the network errors with respect to weights; e is a vector of the network errors, $\nabla E(w)$ is the gradient of function $E(w)$, which needs to be minimized with respect with the weighting parameter w. Then the Gauss-Newton method is expressed by:

$$\Delta w = -\left[J^T(w) J(w) \right]^{-1} J^T(w) e(w) . \tag{26}$$

Finally, the Levenberg-Marquardt modification of Gauss-Newton method can be written in the following form:

$$\Delta w = -\left[J^T(w) J(w) + \lambda I \right]^{-1} J^T(w) e(w) , \tag{27}$$

where I is the identity matrix, λ is a Lagrange multiplier, which regulates whether Newton ($\lambda = 0$) or the gradient descent method (λ is large) is performed. The algorithm has a good performance on nonlinear function fitting problems. In spite, a large memory consumption, Levenberg-Marquardt is the fastest supervised feedforward neural network (up to few hundred weights) optimization algorithm with an efficient implementation in MATLAB [56].

Paper 4 presents a new developed model based on this approach. The results of this model application are also discussed in section 2.6.2.

2.6 Contributions to objective 1

The research questions related to the sample strategy have been formulated in section 1.2.1. The objective related to the sample size has been stated in section 1.3.1. The choice of the sample size is a decision which must be taken by the CMM operator while planning the measurement procedure. A comprehensive standardized guide for the sample strategy and sample size has not been provided so far. Some recommendations can be found in British standard BS 7172 [24] and the old German standards TGL 39093 to TGL 39098, and TGL 43041 to TGL 43045 [1]. The sample strategy

depends on the uncertainty of the measuring system, the tolerance types, and the workpiece shape. As an example, the measurement of cylindricity requires more measured points than the measurement of straightness of the cylinder axis.

Especially, there is a difference between assessment of geometrical deviations (form, orientation, location) and size deviations. In case of form deviation assessment (e.g. roundness) it is necessary to detect the difference between the smallest and the largest value (Paper 1 [11]), while for the assessment of the size deviation, it may be enough to estimate only the mean value (Paper 3 [10]). To be able to define which measuring instrument or method that should be applied, we need an assumption about the maximum value of geometry deviations and measurements errors (Paper 4 [12]).

2.6.1 Geometrical deviation.

The assessment of the form deviation (e.g. roundness) depends on detection of maximum and minimum variables (e.g. radius extreme values). Paper 2 is based on the research of radius variable distribution derived from CMM measurements of circular sections of internal cylinders manufactured by turning processes.

In this section, for demonstration of the universality of the proposed approach, we present the additional results (non-published) based on eight measured circular profiles (four profiles with nominal diameter of 100 mm and four profiles with varying nominal diameters of 200-500 mm) manufactured by milling processes. Each section has very specific probability density function (pdf), different from pdfs of the other section profiles. Each pdf was estimated by Kernel estimator based on 480 measured radii according to (2) in section 2.2. The coordinates of the circle center (x_c, y_c) were calculated with LSC method based on the coordinates of 480 measured points equally spaced around a circle. The radius variables were calculated as a follows:

$$r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \tag{28}$$

The estimated pdf functions for four circle section profiles with nominal diameter 100 mm are illustrated in **Fig. 7**.

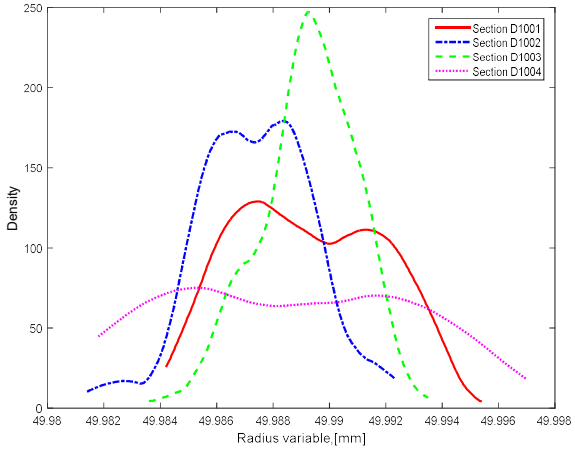


Fig. 7. Estimated pdf based on 480 measured radii for four circular sections with nominal diameter 100 mm

The estimated pdf for normalized radius variables of four other circular sections with nominal diameters 200 mm, 300 mm, 400 mm and 500 mm are illustrated in **Fig. 8**.

In order to define the minimum sample size related to the assessments of the form deviation, the statistical simulations have been developed in MATLAB. The main idea of simulation is motivated by the distribution-free analytical model (12), proposed by Wilks (section 2.4). It means that a new robust approach is not based on assumptions of the normal distribution, or other statistical distributions.

The coordinates (x_i, y_i) of the measured points were used as the *input* of the algorithm. The estimated pdfs of random radius variables (**Fig. 7**, **Fig. 8**) were computed by eq. (2). The following sample sizes were considered n = [5; 10; 15; 34; 60; 90; 93; 95]. The sample size of **34** and **93** measured points have a special interest.

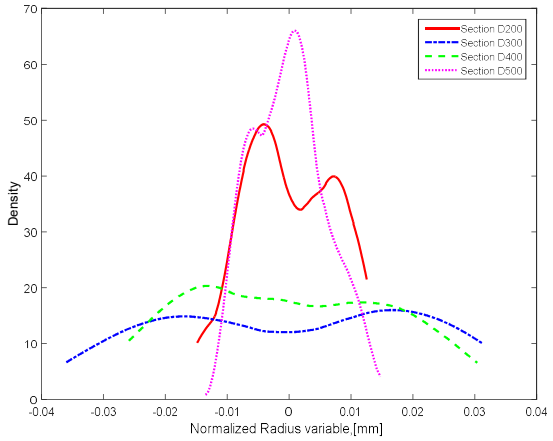


Fig. 8. Estimated pdf based on 480 normalized radii for nominal diameters 200 mm, 300 mm, 400 mm, 500 mm

According to **Table 1**, thirty-four-point sample allows to detect 95 % of the population with 50 % confidence level (CL), while ninety-three-point sample should have at least 95% CL for detecting 95 % of the population. Thus, these samples can be further used as the references for verification of the algorithm. We employ the random generators based on the estimated pdfs to simulate random radius variables. There are 10^5 iterations applied for each sample size in our simulation algorithm. The maximum and minimum values of the radius (r_{max}, r_{min}) are defined in each iteration. Then, the proportion p of the population, which is the area under the corresponding pdf curve (**Fig. 4**) is bounded by the two defined extreme ordinates (r_{max}, r_{min}). Thus, for all types of probability distribution curves, p is represented by the area, which is a difference of the two cumulative distribution functions (cdf):

$$p_i(r_{\min} < r < r_{\max}) = \int_{-\infty}^{r_{\max}} K(\mu_k, \sigma_k)dr - \int_{-\infty}^{r_{\min}} K(\mu_k, \sigma_k)dr \tag{29}$$

where $K(\mu_s, \sigma_s)$ is a distribution estimated by the kernel (shown in **Fig. 7**, **Fig. 8**) that corresponds to each section profile. Then, the condition $p_i \geq 0.95$ is checked for each iteration. Whether the condition is satisfied or not the sum S will be updated either as $S = S + 1$ (success/Yes) or $S = S + 0$ (failure/NO). Eventually, the confidence level (CL) for each sample size n_j ($j = 1,..8$) is calculated as

$$CL = \left(\frac{1}{M} \sum_{i=1}^{M} S_i \right) \cdot 100\% \tag{30}$$

where M is the total number of iterations. The block-diagram of the simulation algorithm to estimate the CL for detecting 95 % of the population due to various sample sizes n_j is shown in **Fig. 9**.

The robust estimations of CL due to the sample size to detect 95% of the population for four circular profiles with the same nominal diameter 100 mm are given in **Table 2**. The developed approach can be also applied for profiles with larger diameters as shown in **Table 3**. As can be observed from the computation results in both tables, the CL is very similar and thus, it is independent on the size.

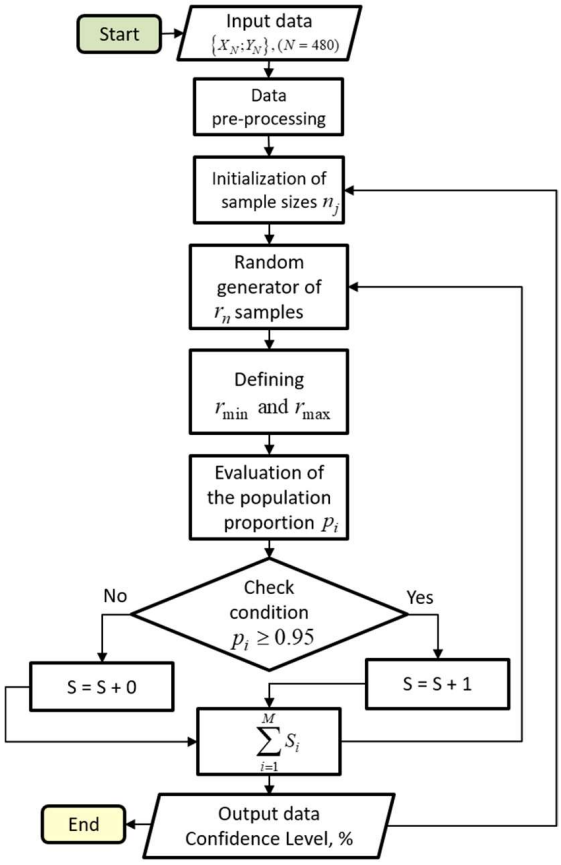


Fig. 9. The block-diagram of the simulation algorithm

Table 2. The confidence level $CL = (1-\alpha)100\%$ versus the sample size n to detect at least 95% ($p \geq 0.95$) of the population (for the circle profiles with **diameter of 100 mm**)

				Section D1001				
Sample size	5	10	15	**34**	60	90	**93**	95
CL , %	2.3	7.9	17.5	51.4	80.0	94.5	94.9	95.7
				Section D1002				
Sample size	5	10	15	**34**	60	90	**93**	95
CL , %	2.0	9.0	16.6	51.1	81	94.2	94.6	95.5
				Section D1003				
Sample size	5	10	15	**34**	60	90	**93**	95
CL , %	2.1	8.3	17.4	50.9	80.6	94.6	95.0	95.2
				Section D1004				
Sample size	5	10	15	**34**	60	90	**93**	95
CL , %	2.1	8.5	17.2	51.2	80.7	94.2	94.9	95.4

Table 3. The confidence level $CL = (1-\alpha)100\%$ versus the sample size n to detect at least 95% ($p \geq 0.95$) of the population (for the circle profiles with **diameter of 200 mm, 300 mm, 400 mm, 500 mm**)

				Section D200				
Sample size	5	10	15	**34**	60	90	**93**	95
CL , %	2.2	7.9	16.9	49.9	80.6	94.2	94.8	95.6
				Section D300				
Sample size	5	10	15	**34**	60	90	**93**	95
CL , %	2.1	8.8	16.5	51.5	80.7	94.7	95.0	95.5
				Section D400				
Sample size	5	10	15	**34**	60	90	**93**	95
CL , %	2.2	8.7	16.5	51.0	81.2	94.5	95.0	95.7
				Section D500				
Sample size	5	10	15	**34**	60	90	**93**	95
CL , %	2.2	8.7	17.2	50.7	81.2	94.4	95.3	95.6

According to the simulation results, the optimal sample size is 94 measured points. This sample size can guarantee at least 95 % of the population content with 95 % confidence level. The minimum 35 measured points can ensure at least 95 % of the

population content with at least 50 % CL. These results are consistent with Wilks criteria and computation results (**Table 1**) based on the analytical approach (12). All sample sizes below 34 measured points have CL less than 50 % for detecting 0.95 population proportion. It should be emphasized that the developed approach does not have any constraints concerning to the type of the distribution and the nominal size. The simulation results presented here and in Paper 2 leads to the conclusion that workpieces with different materials, sizes, produced by other machine processes will provide the similar results.

2.6.2 Maximum estimated error of roundness deviation

The previous approach has a number of limitations regarding to the real practice conditions. It was based on the assumptions of a single circle center and given distribution form of the radius variable. The approach proposed in Paper 4 [12] does not have any of these constraints.

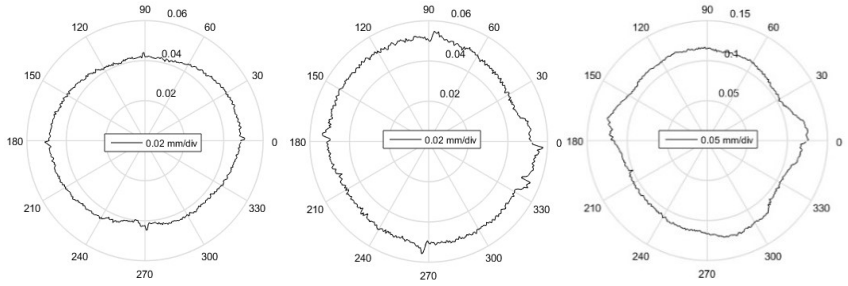
The form of a workpiece profile is not known before it is measured, thus the profile is characterized as nondeterministic. This approach deals with the nondeterministic profiles derived from coordinate measurements of real workpieces. In order to investigate the influence of the sample size, Artificial Neural Network (ANN) was employed to create the continues nondeterministic profiles based on the discrete CMM data. When the relevant factors of the nonlinear correlation are not known and hence have to be assumed, the ANN approach is more preferable than the conventional regression approaches. Besides, the ANN approach is more versatile, and it can be easier adopted to the new profiles generated by other machine processes.

In order to implement the new developed model, 9 holes with various diameters from 40 mm to 500 mm have been milled in a 20 mm aluminium plate. The inspection was performed in a Leitz PMM-C-600 coordinate measuring machine with an analogue probe and *PC-DMIS* software. The middle section of each hole was measured with 480 uniformly distributed points. The least squared circle (LSC) method was applied to calculate the circle centre coordinates (X_c, Y_c) and radius values R_k of each section.

The radius distance R_k from the circle centre (X_c, Y_c) to each individual measured point (X_k, Y_k) was calculated by:

$$R_k = \sqrt{(X_k - X_c)^2 + (Y_k - Y_c)^2} \ . \tag{31}$$

The angle between points is $\varphi_k = 2\pi k/480$, where k is the index number of the points (k = 0…479). Examples of measurements are shown on the polar plots in **Fig. 10**.



(a) (b) (c)

Fig. 10. Examples of measured radial sections: (a) $D_1 = 40$ mm; (b) $D_3 = 100$ mm; (c) $D_9 = 500$ mm.

The back propagation (BP) ANN was applied for profile fitting of the circular features. The theoretical background for design and processing of the BP ANN is given in section 2.5. The multilayered BP ANN developed with MATLAB is illustrated in **Fig. 11**.

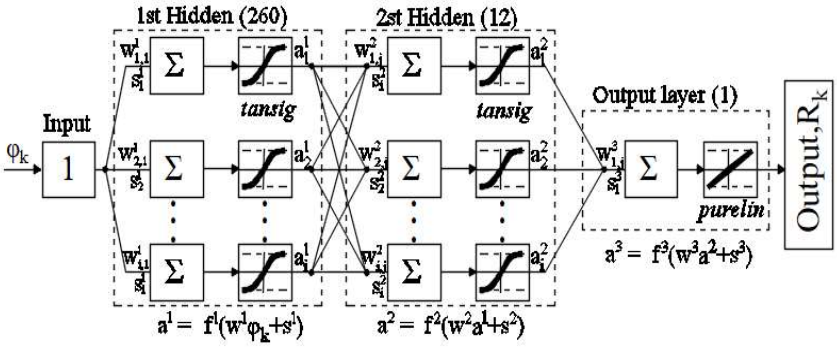


Fig. 11. A 1-260-12-1 feedforward BP ANN for approximation of the measured profiles

The network input is denoted as φ_k and the target as R_k, thus we have a network with one input and one output. In order to achieve a better accuracy of approximation we apply a deep learning strategy in this work. There are two hidden layers, with 260 neurons in the first and 12 neurons in the second layer. The chosen number of layers and neurons is the result of a trial-and-error procedure providing the best-experienced performance. The data set was divided in the following groups: training – 85 % ; validation – 10 %; test – 5 %. The training process was repeated until the maximum absolute error reached a value below the certain level $|\varepsilon_{max}| < 2.5\ \mu m$.

An example of the approximated nondeterministic profile is illustrated in **Fig. 12**. The lowest graph of the figure shows the approximations error in each particular point. The range of the fit errors ε_i is $[-0.9 \cdot 10^{-4}; 0.9 \cdot 10^{4}]$ mm for the presented profile (**Fig. 12**). The nondeterministic profile equivalents to a continuous function

that provides an opportunity to simulate the measuring strategies based on a typical measuring procedure and perfect repeatability conditions.

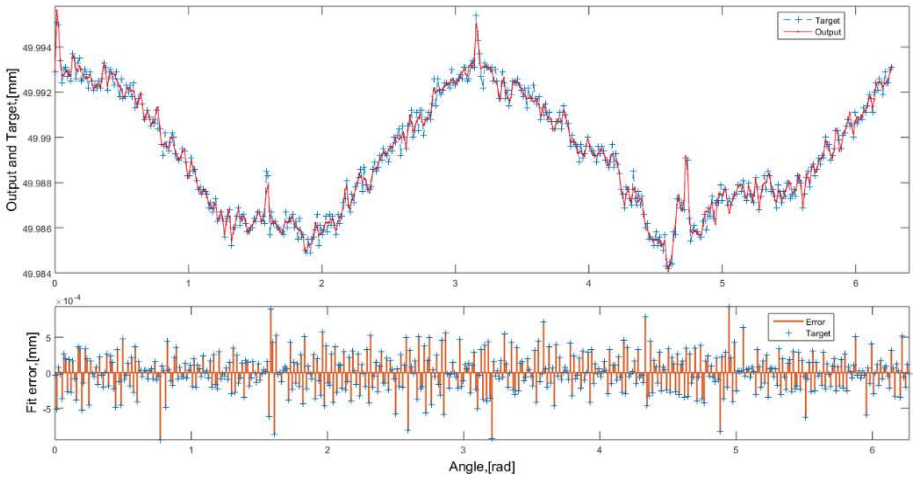


Fig. 12. The continues nondeterministic profile approximated with ANN ($D_3 = 100$ mm)

In order to determine a maximum measuring error due to the sample size, an additional simulation algorithm has been developed. According to common practice in CMM, measuring points are uniformly spaced around a circle profile, and the LSC method is utilized as default. An example of the simulation, using a five-point sample (n = 5), is illustrated in **Fig. 13**. A sample of n equally distributed points is taken from the profile approximated with the designed ANN, and the n-point sample is rotated clockwise with m = 1000 iterations. In each iteration the sample is rotated by the angular step $s = 2\pi/nm$. When the first point p_1 position is defined, the other $(n-1)$ sample points $(p_2, p_3, \dots p_n)$ are determined uniquely with the equal spacing $2\pi/n$.

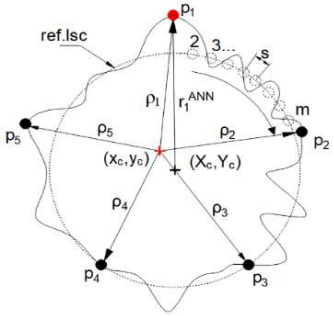


Fig. 13. The five-point sample: $p_2, p_3, \dots p_5$ are the measured points; $\rho_1, \rho_2, \dots \rho_5$ are the estimated radius variables; *ref.lsc* is the reference least squares circle; r_1^{ANN} is the radius of the original reference circle; (X_c, Y_c) is the reference circle center based on 480 points; (x_c, y_c) is the new circle center based on 5 points.

Then, the sample of n radius values r_k^{ANN} $(k = 0, \dots n-1)$ is generated from the trained network. The x_k, y_k coordinates are calculated from radius variables r_k^{ANN} by the following equations:

$$\begin{cases} x_k = X_c + r_k^{ANN} \cos(\varphi_k) \\ y_k = Y_c + r_k^{ANN} \sin(\varphi_k) \end{cases} \tag{32}$$

A new circle centre (x_c, y_c) is calculated with the LSC method to simulate the measuring routine. Then, the radius for each point is calculated with new centre coordinates again by:

$$\rho_k = \sqrt{(x_k - x_c)^2 + (y_k - y_c)^2} \,. \tag{33}$$

The new circle centre (x_c, y_c) and radius values ρ_k were calculated in each iteration. Then the radius variation range of the *n*-sample for each particular location was estimated as $\Delta\rho = \rho_{max} - \rho_{min}$. Eventually, after all iterations were completed, the smallest estimated radius variation range $\Delta\rho_{min}$ for the particular sample size n was defined. The maximum estimation error δ_{max} was calculated as $\delta_{max} = \Delta R^{ANN} - \Delta\rho_{min}$, where $\Delta \mathrm{R}^{ANN} = R_{max}^{ANN} - \mathrm{R}_{min}^{ANN}$ is the precise radius variation range based on 480 variables, which were simulated with the continuous virtual profile.

The simulation procedure described above was applied with different sample sizes from 5 to 400 measuring points and 9 circle sections with nominal diameter from 40 mm to 500 mm. The final simulation results are tabulated in **Table 4**.

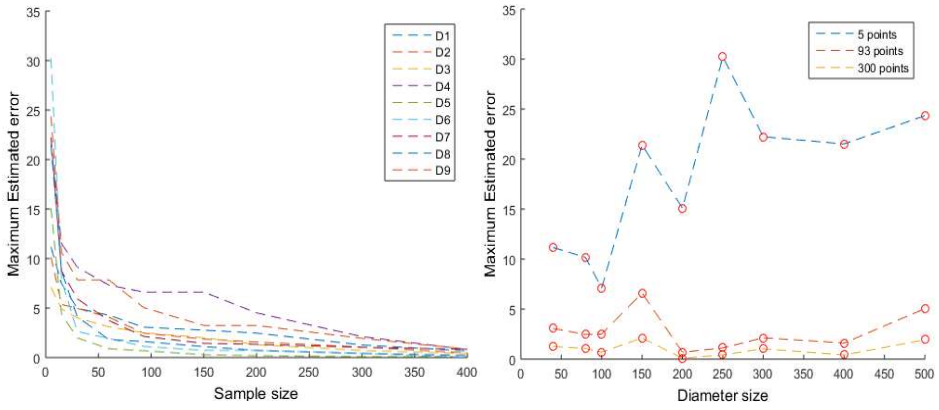
Table 4. The maximum estimated error (δ_{max}) due to the sample size for various diameters

Sample size *n*	**5**	**15**	**30**	**60**	**93**	**150**	**200**	**300**	**400**
D_1 = 40 mm	11.2*	7.5	5.0	4.3	3.1	2.8	2.5	1.3	0.7
D_2 = 80 mm	10.2	5.4	5.0	4.1	2.5	1.9	1.6	1.1	0.5
D_3 = 100 mm	7.1	4.9	4.0	3.1	2.5	2.0	1.3	0.7	0.4
D_4 = 150 mm	21.4	11.5	9.1	7.3	6.6	6.6	4.5	2.1	0.8
D_5 = 200 mm	15.1	4.4	2.0	0.9	0.7	0.3	0.2	0.1	0.1
D_6 = 250 mm	30.2	7.8	2.6	1.9	1.1	0.7	0.7	0.4	0.0
D_7 = 300 mm	22.2	8.7	5.9	3.7	2.1	1.4	1.3	1.0	0.8
D_8 = 400 mm	21.5	8.7	4.1	1.8	1.6	1.1	0.7	0.4	0.2
D_9 = 500 mm	24.3	10.6	7.8	7.8	5.0	3.2	3.2	1.9	0.8

* δ_{max} is given in µm

A plot of the results (see **Fig. 14**a) shows that the relation between the maximum estimated error δ_{max} and the sample size n has nonlinear, asymptotic behavior. This behavior appears relatively predictable. However, the relation between the maximum

estimated error and the diameter size for various sample sizes does not follow a clear trend, as shown in **Fig. 14**b.



(a) (b)

Fig. 14. Simulation of error in roundness assessment. (a): The maximum estimated error vs sample size for various diameters $D_1, D_2, \ldots D_9$; (b): The maximum estimated error δ_{max} vs diameter size D_i for sample sizes with 5, 93, 300 points

The maximum estimated error for a five-point sample (**Fig. 14**b) varies between 7.1 µm and 30.2 µm. The maximum error for ninety-three-point sample does not exceed 6.6 µm, and for three-hundred-point sample size the error does not exceed 2.1 µm. However, for most of the diameters, there is not a substantial improvement of the maximum estimated error with the three-hundred-point sample size relative to the ninety-three points.

According to the simulation results, the error due to the sample size can give a significant contribution to the measurement uncertainty and thus it must be considered in the measuring strategy. The computed values of the maximum estimated errors can be used in calculation of a worst-case scenario in the design and inspection stage.

As shown with the measurements and the simulation, the diameter size is not the main factor for defining the sample strategy. The actual form of the measured workpiece is a more dominant factor than the nominal diameter. Thus, increasing the sample size due to larger diameter is not always necessary.

The presented ANN approach can be adapted to profile forms generated by any machining operations. The approximated nondeterministic profile can be further used as the continuous function for simulations of sample strategies, alignments, filtration methods and measurement uncertainty.

2.6.3 Dimensional (Size) deviation

This section presents an algorithm for evaluation of the effect of sample size in two-point diameter verification of machined features. The algorithm is based on the calculation of minimum sample size presented in section 2.3. According to ISO 14405-1, the dimensions specified on the drawing are defined as two-point size. The following definition applies: "*the two point size is the distance between two opposite points on an extracted integral linear feature of size*" [62]. In Paper 3, we propose a method for optimizing the sample size for diameter verification of cross-sections of cylindrical components. The two-point size of such circular features are also called "*two-point diameter*". The illustration of the two-point diameter is given in **Fig. 15**.

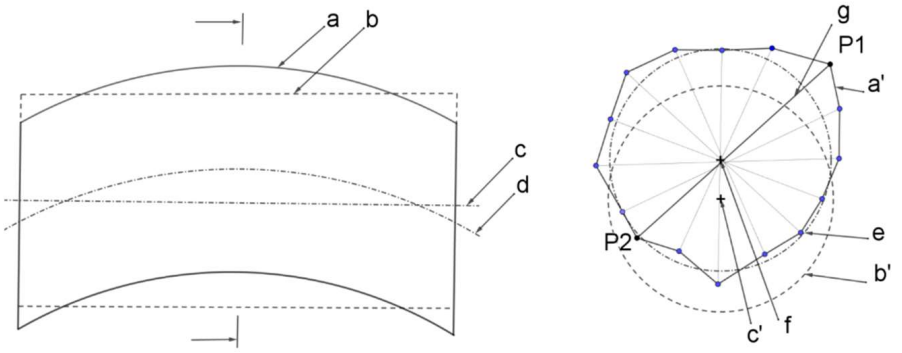


Fig. 15. Two-point diameter: **a** – extracted feature; **a'** – section profile; **b**, **b'** – *LS* associated cylinder; **c**, **c'** – axis of *LS* associated cylinder; **d** – extracted median line; **e** – *LS* associated circle (of section); **f** – *LS* associated circle center of **e**; **g** – actual local size (two-point diameter), the straight line between two opposite points ***P1*** and ***P2***, which goes through the center **f**

In this work, it has been considered the particular case, where the dimensional tolerance interval is larger than the variation of the measurements of the two-point diameter. Three cross-sections *A*, *B* and *C* of an internal cylinder produced by turning operation with nominal diameter of 60 mm were measured by a coordinate measuring machine (CMM). It is assumed that the connecting line between two opposite points includes the associated circle centre. There are 500 measured points in each cross-section, which provides 250 two-point diameters D_i:

$$D_i = \sqrt{(x_i - x_{i+250})^2 + (y_i - y_{i+250})^2} \text{ , with } (i = 1, \dots 250) \tag{34}$$

For convenience of the result presentation and further data processing, the values of D_i has been transformed into $\xi_i = (\overline{D} - D_i)1000$, where $\overline{D}$ is the mean value of the diameter measurements in each cross-section. In order to derive the shape of the probability density function (pdf) $f(\xi)$ of the standardized variable ξ_i, the kernel density estimator (KDE) has been used according to (2) and (5) given in section 2.2.

The estimation results of pdfs $f_A(\xi), f_B(\xi), f_C(\xi)$ for all three sections are shown in **Fig. 16**.

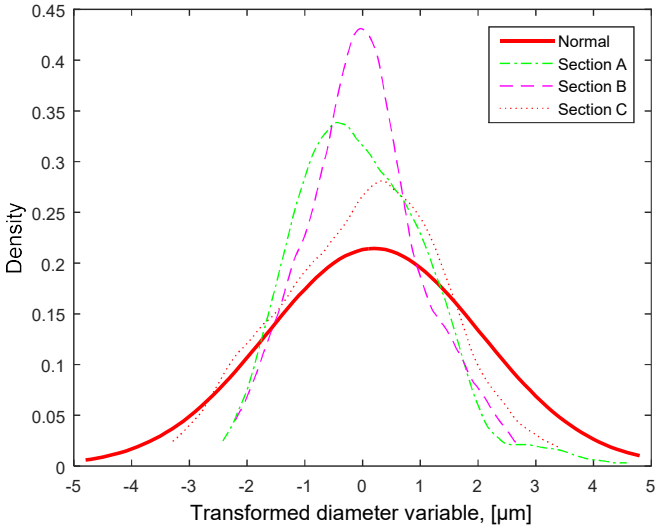


Fig. 16. The pdfs of transformed variables ξ_i estimated by KDE for sections *A*, *B* and *C* and the normal distribution adjusted according to the six-sigma interval

In order to combine the parametric and non-parametric statistics within a new developed model, we need to adjust the standard deviation σ_0 of the reference normal distribution corresponding to the null hypothesis in such way that the six-sigma interval could cover any of the estimated pdf shown in **Fig. 16**. The standard deviation $\sigma_0 =$ 1.6 μm is satisfied for this condition. These estimated pdfs will be further used in the statistical simulation model.

The developed method evaluates whether the given sample size is sufficient or not for two-point diameter verification. We are applying a hypothesis test as illustrated in **Fig. 17**. Examples of distributions of the workpiece diameters are depicted as 5, 6, and 7. As long as we do not know in advanced in which area of the tolerance interval the deviation might be located (upper or lower), then two independent hypothesis tests (denoted as 3 and 4) must be formulated. The mean values μ_0^L and μ_0^U of each corresponding normal distribution $N(\mu_0^L, \sigma_0)$ and $N(\mu_0^U, \sigma_0)$ are located in $3\sigma_0$ distance from the tolerance limits (1 – lower, 2 – upper, **Fig. 17**).

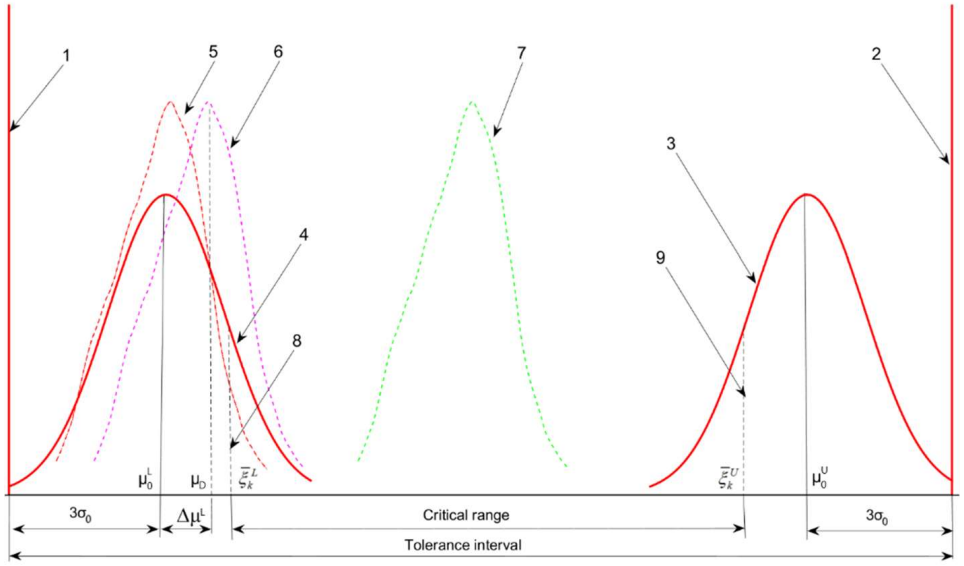


Fig. 17. Hypothesis tests for verification of two-point diameter: 1 – lower tolerance limit; 2 – upper tolerance limit; 3 – upper Gauss (null hypothesis); 4 – lower Gauss (null hypothesis); 5 – KDE object with large deviation; 6 – KDE object with medium deviation; 7 – KDE object with small deviation; 8 – lower boundary of the statistical test; 9 – upper boundary of the statistical test.

The cases when the sample mean is equal to one of the reference means μ_0^L, μ_0^U correspond to the null hypotheses H_0^L or H_0^U. If one of the null hypotheses is not rejected, then a larger sample size will be recommended.

As an example, we consider the dimensional tolerance H7 for nominal diameter 60 mm of an internal cylinder. According to ISO 286-1 the tolerance limits are $EI = 0, ES = +30\ \mu m$. The values of the reference means can be calculated as $\mu_0^L = LTL + 3\sigma_0$ and $\mu_0^U = UTL + 3\sigma_0$ respectively. Finally, the null hypotheses can be formulated as the following $H_0^L: \mu_D = \mu_0^L$ and $H_0^U: \mu_D = \mu_0^U$, where μ_D is the diameter mean estimated as $\bar{\xi} = \frac{1}{n}\sum_{i=1}^{n} \xi_i$. Then the alternative hypotheses are formulated as $H_1^L: \mu_D > \mu_0^L$ and $H_1^U: \mu_D < \mu_0^U$ respectively. The sample means are computed from the sample data generated by random generator based on KDE, which has unknown non-normal distribution $K(\mu_D, \sigma_D)$. The alternative hypotheses H_1^L, H_1^U are accepted, when the sample mean is inside of the critical range D_k. Then, according to (6), the critical range is defined by both, the lower bound:

$$\bar{\xi}_k^L = u_{1-\alpha} \cdot \sqrt{\sigma_0^2 / n} + \mu_0^L, \tag{35}$$

and the upper bound

$$\bar{\xi}_k^U = u_{\alpha} \cdot \sqrt{\sigma_0^2 / n} + \mu_0^U. \tag{36}$$

Where n is the sample size, u_α and $u_{1-\alpha}$ are the quantiles of $N(0,1)$ distribution. The significance level $\alpha = 0.05$ (i.e. $u_{0.05} = -1.645$ and $u_{0.95} = 1.645$) was applied for hypothesis tests. The calculation results of the critical range and boundary values due to the sample size are shown in **Table 5**.

Table 5. The critical range for 30 µm tolerance interval with $\alpha = 0.05, \sigma_0 = 1.6$

Sample size, n	**Lower bound** $\bar{\xi}_k^L$ **, µm**	**Upper bound** $\bar{\xi}_k^U$ **, µm**	**Critical range** D_k **, µm**	**Tolerance reduction,%**
5	6.0	24.0	18.0	40.00
10	5.6	24.4	18.8	37.33
15	5.5	24.5	19.0	36.67
20	5.4	24.6	19.2	36.00
30	5.3	24.7	19.4	35.33
40	5.2	24.8	19.6	34.67
50	5.2	24.8	19.6	34.67
60	5.1	24.9	19.8	34.00

After the initial conditions are specified, we can proceed with the simulation procedure. A number of $N = 10^5$ iterations were used for each sample size of the simulated diameter measurements. In order to choose which of the hypothesis tests must be carried out, we simply check the following condition $\Delta\mu^L < \Delta\mu^U$ (where $\Delta\mu^L = \mu_D - \mu_0^L$ or $\Delta\mu^U = \mu_0^U - \mu_D$). The general principle of the algorithm is shown in **Fig. 18**. In this work, we would like to demonstrate simulation results only for three values of the mean differences $\Delta\mu_1^L = \sigma_0$, $\Delta\mu_2^L = 0.5\sigma_0$ and $\Delta\mu_3^L = 0$. The estimated sample mean $\bar{\xi}$ is compared with either the lower bound $\bar{\xi}_k^L$ or the upper bound $\bar{\xi}_k^U$ (**Table 5**). When the conditions of $\bar{\xi} > \bar{\xi}_k^L$ or $\bar{\xi} > \bar{\xi}_k^U$ are met, the iteration is assigned as 1 (0 otherwise) and summed up as the counters C^L or C^U, then the confidence levels (CLs) η_0^L or η_0^U are calculated as C^L/N or C^U/N respectively. The simulation results for each section *A*, *B* and *C* are presented in **Table 6**, **Table 7** and **Table 8**. The results for the opposite side *UTL* are similar, and they are not presented here.

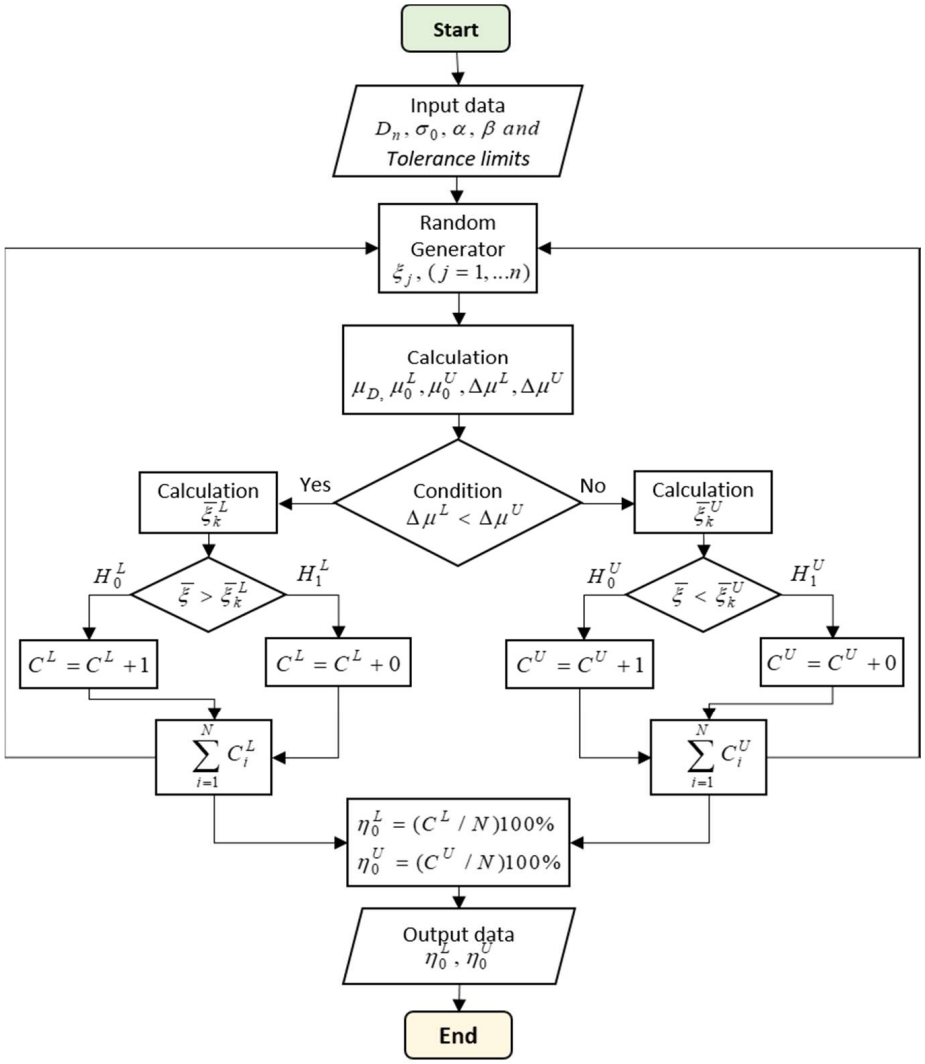


Fig. 18. The algorithm for evaluation of the sample size for two-point diameter verification

Table 6. Estimation of the CL η_0^L for various sample sizes and different $\Delta\mu^L$, for ***Section A***

Mean difference	Sample size 5	10	15	20	30	40	50
$\Delta\mu_1^L = \sigma_0$	0.78	**0.98**	1	1	1	1	1
$\Delta\mu_2^L = 0.5\sigma_0$	0.22	0.46	0.65	0.79	0.93	0.98	1
$\Delta\mu_3^L = 0$	0.02	0.02	0.02	0.02	0.02	0.02	0.02

Table 7. Estimation of the CL η_0^L for various sample sizes and different $\Delta\mu^L$, for ***Section B***

Mean difference	Sample size 5	10	15	20	30	40	50
$\Delta\mu_1^L = \sigma_0$	0.82	**0.99**	1	1	1	1	1
$\Delta\mu_2^L = 0.5\sigma_0$	0.21	0.46	0.66	0.82	0.95	0.99	1
$\Delta\mu_3^L = 0$	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 8. Estimation of the CL η_0^L for various sample sizes and different $\Delta\mu^L$, for ***Section C***

Mean difference	Sample size 5	10	15	20	30	40	50
$\Delta\mu_1^L = \sigma_0$	0.75	0.95	0.99	1.00	1.00	1.00	1.00
$\Delta\mu_2^L = 0.5\sigma_0$	0.29	0.47	0.63	0.75	0.88	0.95	0.98
$\Delta\mu_3^L = 0$	0.03	0.03	0.03	0.03	0.03	0.03	0.03

From the simulations, we get three categories of results. The first category *G* (good parts) corresponds to the intersection of two subsets $\{\mu_D \geq \mu_0^L + \sigma_0\} \cap \{\mu_D \leq \mu_0^U - \sigma_0\}$. The second category *T* (not confirmed) includes the subsets $\{\mu_0^L < \mu_D < \mu_0^L + \sigma_0\}$ and $\{\mu_0^U - \sigma_0 < \mu_D < \mu_0^U\}$, and the third category *S* (suspected parts) belongs to $\{\mu_D \leq \mu_0^L\}$ and $\{\mu_D \geq \mu_0^U\}$. In addition, we have parts which are definitely out of tolerances (*F*, fail parts), $\{\mu_D < LTL\}$ and $\{\mu_D > UTL\}$. For illustrational purposes, we presume a uniform distribution *U(0,30)* of the manufacturing process over a long time period. The tolerance interval is illustrated in **Fig. 19**. The content of each region (*S*, *T*, and *G*) can be easily evaluated:

$$\int_0^{3\sigma_0} \frac{1}{30} du = 0.16, \quad \int_{3\sigma_0}^{4\sigma_0} \frac{1}{30} du = 0.05, \quad \int_{\mu_0^L+\sigma_0}^{\mu_0^U-\sigma_0} \frac{1}{30} du = 0.57 \qquad (37, 38, 39)$$

In the S category ($\mu_3^L = 0$), the correct decision can be taken with at least 95 % probability ($1 - \eta_0^L$) with the five-observation sample, according to **Table 6**, **Table 7** and **Table 8**.

Similar for the *G* category, the correct decision can be taken with at least 95 % probability (η_0^L) with the ten-observation sample for $\mu_1^L = \sigma_0$.

According to **Table 6**, **Table 7** and **Table 8** for the category *T* ($\mu_2^L = 0.5 \cdot \sigma_0$), more than 40 observations might be required to confirm the compliance of the diame-

ter measurements with the tolerance limits, with 95 % confidence level. Nevertheless, the *T* category is only 10 % of whole area of the uniform distribution (**Fig. 19**) according to (38).

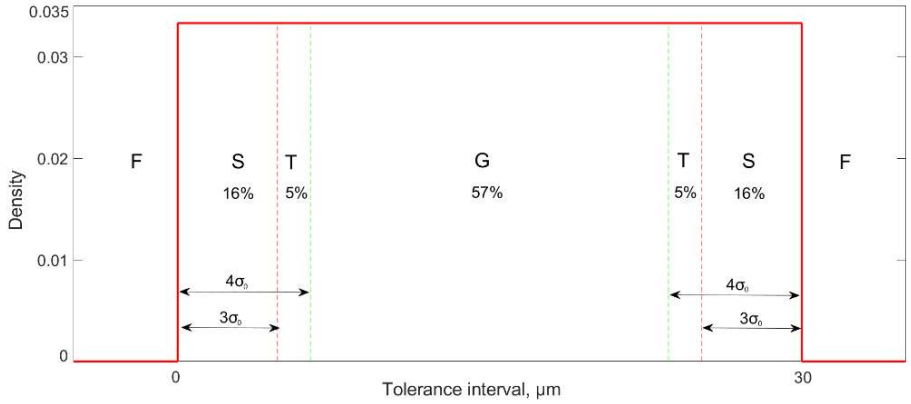


Fig. 19. The dimension tolerance interval based on the uniform distribution assumption.

The presented method reduces the tolerance interval, but this is compensated by the significant minimization of the sample size. By using the ten-observation sample (corresponding to 20 measuring points for the two-point diameter), we are able to make the right decision in about 95 % of the cases.

3 Outlier detection

3.1 Outlier detection methods

A great number of methods for outlier detection are available today [21, 63]. It is always advisable to use a combination of graphical [64] and analytical approaches [47, 48] for investigation of outliers in data sets. There are also hybrid methods based on both analytical and graphical interpretations, which are very efficient for detection of outliers. The examples of such methods are kernel density estimator [43] (described recently above in section 2.2) and box plot [65] (will be given later in section 3.2.1). In this chapter, two analytical methods based on parametric statistics [13] and the hypothesis test (described in previous section 2.3) are presented. These methods must meet the following conditions:

- flexibility to the sample size (finite or large)
- more than one outlier can be handled
- not strictly constrained to a specific statistical distributions of the data set
- robustness to masking and swapping effects

The *masking* and *swamping effects* can occur during the data analysis with parametric statistical test. The *masking effect* can happen when too few outliers are specified in the outlier detection procedure. Then, the test performance can be influenced by the other outliers and the result is that no outliers will be detected. On the other hand, if too many outliers are specified in the parameters of outlier test, then some valid observations can be incorrectly labelled as outliers, which is the *swamping effect*.

Paper 2 deals with an investigation of the outlier detections procedures based on both graphical and analytical approaches. The theoretical background of two analytical methods is given in next two sections. The contributions based on these approaches are presented further in section 3.2.

3.1.1 Grubbs method

The Grubbs method can detect a single outlier in a normally distributed data set. In order to detect more than one outlier, Grubbs method must be sequentially applied on data sample [66, 47]. The outlier detecting procedure tests two types of hypotheses:

null hypothesis H_0 is no outliers in the data set, the alternative hypothesis H_1 is that there is one outlier in the sample. The test statistic for the two-sided case is

$$G = \frac{\max\left|\rho_i - \bar{\rho}\right|}{s}, \tag{40}$$

where ρ_i is the individual observations, $\bar{\rho}$ is the sample mean, and s is the standard deviation. The following conditions must be checked for two-sided case:

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{\left(t_{\left(\frac{\alpha}{2n},n-2\right)}\right)^2}{n-2+\left(t_{\left(\frac{\alpha}{2n},n-2\right)}\right)^2}}, \tag{41}$$

where $t_{\left(\frac{\alpha}{2n},n-2\right)}$ is the Student's quantile given at probability $\frac{\alpha}{2n}$ with $n-2$ degrees of freedom in the data set and *n* sample size. The null hypothesis is rejected if the condition (41) is satisfied. The significance level α is the probability of rejection of a true H_0, i.e. the *type I error*. The test result might be affected by the masking effect.

3.1.2 Rosner method

The original name of this method is Generalized Extreme Studentized Deviate (GESD) procedure, proposed by Rosner [48]. The method was developed to detect an unknown number of outliers. The only parameter required is an upper number *m* of expected outliers (to avoid the masking effect, the number should not be too small).

The method tests two types of hypotheses: null hypothesis H_0 – no outliers in the sample, alternative hypothesis H_1 – the sample has up to *m* outliers. The test statistic for two-sided case is [48]:

$$R_i = \frac{\max_i\left|\rho_i - \bar{\rho}\right|}{s}. \tag{42}$$

The observation that maximizes $|\rho_i - \bar{\rho}|$ is removed, and then the statistic R_i is recomputed with $n-1$ observations. The process is repeated until *m* observations have been removed. The result of the computation will be an array of $R_1, R_2, \ldots R_m$. Then for each single element of the array R_i, the critical value k_i for the two-sided case is calculated:

$$k_i = \frac{t_{(p,n-i-1)}\left(n-i\right)}{\sqrt{\left(n-i-1+\left(t_{(p,n-i-1)}\right)^2\right)\left(n-i+1\right)}},\ i = 1,2,\ldots m\ . \tag{43}$$

Where $t_{(p,n-i-1)}$ is the quantile of Student's distribution with $(p, n-i-1)$ degrees of freedom and probability value $p = 1 - \frac{\alpha/2}{n-i+1}$. Thus the total number of outliers is

the largest i such that $R_i > k_i$. The GESD test was created to reduce the masking effect.

3.2 Contributions to objective 2

The research questions and objectives related to the data outlier detection have been formulated in section 1.2.2 and 1.3.1, respectively. The stated questions are relevant for those who deal with measuring data processing and analyses. This contribution provides a criteria for assessment of suspected measurements in the data sample. The presence of outliers can significantly affect the result of data analysis such as data statistics, computation of substitute geometrical elements and data filtration. Therefore, the detection of outliers has a decisive role for the GD&T inspection.

In CMM inspection, outliers are not necessarily erroneous measurements. The presence of outliers can point out that an additional investigation might be required. A combination of both graphical and analytical approaches provides the best investigation result.

3.2.1 Graphical approach

It is always recommended to look at the graphical interpretation of the data before any analytical method is applied. There are many options available such as histograms, scatter diagrams, dot plots, etc. [64]. There is also hybrid solution such as box plots, which is based both on the analytical model and the graphical interpretation [65]. Analyses of the data with the box plot helps to decide which method is the most appropriate one (for single or multiple outliers), if any outliers are presented. An example of a box plot is shown on **Fig. 20**.

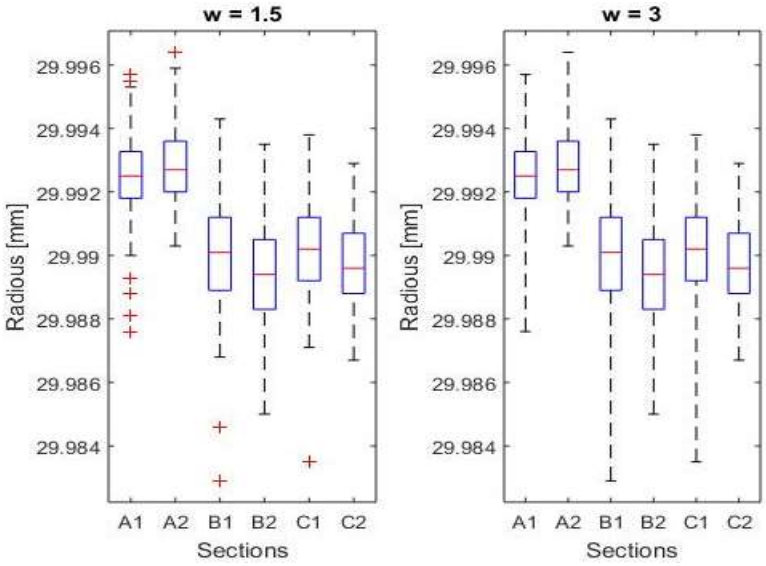


Fig. 20. The six data sets with outliers (*A1, B1, C1*) and without (*A2, B2, C2*)

Six data sets were derived from CMM measurements of sections of an internal cylinder, with 475 observations in each set. The data with outliers is denoted as *A1, B1,* and *C1,* and the data sets without outliers as *A2, B2,* and *C2*. (In section *A2* one outlier has not been removed.) The section *A1* represents the measured sample with multiple potential outliers. The section *B1* has two potential outliers, and the section *C1* has only one potential outlier. The influence of outliers on statistics (sample mean, median, skewness, IQR – interquartile range, etc.) are indicated by the box plots. The IQR of data, the different between lower and upper quartiles, is measured along the vertical axis (Radius value) on the diagram.

The analytical part of the box plot is expressed as the following:

$$LF = q_1 - w(q_3 - q_1)\,, \tag{44}$$

where *LF* is the *lower fence*, q_1, q_3 are the first and the third quartiles of data sample, and *w* is the *significant factor*. And for *upper fence UF* [67]:

$$UF = q_3 + w(q_3 - q_1). \tag{45}$$

The left part of **Fig. 20** corresponds to significant factor $w = 1.5$. The extreme observations are indicated by the red dots, and they can be classified as *suspected outliers*. The right side of **Fig. 20** corresponds to the significant factor $w = 3$, which may indicate *extreme outliers.* Thus, the extreme outliers are not present in the data set.

The coordinates of the reference circle center (x_c, y_c) of each circle section were computed by the least square (LS) method based on 475 measured points. The radius variable r_i for each measured point was calculated according to (28). For further data processing and simulation, a standardized radius was used:

$$\rho_i = (\overline{r} - r_i)1000\,, \tag{46}$$

where $\bar{r}$ is the average value of radius. The probability density function (pdf) $f(\rho)$ was estimated with the kernel density estimator [43] according to (2) and (5) for $N = 475$ observations. The estimation results of pdf for sections *A, B, C* are shown on **Fig. 21** and **Fig. 22**. As can be observed from the figures, the variables are not normally distributed.

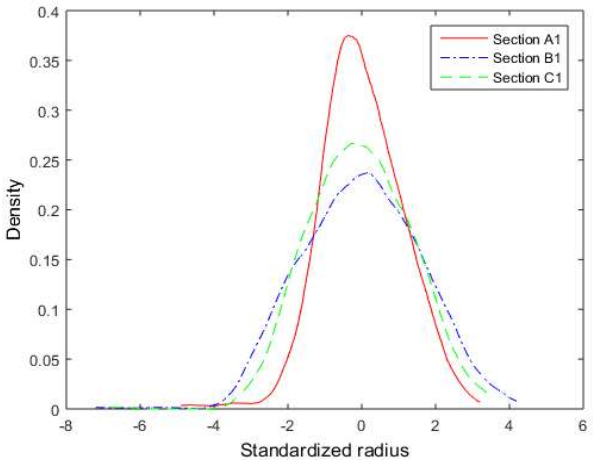


Fig. 21. Estimated pdf $\hat{f}_{A1}(\rho), \hat{f}_{B1}(\rho), \hat{f}_{C1}(\rho)$ with outliers

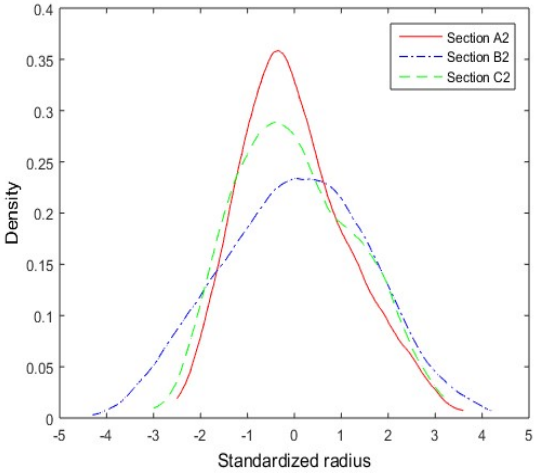


Fig. 22. Estimated pdf $\hat{f}_{A2}(\rho), \hat{f}_{B2}(\rho), \hat{f}_{C2}(\rho)$ without outliers

3.2.2 Analytical approaches

Several parametric tests for normality of distributions are available e.g. [68, 69]. The normality test of Andersen Darling method applied with data sets *A2, B2, C2* gives the following p-values: 0.001, 0.138, 0.001. Only section *B2* has a significant p-value (> 0.05), and the normal distribution can be assumed.

By knowing all that, we can start selecting a method. Many methods for detection of outliers are available today [63]. Practically, the existing methods differ from each other with respect to the following properties:

- the sample size for which the method is applicable
- the assumed distribution of the population
- detection of single or multiple outliers
- detection of an exact number of outliers, or an upper bound

According to these conditions, two of the most suitable methods such as Grubbs and Rosner were chosen for our study. The methods are recommended by ISO 16269-4, ISO 5725-2 [21, 66]. Both methods are based on estimation of the deviation from the sample mean and both have an assumption about the normal distribution. The strictness of this assumption and a resistance to the masking / swamping effects has been examined in the Paper 2 [13].

3.2.3 Simulation of outlier procedures

The estimators of pdf $\widehat{f}_{A2}(\rho), \widehat{f}_{B2}(\rho), \widehat{f}_{C2}(\rho)$ based on the data without outliers, are used as a basis for simulation of outlier detection. The estimated standard deviations for the data sets are: $S_A = 1.15, S_B = 1.54, S_C = 1.27$. We have defined two cases for outliers, *medium* and *large*, based on the standard deviation of the data sets and uniform distribution. The medium and large values for the outliers are generated in the intervals

$$\rho_m \in [3.9s - 0.01, 3.9s + 0.01] \cup [-3.9s - 0.01, -3.9s + 0.01]$$

and

$$\rho_l \in [4.5s - 0.1, 4.5s + 0.1] \cup [-4.5s - 0.1, -4.5s + 0.1]$$

respectively. The successfulness of Rosner and Grubbs methods was estimated with 10^5 iterations by summing up two possible results: $S = S + 0$ as a failure, or $S = S + 1$ as a success. If exactly the same number of outliers with the same indexes are detected then such iteration is assigned by 1 (success). On the other hand, if only $m-1$ or less from a total number m of outliers were detected correctly, then the iteration is assigned by 0 (fail). The efficiencies of Grubbs method e_G and Rosner method e_R were estimated simultaneously as a rate of the sum of successful iterations S to the total iteration number M: $e_G = S_G/M$ and $e_R = S_R/M$.

Several modifications of outlier simulations have been developed in this work. All simulation procedures were carried out in MATLAB and results were tabulated in **Table 9**, **Table 10**, **Table 11**, and **Table 12**. The simulation results were rounded to two decimals. The significance level $\alpha = 0.05$ was applied in all tests. The following

factors and conditions were used in the simulation to evaluate the efficiency of the Grubbs and the Rosner methods:

- non-normal distribution of randomly generated data sample with 100 observations were applied (**Table 9**, **Table 10**, and **Table 11**);
- various number of outliers (from 1 to 4) were randomly distributed around a mean value of data with specified deviation values (**Table 9** and **Table 10**);
- various number of outliers (from 2 to 4) were integrated as a block with deviation values corresponding to the large value (**Table 11**);
- two outliers with random locations and large values were randomly integrated into data of various sample sizes of 15, 30, 60 and 100 observations (**Table 12**).

Table 9. Efficiency rates e_G, e_R of outlier detection methods (randomly located outliers, ***medium values***, 100 observations)

Number of outliers	Method	Section A $\bar{\rho}_m^A = 4.49$	Section B $\bar{\rho}_m^B = 6.00$	Section C $\bar{\rho}_m^C = 4.94$
1	Grubbs	0.66	0.65	0.67
	Rosner	0.66	0.65	0.67
2	Grubbs	0.36	0.33	0.34
	Rosner	0.60	0.58	0.61
3	Grubbs	0.12	0.09	0.09
	Rosner	0.58	0.57	0.58
4	Grubbs	0.02	0.01	0.01
	Rosner	0.57	0.55	0.57

Table 10. Efficiency rates e_G, e_R of outlier detection methods (randomly located outliers, ***large values***, 100 observations)

Number of outliers	Method	Section A $\bar{\rho}_l^A = 5.2$	Section B $\bar{\rho}_l^B = 6.9$	Section C $\bar{\rho}_l^C = 5.7$
1	Grubbs	0.99	0.99	1.00
	Rosner	0.99	0.99	1.00
2	Grubbs	0.95	0.94	0.96
	Rosner	0.99	0.99	1.00
3	Grubbs	0.77	0.72	0.76
	Rosner	0.99	0.99	0.99
4	Grubbs	0.37	0.32	0.34
	Rosner	0.99	0.99	0.99

The simulation procedure code corresponding to **Table 9** (similar as for **Table 10**) is shown in **Fig. 23**.

```
clear;
clc;
                        %   Simulation regarding to   %
                        %   Random Outlier locations  %
%
% Data
load stdataN % measured data (standartized), sections [A2, B2, C2] without outliers
pd = fitdist(stdataN(:,1),'Kernel','Kernel','epanechnikov'); % estimated pdf
s_data = std(stdataN(:,1)); % Data estimated Standard deviation
Ov = round(3.9*s_data,2);   % Medium Value of outlier
%Ov = round(4.5*s_data,1);  % Large Value of outlier
M = 10000;                  % Number of iterations
N = 100;                    % Sample size
alpha = 0.05;               % Significance level
k=1;                        % Table index
sec_GR = zeros(8,1);        % Initilizing of column vector for results
for On = 1:4        % Number of outliers
flag_esd = 0;       % Number of success for Rosner
flag_grabbs = 0;    % Number of success for Grubbs
    for j = 1:M
    x = random(pd,1,N);     % Random generator based on Kernel distribution
    Oi = randperm(N,On);    % Sample indexes replaced by outliers
    % Integrate outliers into data sample
    x(Oi) = (-1).^unidrnd(2,1,On).*random(makedist('Uniform','lower',Ov-0.01,'upper',Ov+0.01),1,On);
    [~,~,i_esd] = g_esd(x, alpha, On);                  % Determine outlier indexes with Rosner
    [~,i_grabbs,~] = grubbs(x, alpha);                  % Determine outlier indexes with Grubbs
    flag_esd=flag_esd+min(ismember(Oi,i_esd));          % counting numbers of success for Rosner
    flag_grabbs=flag_grabbs+min(ismember(Oi,i_grabbs)); % counting numbers of success for Grubbs
    end
% Simulation results, column A, B, C for Table format
sec_GR(k,1) = flag_grabbs/M;
sec_GR(k+1,1) = flag_esd/M;
k = k+2;
end
```

Fig. 23. Code of the simulation procedure for **Table 9**

For the simulation with locations of outliers as a block, only negative large values corresponding to the interval $\rho_l \in [-4.5s - 0.1, -4.5s + 0.1]$ were applied. Due to low skewness of data (**Fig. 22**) the simulation results for positive values were similar, and they are not presented in the report. The simulation results are given in **Table 11**.

Table 11. Efficiency rates e_G, e_R of outlier detection methods (***located as a block***, large values, 100 observations)

Number of outliers	Method	Section A $\overline{\rho}_l^A = 5.2$	Section B $\overline{\rho}_l^B = 6.9$	Section C $\overline{\rho}_l^C = 5.7$
2	Grubbs	0.95	0.92	0.96
	Rosner	1.00	1	1
3	Grubbs	0.6	0.52	0.55
	Rosner	1.00	0.99	1
4	Grubbs	0.07	0.07	0.06
	Rosner	1.00	0.99	1

The code for outlier detection procedures corresponding to **Table 11** is shown in **Fig. 24**.

```
clear;
clc;
                %    Simulation regarding to        %
                %    Outlier locations as a Block   %
%
% Data
load stdataN % measured data (standartized), sections [A2, B2, C2] without outliers
pd = fitdist(stdataN(:,2),'Kernel','Kernel','epanechnikov'); % Estimated pdf
s_data = std(stdataN(:,2)); % Data estimated Standard deviation
Ov = round(4.5*s_data,1);   % Large Value of outlier
M = 10000;                  % Number of iterations
N = 100;                    % Sample size
alpha = 0.05;               % Significance level
k=1;                        % Indexing for Table
sec_GR = zeros(6,1);        % Initilizing of column vector for results
for On = 2:4        % Number of outliers
flag_esd = 0;       % Number of success for Rosner
flag_grabbs = 0;    % Number of success for Grubbs
    for j = 1:M
    x = random(pd,1,N);         % Random generator based on Kernel distribution
    first = randperm(N,1);      % First indexe replaced by outlier
    Oi = first:1:first+On-1;    % Rest Block Outliers
    % Integrate outliers into data sample
    x(Oi) = (-1).*random(makedist('Uniform','lower',Ov-0.1,'upper',Ov+0.1),1,On);
    [~,~,i_esd] = g_esd(x, alpha, On);                  % Determine outlier indexes with Rosner
    [~,i_grabbs,~] = grubbs(x, alpha);                  % Determine outlier indexes with Grubbs
    flag_esd=flag_esd+min(ismember(Oi,i_esd));          % counting numbers of success for Rosner
    flag_grabbs=flag_grabbs+min(ismember(Oi,i_grabbs)); % counting numbers of success for Grubbs
    end
% Simulation results, column A, B, C for Table format
sec_GR(k,1) = flag_grabbs/M;
sec_GR(k+1,1) = flag_esd/M;
k = k+2;
end
```

Fig. 24. Code of the simulation procedure for location of outliers as a block (**Table 11**)

Finally, data of various sample sizes from 15 to 100 observations were randomly generated based on Kernel pdf estimators $\widehat{f}_{A2}(\rho), \widehat{f}_{B2}(\rho), \widehat{f}_{C2}(\rho)$. In order to equally test both methods, a low number of outliers were chosen (two outliers with large values). The outliers were integrated into data samples with random locations within intervals: $\rho_l \in [4.5s - 0.1, 4.5s + 0.1] \cup [-4.5s - 0.1, -4.5s + 0.1]$. The simulation results are given in **Table 12**.

Table 12. Efficiency rates e_G, e_R of outlier detection methods for various sample sizes (two outliers with random locations and large values)

Sample size	Method	Section A $\overline{\rho}_l^A = 5.2$	Section B $\overline{\rho}_l^B = 6.9$	Section C $\overline{\rho}_l^C = 5.7$
15	Grubbs	0.07	0.06	0.06
	Rosner	0.75	0.74	0.76
30	Grubbs	0.44	0.42	0.42
	Rosner	0.92	0.92	0.94
60	Grubbs	0.84	0.83	0.85
	Rosner	0.98	0.98	0.99
100	Grubbs	0.95	0.94	0.96
	Rosner	0.99	0.99	1.00

The code for outlier detection procedures corresponding to **Table 12** is illustrated in **Fig. 25**.

```
clear;
clc;
          %     Simulation regarding to                %
          %     Two Outliers and various sample sizes  %
%
% Data
load stdataN % measured data (standartized), sections [A2, B2, C2] without outliers
pd = fitdist(stdataN(:,3),'Kernel','Kernel','epanechnikov'); % estimated pdf
s_data = std(stdataN(:,3)); % Data estimated Standard deviation
Ov = round(4.5*s_data,1);   % Large Value of outlier
M = 10000;                  % Number of iterations
Nsize = [15 30 60 100];     % Various Sample sizes
On = 2;                     % Number of outliers
alpha = 0.05;               % Significance level
k=1;                        % indexing for Table
sec_GR = zeros(8,1);        % Initilizing of column vector for results
for s = 1:4
N = Nsize(s);
flag_esd = 0;       % Number of success for Rosner
flag_grabbs = 0;    % Number of success for Grubbs
    for j = 1:M
    x = random(pd,1,N);     % Random generator based on Kernel distribution
    Oi = randperm(N,On);    % Sample indexes replaced by outliers
    % Integrate outliers into data sample
    x(Oi) = (-1).^unidrnd(2,1,On).*random(makedist('Uniform','lower',Ov-0.01,'upper',Ov+0.01),1,On);
    [~,~,i_esd] = g_esd(x, alpha, On);                % Determine outlier indexes with Rosner
    [~,i_grabbs,~] = grubbs(x, alpha);                % Determine outlier indexes with Grubbs
    flag_esd=flag_esd+min(ismember(Oi,i_esd));        % Counting numbers of success for Rosner
    flag_grabbs=flag_grabbs+min(ismember(Oi,i_grabbs)); % Counting numbers of success for Grubbs
    end
% sprintf('Rosner success rate %.3f, Grubbs success rate %.3f',flag_esd/M, flag_grabbs/M)
% Simulation results, column A, B, C for Table format
sec_GR(k,1) = flag_grabbs/M;
sec_GR(k+1,1) = flag_esd/M;
k = k+2;
end
```

Fig. 25. Code of the simulation procedure for outliers integrated into data of various sample sizes

3.2.4 Implementation

The Rosner and Grubbs methods were also applied with the data set *A1, B1, C1* with outliers. The result was that two outliers were removed in section *A1* and one outlier was removed in sections *B1* and *C1*. Some outliers (medium) were not detected by any of the methods e.g. sections *A1, B1* though some of these suspected points disappeared after measurements were repeated e.g. section *A2* (see **Fig. 26**, left). The implementation of methods is shown in **Table 13**.

Table 13. Implementation of outlier detection methods with real measurement data based on 475 observations

Outliers no	Outlier parameters	Section A1*		Section B1*		Section C1*	
		Rosner	Grubbs	Rosner	Grubbs	Rosner	Grubbs
1	**Index**	60	60	459	459	2	2
	Values [mm]	29.9881	29.9881	29.9829	29.9829	29.9835	29.9835
2	**Index**	61	61	-	-	-	-
	Values [mm]	29.9876	29.9876	-	-	-	-

The implementation results are quite consistent with the simulation. A comparison of data in sections *A1*, B1*, C1** (after outliers were removed by the analytical methods) with repeated measurements of the same sections *A2, B2, C2* is given in **Fig. 26**.

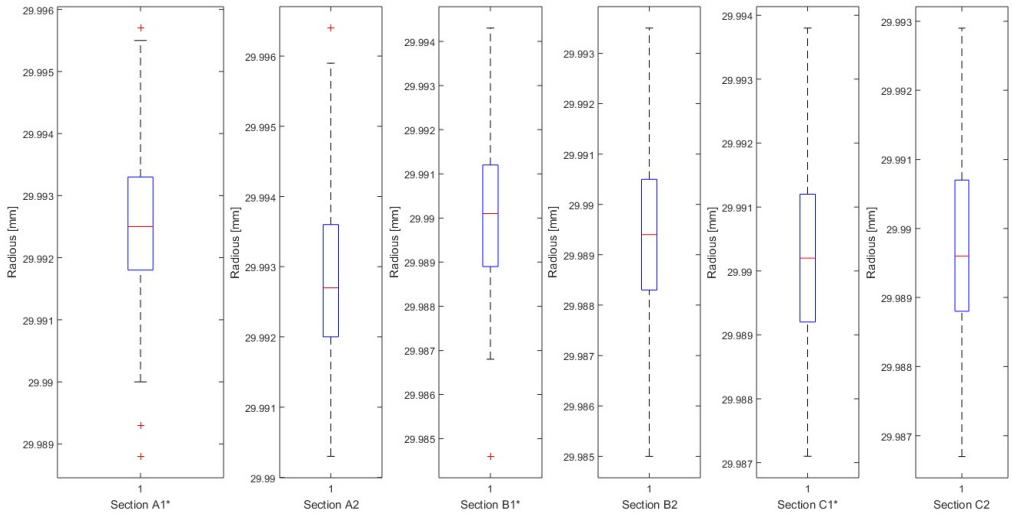


Fig. 26. Boxplot of the measurement data: *A1*, B1*, C1** - after outliers removed by analytical methods; *A2, B2, C2* – after repeated measurements

According to the simulation and implementation, analytical methods are efficient for the large value outliers but less efficient for medium value outliers. Regarding to the results given in **Table 9** and **Table 10**, the Grubbs method has a low efficiency rate

e_G relative to the Rosner method. The Rosner method is to a very small degree affected by the masking effect, even for higher number of outliers in the data sample. The Rosner method is also stable to the block location of outliers, what is not the situation for the Grubbs method. The Rosner method has still efficiency rate e_R about 0.75 for the fifteen-observation sample size while the Grubbs method has efficiency rate e_G below 0.45 for the thirteen-observation sample. Any notable difference in the method efficiencies for the different sections (A, B, C) was not observed. Relaying on the simulation results, the swamping effect was not observed.

The following results were observed with the Rosner (GESD) method in this study:

- ability to work with various sample size including the small one (15 observations)
- ability to detect multi outliers when maximum outlier number is unknown for both the random and block locations of outlier group
- efficiency above 90 % to detect large value outliers, which bring the most significance influence on the data analyses
- stability to the masking and swamping effects

The experimental measurements with CMM also revealed that groups of multiple outliers can be expected at GD&T inspection. The Rosner (GESD) outlier detection procedure can be integrated into software application for coordinate measurements.

4 Substitute Elements (Minimum Volume Bounding Box)

4.1 Overview of MVBB problem

Measurement data is usually provided by the CMM as Cartesian coordinates (x, y, z). In order to evaluate geometrical deviations, the coordinate points must be approximated by some basic substitute (reference) elements (e.g. lines, planes, circles, cylinders, etc.). The method used for defining a substitute element may have significant influence on the measurement result.

In this chapter, we discuss the circumscribing substitute methods for objects with *rectangular parallelepiped* form, which is relevant in calibration of standards for geometrical metrology. It was suggested by Dupuis [70] to use the term *cuboid* when referring to a rectangular parallelepiped. However, in the literature of the computational geometry, the term *box* is commonly associated with the rectangular parallelepiped. In this text, we use the term *side* for the bounding box face. This term may be also used while referring to the physical cuboid object side. The term *face* is mainly used for the inscribed convex polyhedron faces, which are the product of 3D convex hull operation. All six sides (faces) of the box are rectangles and each side is parallel with the opposite side and orthogonal with the other four *adjacent sides*. These four *adjacent sides* comprise a "*closed loop*". For example, the *Top* side has a "*closed loop*" of *adjacent sides* that consists of: *Front, Left, Right, and Back*. The opposite, *Bottom* side has the same "*closed loop*" of *adjacent sides* as the *Top* side.

Together with other association criteria (e.g. minimum zone, least squares), the minimum volume criterion can be applied for estimation of the flatness deviation of mechanical parts in industry [71]. A minimum volume bounding box (MVBB) is an alternative principle for assessment of objects with the cuboid form. An estimation of the MVBB often includes an estimation of minimal area bounding rectangle (MABR). There are many applications in computer graphics, image processing, medicine, metrology, etc. based on these methods.

Based on the proposals of Shamos [72] and Freeman and Shapira [73], Toussaint presented an elegant unambiguous MABR solution in [74]. This exact solution of the MABR problem has $O(n^2)$ computing time with the use of the rotating caliper algorithm for n-point set in $\mathbb{R}^2$, and $O(n)$ time with the use of two pairs of rotating calipers orthogonal to each other. A number of approximation algorithms and heuristic alternatives are suggested to solve the two-dimension problem. Among them, the

searching algorithms based on the R-tree data structures [75-77] and the principle components [78, 79].

The most exact solution of the MVBB problem for n-point set in $\mathbb{R}^3$ with computation time $O(n^3)$ was provided by O'Rourke [80], which remains the state-of-the-art so far. Alternative approximation algorithms have been developed to reduce the computation time. Bespamyatnikh and Segal [81] suggested an efficient $O(n^2)$ approximation algorithm. A search based on Powell's quadratic convergent method was proposed by Lahanas et al. [82]. Later, Barequet and Har-Peled [83] presented an approximating algorithm with $O(n + 1/\varepsilon^{4.5})$ computation time, and a simplified version with $O(n \log n + n/\varepsilon^3)$, where $0 < \varepsilon \leq 1$. Recently, Dimitrov et al. developed a faster algorithm based on the discrete and the continuous versions of principal component analysis (PCA) [79, 84]. The continuous version guarantees a constant approximation factor but it is still limited by $O(n \log n)$ – time required for computation of a convex hull. The commonly used solutions for MABR and MVBB are based on the convex hull operation [72, 85] in order to reduce the number of considered points and avoid redundant computation.

4.1.1 Minimum-Area Bounding Rectangle

As it was mentioned already, the solution of the three-dimension MVBB problem involves the two-dimension case. After the orientation of one side of the bounding box is locked in the MVBB algorithm, all points are projected onto this plane, and the orientations of other (closed-loop) adjacent sides of the bounding box can be found by the MABR algorithm as the two-dimension problem.

The earliest known solution of the MABR problem was presented by Freeman and Shapira [73]. They presented and proved the following theorem, which is the basis for minimum bounding rectangle algorithms: *The rectangle of minimum area enclosing a convex polygon has a side collinear with one of the edges of the polygon.*

The MABR solution is based on the 2D convex hull operation [72], which is applied as the first step. In the second step, we search for the minimum-area bounding rectangle circumscribing the convex polygon constructed by the convex hull algorithm in the first step. The theorem mentioned above limits the number of bounding rectangles that are candidates for the minimum-area bounding rectangle. However, a solution is not always unique. An example with a unique MABR solution is shown in **Fig. 27**.

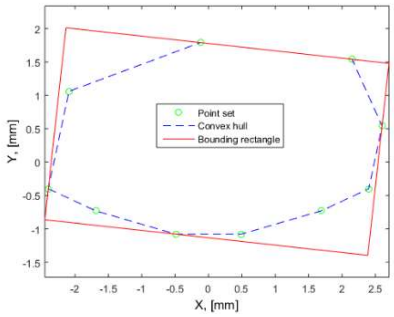
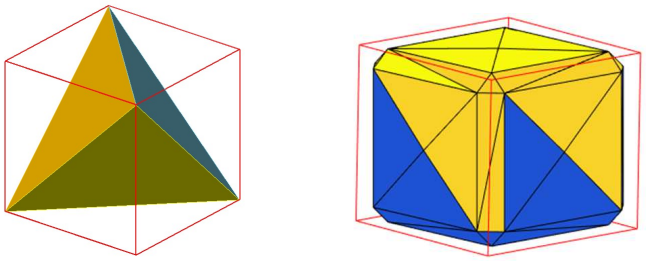


Fig. 27. An example of the MABR solution for a point set in $\mathbb{R}^2$

4.1.2 Minimum-Volume Bounding Box

The second theorem presented here was formulated and proved for the MVBB problem by O'Rourke [80]: *A box of minimal volume circumscribing a convex polyhedron must have at least two adjacent faces (sides) flush with edges of the polyhedron.*



(a) (b)

Fig. 28. The convex polyhedrons enclosed by MVBB: a) a regular tetrahedron with edge length $\sqrt{2}$ circumscribed by MVBB with edge length 1 (conventional units); b) the convex polyhedron based on the regular cube with edge length 1 formed by chamfers with the distances 0.1x0.1, the middle point of each side is above the other four points by 0.1, that is circumscribed by MVBB with edge length 1.02.

It is not necessary that one of the sides of the bounding box is coplanar with one of the faces of the convex polyhedron. In fact, the bounding box with minimal volume circumscribing a regular tetrahedron has all six sides coplanar with the tetrahedron edges without flushing with any tetrahedron faces (**Fig. 28**a).

However, in practise (to be shown in the experimental results, section 4.3.8), the minimal solution may also correspond to the case when one or more sides of the bounding box are coplanar with faces of the convex polyhedron. An example of the bounding box is shown in **Fig. 28**b, that has the minimal volume, and it has four sides coplanar with four faces (12 edges total) of the convex polyhedron.

4.2 Matrix Linear Transformations

The linear transformations are frequently used technique in geometrical metrology, which may include translations, rotations and sometimes reflections (reflection is not considered here). The basic objects of geometry are points, lines (vectors) and planes, which are presented by linear equations. In this work, we use *Cartesian and Homogeneous coordinates* to represent a position and an orientation of the basic objects in both $\mathbb{R}^2$ and $\mathbb{R}^3$ spaces.

4.2.1 A rotation in space around the origin

In order to perform a counterclockwise rotation of a vector $\mathbf{v}\{x, y\}$ with an angle θ around the origin in $\mathbb{R}^2$ (two-dimension) space, a following rotation matrix $R(\theta)$ can be used [86]:

$$R(\theta) = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \tag{47}$$

then, the complete operation can be written as:

$$\mathbf{v}' = [x, y] \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} = [x\cos(\theta) - y\sin(\theta),\, x\cos(\theta) + y\sin(\theta)] = [x', y']. \tag{48}$$

The operation of the vector rotation is illustrated in **Fig. 29**.

The transformation technique in homogeneous coordinates for $\mathbb{R}^2$ is similar, and it is not considered here. The rotation around an arbitrary axis can be done in homogeneous coordinates for $\mathbb{R}^3$ space, and it is described in the next sections.

4.2.2 Homogeneous transformation in space

In order to describe a position and an orientation of point set in $\mathbb{R}^3$ (three-dimension) space for a new coordinate system (CS) relative to an old CS, homogeneous coordinates can be employed. A general homogeneous transformation matrix for $\mathbb{R}^3$ is represented as a four-by-four matrix. There are two types of operations to be considered here: translation and rotation.

4.2.2.1 Translation operation.

The translation is a transformation without any fixed points including the origin. While matrix multiplications have the origin as the fixed point. In order to work around this inconsistency, transformation matrices in homogeneous coordinates are used. The translation matrix in homogeneous coordinates for $\mathbb{R}^3$ can be written in the following form:

$$[T_{xyz}] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ l & m & n & 1 \end{bmatrix}, \tag{49}$$

where *l, m* and *n* are the corresponding movements in x-, y-, and z-directions. Let us consider a vector $\mathbf{v}\{x, y, z\}$ drawn from the origin O to some point *P(x, y, z)* in the old CS (see **Fig. 30**). The new CS is the result of a movement (without any rotations) the origin $O(0,0,0)$ into the new origin $O'(-l, -m, -n)$ by the vector $\overrightarrow{OO'}$ denoted as **u**. The axes of the CS O are parallel to the axes of the CS O'. Then the translation operation in homogenous coordinates for the vector $\mathbf{v}\{x, y, z, 1\}$ with the translation matrix *T* can be written as follows:

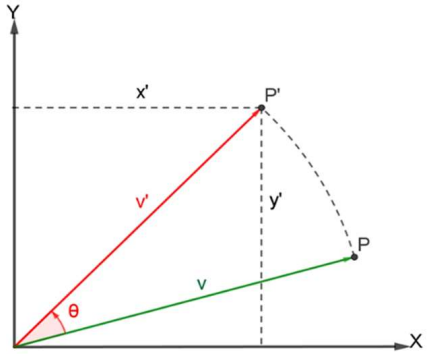


Fig. 29. The rotation of a vector **v** in $\mathbb{R}^2$

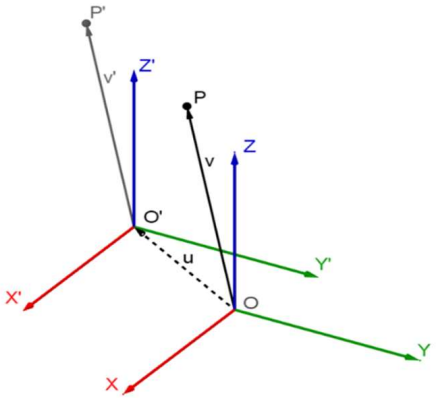


Fig. 30. The translation of a vector **v** into a new CS in $\mathbb{R}^3$

$$\mathbf{v}' = [x\ y\ z\ 1]\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -l & -m & -n & 1 \end{bmatrix} = [(x-l)\ (y-m)\ (z-l)\ 1] = [x'\ y'\ z'\ 1] \tag{50}$$

4.2.2.2 Basic rotation matrices.

A rotation operation around one of the CS axis is often called a *basic* or *unit rotation*. Thus, there are three basic rotations taking place. The *roll* is a rotation around x-axis in the yz-plane. The *pitch* is a rotation around y-axis in the xz-plane. The *yaw* is a rotation around z-axis in the xy-plane [87]. All three rotations are simply two-dimension rotation based on (47). Then, the matrix of the counterclockwise rotations around x-axis, with an angle α in homogeneous coordinates can be written as (roll):

$$[R_x(\alpha)] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) & 0 \\ 0 & -\sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{51}$$

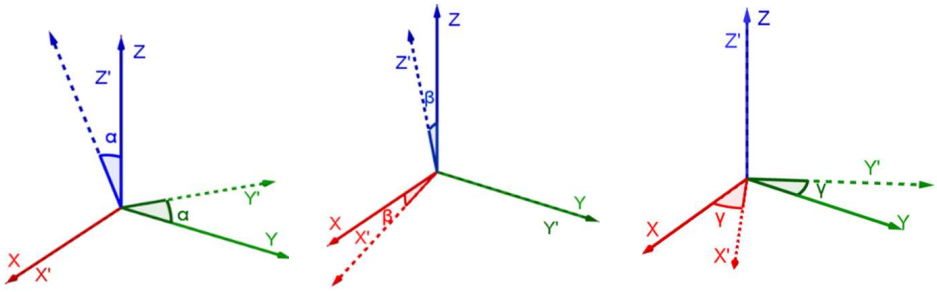
around y-axis with an angle β (pitch)

$$[R_y(\beta)] = \begin{bmatrix} \cos(\beta) & 0 & -\sin(\beta) & 0 \\ 0 & 1 & 0 & 0 \\ \sin(\beta) & 0 & \cos(\beta) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{52}$$

and around z-axis with an angle γ (yaw)

$$[R_z(\gamma)] = \begin{bmatrix} \cos(\gamma) & \sin(\gamma) & 0 & 0 \\ -\sin(\gamma) & \cos(\gamma) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{53}$$

The basic rotations around x, y, z-axes are shown in **Fig. 31**. The homogeneous coordinates for rotations are required only if the rotations matrices will be further used together with other homogeneous transformation matrices e.g. the translation matrix.



(a) Roll (b) Pitch (c) Yaw

Fig. 31. Basic Rotations: a) Rotation around the x-axis with an angle α; b) Rotation around the y-axis with an angle β; c) Rotation around the z-axis with an angle γ.

4.2.2.3 *Compound transformations.*

Using the basic matrices given above as the building blocks, a more complex compound transformation can be constructed. A frequently used operation in Geometrical Metrology is the rotation of objects around an arbitrary axis in $\mathbb{R}^3$. As long as the matrix multiplication is not a commutative operation, then the order of operations (transformation matrices) is important. The main idea is an alignment of an arbitrary axis (vector) with one of the CS-axes.

If some point $P(x_0, y_0, z_0)$ in $\mathbb{R}^3$ belongs to an arbitrary vector **u**, then the rotation around **u** with an angle θ is carried out as follows:

- apply the translation transformation in such way that the point $P(x_0, y_0, z_0)$ is coincident with the origin of the CS;
- apply the subsequent rotations around x-axis and y-axis to get the vector **u** alignment with z-axis;
- apply the rotation around z-axis with the angle θ, which will correspond to the rotation around the vector **u**.

Mathematically, all these transformations can be written as the product of translation T_{xyz}, rotations R_x , R_y and R_z:

$$[C_u] = [T_{xyz}][R_x(\alpha)][R_y(\beta)][R_z(\theta)] \tag{54}$$

4.2.2.4 *A practical example of the compound transformation.*

The transformation procedure described below is representing a substantial part of the new developed metrological version of the Minimum Volume Bounding Box (MVBB) algorithm proposed in Paper 6 [23] (described in Section 4.3).

There are two non-coplanar planes φ_1 and φ_2 based on four points $P_0(x_0, y_0, z_0)$, $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$ and $P_3(x_3, y_3, z_3)$ defined in $\mathbb{R}^3$as shown in **Fig. 32**.

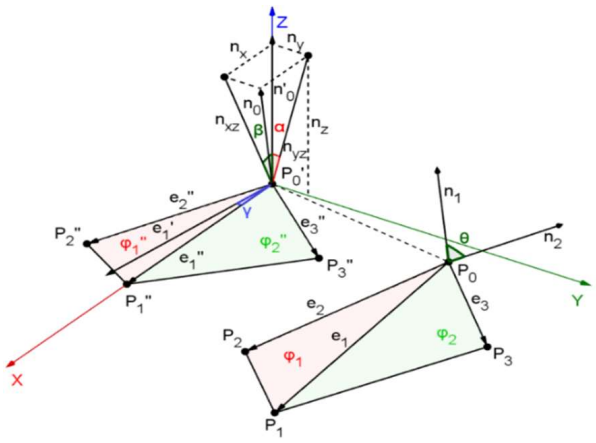


Fig. 32. Coordinate transformation of two adjacent planes φ_1, φ_2 with the common edge e_1

These two planes have the common edge P_0P_1, and we will therefore perform the following operations:

1. The alignment of the plane φ_1 with xy-plane of the CS;
2. The rotation of the planes φ_1, φ_2 around the edge P_0P_1 to combine the plane φ_2 with xy-plane of the CS.

First, we would like to define the smallest angle between the two planes, which can be found as the angle θ between two corresponding normal vectors $\boldsymbol{n}_1, \boldsymbol{n}_2$ to the planes. Let us to define three non-collinear vectors $\overrightarrow{P_0P_1}$, $\overrightarrow{P_0P_2}$ and $\overrightarrow{P_0P_3}$ based on the given points as the following $\boldsymbol{e}_1\{x_1 - x_0, y_1 - y_0, z_1 - z_0\}$, $\boldsymbol{e}_2\{x_2 - x_0, y_2 - y_0, z_2 - z_0\}$ and $\boldsymbol{e}_3\{x_3 - x_0, y_3 - y_0, z_3 - z_0\}$ or in the shorter form $\boldsymbol{e}_1\{a_1, b_1, c_1\}$, $\boldsymbol{e}_2\{a_2, b_2, c_2\}$ and $\boldsymbol{e}_3\{a_3, b_3, c_3\}$. Thus, the planes φ_1 and φ_2 can be also defined by pairs of two non-collinear vectors $\boldsymbol{e}_1, \boldsymbol{e}_2$ and $\boldsymbol{e}_1, \boldsymbol{e}_3$. Then the normal vectors $\boldsymbol{n}_1, \boldsymbol{n}_2$ to the planes can be found as the cross product of these vector pairs:

$$\boldsymbol{n}_1\{A_1, B_1, C_1\} = \boldsymbol{e}_2 \times \boldsymbol{e}_1 = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ a_2 & b_2 & c_2 \\ a_1 & b_1 & c_1 \end{vmatrix}, \tag{55}$$

and

$$\boldsymbol{n}_2\{A_2, B_2, C_2\} = \boldsymbol{e}_1 \times \boldsymbol{e}_3 = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ a_1 & b_1 & c_1 \\ a_3 & b_3 & c_3 \end{vmatrix}. \tag{56}$$

Where A_1, B_1, C_1, A_2, B_2, C_2 are corresponding minors of the matrices given by (55) and (56) of the following forms:

$$A_1 = \begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix}, \quad B_1 = \begin{vmatrix} c_1 & a_1 \\ c_2 & a_2 \end{vmatrix}, \quad C_1 = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \tag{57}$$

$$A_2 = \begin{vmatrix} b_3 & c_3 \\ b_1 & c_1 \end{vmatrix}, \quad B_2 = \begin{vmatrix} c_3 & a_3 \\ c_1 & a_1 \end{vmatrix}, \quad C_2 = \begin{vmatrix} a_3 & b_3 \\ a_1 & b_1 \end{vmatrix} \tag{58}$$

Then the sharp angle between the normal vectors $\boldsymbol{n}_1\{A_1, B_1, C_1\}$ and $\boldsymbol{n}_2\{A_2, B_2, C_2\}$ can be calculated as following [86]:

$$\theta = \pi - \arccos\left(\frac{AA_2 + BB_2 + CC_2}{\sqrt{A^2 + B^2 + C^2}\sqrt{A_2^2 + B_2^2 + C_2^2}} \right) \tag{59}$$

In order to combine the plane φ_1 with xy-plane of the CS, we need to align the normal vector $\boldsymbol{n}_1$ with the positive z-axis. Then, the vector $\boldsymbol{n}_1$ must be displaced into the origin of the CS by using the translation matrix T_{xyz} given by (49) and the coordinates of the point $P_0(x_0, y_0, z_0)$:

$$\left[T_{xyz}\right] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -x_0 & -y_0 & -z_0 & 1 \end{bmatrix}. \tag{60}$$

Let us call the result of the transformation of $\boldsymbol{n}_1$ as vector $\boldsymbol{n}_0$. The projections n_{yz}, n_{xz} of the vector $\boldsymbol{n}_0$ on the planes yz ($x=0$) and xz ($y=0$) with their coordinates n_x, n_y, n_z respectively (see **Fig. 32**), give us two angles α and β:

$$\alpha = \arcsin\left(\frac{n_y}{n_{yz}}\right) = \arcsin\left(\frac{n_y}{\sqrt{{n_z}^2 + {n_y}^2}}\right), \tag{61}$$

and

$$\beta = \arcsin\left(\frac{n_x}{n_{xz}}\right) = \arcsin\left(\frac{n_x}{\sqrt{{n_z}^2 + {n_x}^2}}\right). \tag{62}$$

The calculated angles are substitute in the rotation matrices (51) and (52), and then the final transformation matrix to combine the plane φ_1 with xy-plane can be written as:

$$[A_{nz}] = \left[T_{xyz}\right]\left[R_x(\alpha)\right]\left[R_y(-\beta)\right]. \tag{63}$$

The second operation is the rotation of the planes φ_1, φ_2 around the vector (common edge) $\boldsymbol{e}_1$ to combine the plane φ_2 with xy-plane of the CS. In order to do that, the vector $\boldsymbol{e}_1'$ (which is the vector $\boldsymbol{e}_1$ after transformation by (63)) must be aligned with x-axis. As long, as the vector $\boldsymbol{e}_1'$ lays in xy-plane with the beginning of the vector in the origin of CS ($|\boldsymbol{e}_1'| := e_{xy}'$), then the angle γ between $\boldsymbol{e}_1'$ and x-axis can be found by:

$$\gamma = \arcsin\left(\frac{e_x'}{e_{xy}'}\right) = \arcsin\left(\frac{e_x'}{\sqrt{{e_y}'^2 + {e_x}'^2}}\right), \tag{64}$$

where e_x', e_y' are the corresponding coordinates of the vector $\boldsymbol{e}_1'$ on the x- and y-axes respectively (e_x', e_y' are not shown in **Fig. 32**). Then, the vector $\boldsymbol{e}_1'$ can be aligned with x-axis by using the rotation matrix given by (53). The result of the transformation of the two faces φ_1, φ_2 into φ_1'', φ_2'' is illustrated in **Fig. 32**.

After all necessary alignments are completed, the counterclockwise rotation of the planes φ_1'', φ_2'' around the edge $\boldsymbol{e}_1''$ can be proceeded by using the basic rotation matrix given by (51) with the angle θ calculated by eq. (59). Then, the compound transformation including all operations described above can be written as the following:

$$\begin{bmatrix} x_0 & y_0 & z_0 & 1 \\ x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \end{bmatrix} [A_{nz}]\left[R_z(\gamma)\right]\left[R_x(\theta)\right] = \begin{bmatrix} x_0' & y_0' & z_0' & 1 \\ x_1' & y_1' & z_1' & 1 \\ x_2' & y_2' & z_2' & 1 \\ x_3' & y_3' & z_3' & 1 \end{bmatrix}. \tag{65}$$

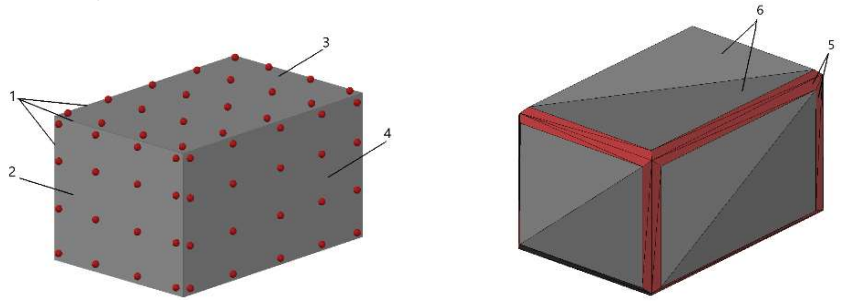
4.3 Contributions to objective 3

The questions and objective related to the MVBB problem have been formulated in section 1.2.3 and 1.3.1. The formulated questions are relevant for those who deal with Computation Geometry related to the metrological applications.

The approaches discussed in section 4.1 was mainly focused on reducing the computation time, but at the expense of accuracy. The developed approach provided in this work for the minimum bounding box on reference standards used for calibration

of dimensional measuring systems; hence, the accuracy must be ensured. The elegant approach provided by O'Rourke is the accurate solution, but it does not take into account some metrological issues related to the discrete point measurement with CMM, which are discussed below.

The physical edges (denoted by 1 in **Fig. 33**a) of the cuboid object are typically not measured and there is always a distance between the edges and the measured points. As a result, there is an intermediate space between the measured points on all pairs of the adjacent sides (e.g. side 2 and side 3, side 3 and side 4 in **Fig. 33**a) of the cuboid object. This intermediate space is transformed into a large number of the convex polyhedron faces after appliance of the convex hull operation. Such newly constructed faces provide acute angles and look similar to "chamfers" faces (denoted by 5 in **Fig. 33**b). These faces cut off the physical cuboid object and they will lead to unnecessary computation in the O'Rourke algorithm. Obviously, these "chamfer" faces cannot be a part of the minimum bounding box solution and these faces should be excluded from the algorithm.



(a) (b)

Fig. 33. An example of the metrological issue: a) a cuboid object with CMM measured points; b) an example of the convex polyhedron with the chamfer faces after convex-hull operation; 1 – edges; 2 – left side; 3 – top side; 4 – front side; 5 – "chamfer" polyhedron faces; 6 – ordinary polyhedron faces.

Paper 6 [23] proposed the algorithms for estimation of MVBB that takes into account the effect of the "chamfer" faces. The MVBB criteria is applied to the physical objects with an actual shape close to the perfectly rectangular bounding box. The developed accurate algorithm searches for the minimum solution according to the conditions defined by two theorems related to the MABR and the MVBB problems presented in sections 4.1.1 and 4.1.2 respectively. The brief description of three conventional geometrical algorithms together with data pre-process algorithm is given further in sections 4.3.1-4.3.5. Implementation of the methods is presented in sections 4.3.6-4.3.8 together with description of the experimental setup and the computational results.

4.3.1 The computation methods

Three methods for finding the Minimum Volume Bounding Box (MVBB) are presented here. The methods are denoted as the "*side-"* method (MVBBS), the "*face-"*

method (MVBBF) and the "*edges-"* method (MVBBE). All three methods differ from each other by accuracy, complexity and hence the computation time. The reason we consider these three methods is that their appliance depends on which measurement system and measuring procedure are going to be used.

All the three methods utilize the MABR algorithm. Two of the methods ("*face-"*, "*edges-"*) include the specific *data pre-processing algorithm* (to be described further), which distinguishes these methods from other known methods. Only the *MVBBE method* completely satisfies both theorems given in sections 4.1.1 and 4.1.2, and therefore it can be used as the *reference* for the other alternative methods.

4.3.2 The Minimum Volume Bounding Box Side (MVBBS) Method

This MVBBS approximation method is well known and often used in practice. It is fast and straightforward, based on an assumption that the test object has one perfectly flat side e.g. *Bottom*, which is aligned with the support surface ($Z_{\min}$). Such assumption allows a substantial simplification, in the measurement procedure and the computation procedure. However, because of such assumption the calculated minimal volume by this method is often overestimated. Groen et al. [88] developed an operational automatic system for measurement of parcels and suitcases on a conveyor belt based on this principle.

The principle of this method is to define the height as $H_{min} = Z_{max} - Z_{min}$ and the smallest area A_{min} of the bounding rectangle of the xy-projection of all measured points. As long as a single 2D projection of the convex polyhedron is considered, then the MABR algorithm is applied only once.

4.3.3 Data pre-processing

The main idea of the new developed approach is based on the data pre-processing algorithm. Generally, the most precise methods include a large number of computation operations and require a longer time. As it was mentioned above, many "chamfer" faces are created by the convex hull operation. These faces never become a part of the minimum bounding box solution. By excluding these faces from the MVBB algorithm, the computation time is compensated. This achievement is especially desired for the "*edges-"* method (MVBBE), which provides the most optimal solution. Let us have a look at some details of the developed approach, which is focused on solving the minimum bounding box problem for physical objects that are rectangular objects close to the perfectly shaped bounding box.

The following measurement procedure with CMM was applied. The measured points of the object sides are given as six sets of points: *Front, Back, Right, Left, Top* and *Bottom*. The data set of each side is a $n \times 3$ matrix containing x- y- and z-coordinates for the n number of points.

The input of the convex hull operation are the point coordinates from the six sets of points jointed together as illustrated in the **Fig. 34**. The output from the convex hull operation is a matrix $S_{m,3}^{Pol}$ with m rows. Each row of the matrix is a convex polyhe-

dron face φ_i defined by its three vertices. The vertices are given as indices that refer to the input data to the convex hull operation.

Some of the faces of the polyhedron described by $S_{m,3}^{Pol}$ will have vertices from two or three adjacent sides of the physical cuboid object. For example, the measured points from the *Top* side may be combined with measured points from the *Front* side and *Left* side into common faces, or "chamfer" faces between the sides (**Fig. 33**b). When defining the minimum bounding box in measurement and calibration of rectangular objects, these combined faces and their edges are not a part of the solution, and thus they can be excluded from the algorithm.

Two data structures are constructed from the output matrix $S_{m,3}^{Pol}$ by the pre-processing algorithm. The first structure represents six matrices $S^F, S^B, S^R, S^L, S^T, S^M$ of face vertices $v_{i,j}$ separated according to the reference object sides (*Front, Back,...Bottom*) without common "chamfer" faces (**Fig. 34**, denoted by I); the second is a data matrix $P_{k,4}$ with x- y- and z-coordinates for face vertices (**Fig. 34**, denoted by II). Such data structure allows to avoid a large number of unnecessary calculations.

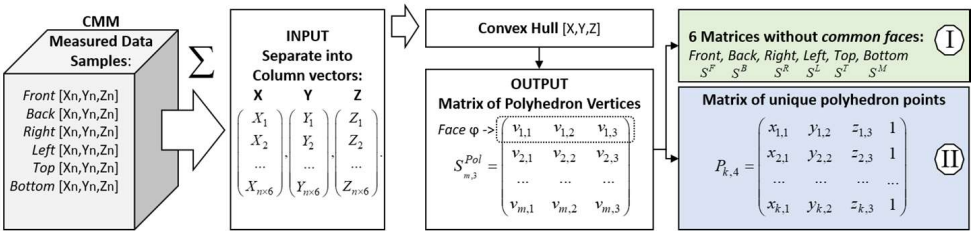


Fig. 34 The flowchart of the data pre-processing with the two data structures I and II

4.3.4 The Minimum Volume Bounding Box Face (MVBBF) Method

The MVBBF method developed by the author is based on the pre-process algorithm and therefor it is more efficient than other MVBBF methods based on the same assumption. The main assumption of this method is that one side of the bounding box is coplanar with a face of the convex polyhedron. The accuracy of the method is better than the previously described MVBBS method, however it is still approximated solution. The flowchart of the algorithm of MVBBF method is shown in **Fig. 35**.

The six matrices (S^F, S^B ... S^M) associated with the cuboid reference object sides are the output of the pre-processing algorithm (**Fig. 34**, denoted by I). The algorithm searches through the six matrices S^F, S^B ... S^M associated with sides of the measured cuboid object and checks all faces within each set.

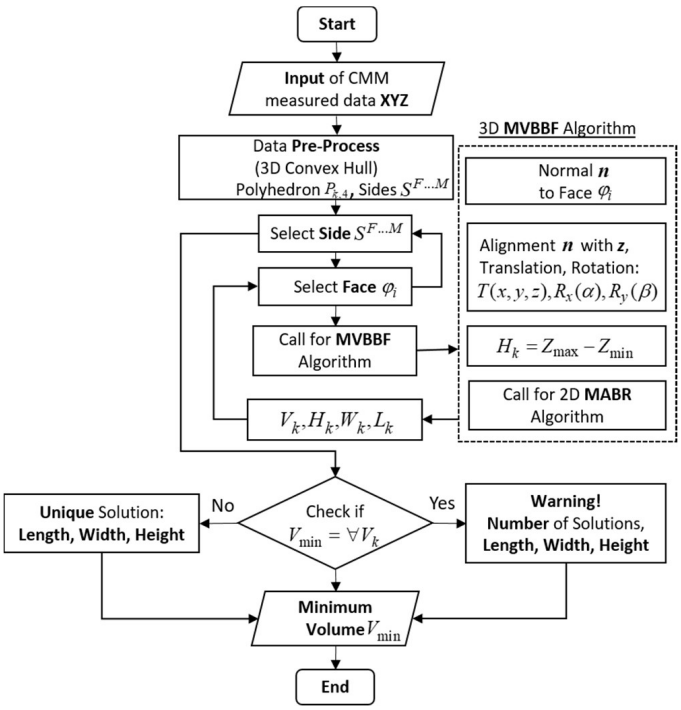


Fig. 35. The flowchart of the MVBBF method

Each newly selected face is aligned with xy-plane of CS by means of the transformation technique described in section 4.2.2.4. After the coordinate transformation is completed, all newly transformed points are projected into the xy-plane. Then the MABR algorithm (section 4.1.1) is applied for the projected points. It defines an orientation of the closed-loop of adjacent sides and, hence the estimation of the width W_k and the length L_k of the minimum bounding box. The height H_k is defined as a difference between maximum and minimum z-values: $Z_{max} - Z_{min}$, related to the current orientation of the point set. Thus, the volume is: $V_k = H_k \cdot W_k \cdot L_k$.

After all faces of all six matrices (S^F, S^B ... S^M) are examined for the volume value V_k, then the smallest value V_{min} is chosen as the final solution.

4.3.5 The Minimum Volume Bounding Box Edge (MVBBE) Method

The MVBBE method completely corresponds to the conditions of the theorem presented in section 4.1.2 and therefore this is the most accurate method, which guarantees the global minimum solution. However, the algorithm is more complex and hence slower than two previous methods. In this case, the data pre-process becomes the most crucial part of the method due to a significant reduction of unnecessary computation of the "chamfer" faces and the corresponding edges of the convex polyhedron. The MVBBE algorithm is shown in **Fig. 36**.

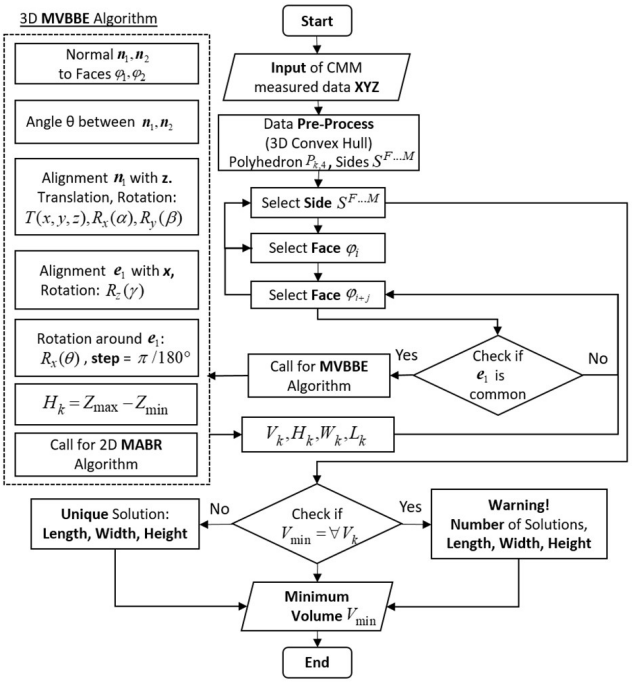


Fig. 36. The flowchart of the MVBBE method

In this case, the algorithm searches through the six matrices S^F, S^B ... S^M of the pre-process algorithm output and examines every single pair of faces with common edge. Each common edge of a newly selected pair of faces is align with x-axis in such way that one of these two faces is aligned with xy-plane at the same time. Then a counterclockwise rotation around the common edge of all points corresponding to the matrix $P_{k,4}$ is applied with one degree step until the second face is compound with xy-plane. The transformation technique utilized for these operations is described in section 4.2.2.4 (*A practical example of the compound transformation*). In each step of the rotation, the height H_k is defined as a difference between maximum and minimum z-values, and all newly transformed points are projected into xy-plane. In order to estimate the width W_k and the length L_k of the minimum bounding box, the MABR algorithm (section 4.1.1) is applied for the projected points. Thus, the volume is calculated as: $V_k = H_k \cdot W_k \cdot L_k$. After all pairs of faces in six matrices (S^F, S^B ... S^M) are examined for the minimum volume solution V_k, then the smallest value V_{min} is selected as the final solution.

4.3.6 Implementation of the MVBB methods

The algorithms described in the previous sections are implemented in MATLAB® programming environment based on CMM measurement. The measurements have been performed in a Leitz PMM-C-600 CMM with an analogue probe. The PC-DMIS software was utilized for operation of the CMM.

4.3.7 Experiment setup

A cuboid object with the following nominal dimensions: length 210 mm, width 140 mm, and height 120 mm was used for the experimental test. The test object is shown in **Fig. 37**. The measured data is arranged into separated data samples according to the cuboid sides: *Front, Back, Right, Left, Top* and *Bottom*. Each sample is a $n \times 3$ matrix with three columns and *n*-rows of xyz-coordinates corresponding to the *n*-measured points as shown in the **Fig. 34**. We have used a uniform distribution of measured points with 15 mm distance between the points. The total number of the measuring points is $N = 650$.

In order to get complete measurements of all six sides of the test object in a common coordinate system, we have measured the object in two setups. The measurements of the two setups can be combined by using common alignment points in the two setups.

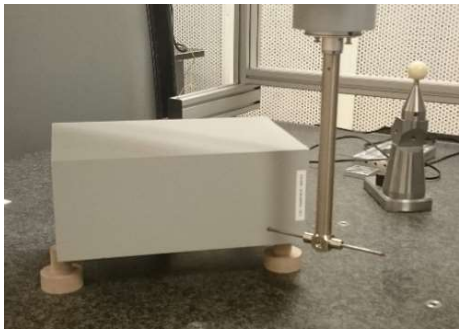


Fig. 37. The cuboid test object in CMM

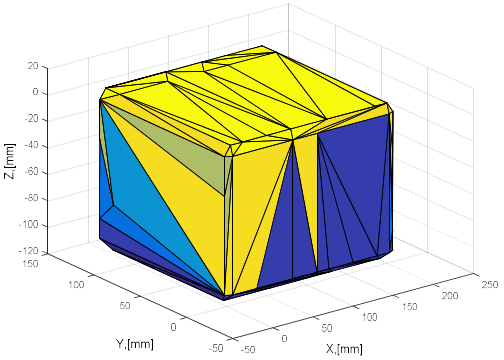


Fig. 38. The convex polyhedron after the convex hull operation

4.3.8 Results

The collected data is further exported to a MATLAB code as an input for the developed algorithms. The result of the convex hull operation for measured data is shown in the **Fig. 38**. There are **166** faces combined together into one convex polyhedron. There are **88** faces are left after applying of the *data pre-process algorithm* (almost 50% of calculation is reduced). The computation results of all three methods are tabulated in **Table 14** (the results are rounded to 1e-3 order).

We consider the Bottom side as the support side for the MVBBS method. The MVBBS method provides a significant overestimation of the volume V_S of the bounding box relative to the other two methods: $\Delta V = V_S - V_E = 951.959$ mm^3.

Table 14. The computation results of the MVBBS, MVBBF, MVBBE methods for the first test

Method	Width, mm	Length, mm	Height, mm	Volume, mm^3	Number of solutions
MVBBS	140.016	210.035	119.980	3528388.312	1
MVBBF	139.997	210.010	119.978	3527436.353	1
MVBBE	139.997	210.010	119.978	3527436.353	1

There is no difference between estimated volumes from the MVBBF and the MVBBE methods. A possible reason for such coincidence may be a small form deviation and as a result, a small, below one-degree, angle between the polyhedron faces.

An extra test was applied for estimation of MVBB for the cuboid object with the same nominal dimensions but with larger flatness deviations. The computation results are given in **Table 15.**

Table 15. The computation results of the MVBBS, MVBBF, MVBBE methods for the second test

Method	Width, mm	Length, mm	Height, mm	Volume, mm^3	Number of solutions
MVBBS	140.016	210.195	120.068	3533662.007	1
MVBBF	139.995	210.195	120.054	3532743.929	1
MVBBE	139.994	210.195	120.055	3532735.210	1

We can observe a difference: $\Delta V = V_F - V_E = 8.675$ mm^3 between the solutions of MVBBF and MVBBE methods. The MVBBE method may provide the minimal solution and yet, it includes all solutions of the MVBBF method and therefor it is more reliable and accurate.

Three methods have been proposed and demonstrated in this section to estimate the minimum volume of bounding box with the proposed *data pre-process algorithm* for the metrological applications. The first two methods are based on a number of assumptions allowing decreasing of the computation time but often with overestimated results. The minimal and the most optimal solution is provided by the MVBBE method. Furthermore, the solution of the MVBBE method is based on theorems presented in this paper and hence, its estimation is the most accurate. Relying on an application, different methods may be applicable while the MVBBE method should stay as the reference.

The original algorithm for the MVBBE may include relatively large number of computations due to the calculations on the “chamfer” faces. The proposed data pre-process algorithm based on the specific metrological conditions (described above)

allows a significant reduction (almost 50%) of the amount of computation, preserving the initial accuracy at the same time.

Thus, the MVBBE method should be used for those metrological tasks, where the accuracy is the critical factor, particularly when a large geometry form deviation is expected. The principles outlined in this work could also improve the functionality of operation software for the measuring systems.

5 Conclusion and Future Research

New approaches and algorithms in geometrical inspection with CMM are proposed in this PhD thesis. In this section, some general comments and recommendations for further research are given.

The following aspects of measurements were considered in this thesis: *sample strategy*, *outlier detection methods*, and *algorithms* for calculation of *substitute elements*. The new methods were developed according to various task definitions due to the tolerance types and the actual workpiece geometry deviations. As a result, five approaches have been developed and presented in this thesis: three methods related to the sample size problem, one approach estimating the most efficient outlier detection method in the CMM measurements, and one approach devoted to the MVBB problem. These approaches have been developed through the research presented in the scientific published papers provided in Part II of this thesis.

The influence of the sample size on the measurement uncertainty has been evaluated by three different approaches. The first approach (Paper 1) based on ISO16269-6 allows to define the desired confidence level of a certain content of the radius variation due to the sample size for the circular profiles. The second approach (Paper 3) is developed in accordance with ISO14405-1 for the two-point diameter problem. The approach is not only able to estimate the necessary sample size for the two-point diameter problem, but also it provides an opportunity for substantial reducing the number of measured points that could be especially useful for online measurements with CNC machine tool. However, these two approaches based on the statistical simulations, have a number of constraints. The third approach (Paper 4), based on the Artificial Neural Network (ANN), do not have these constraints and it can provide the simulation results closer to the real measuring routine.

The approaches described above will improve the reliability and accuracy of measuring strategies for GD&T inspection. The approaches can be used to define the optimal sample size based on the product specifications and manufacturing conditions. The optimal sample size can be set as the default parameter in the CMM operating software. The approaches can be further applied to develop a generalized measuring strategy through the collection of the measured data of workpieces produced by various machine processes and workpiece materials.

Paper 2 considers methods for detection of the outliers in the CMM measurements. The efficiency of the methods has been investigated based on experimental data and statistical simulations.

A new approach for analysis of measurement data of convex polyhedron objects has been presented in Paper 6. Several methods for computing of Minimum Volume Bounding Box (MVBB) have been developed based on the new approach.

All presented methods are based on the experimental data and thus can provide the most realistic estimations of the actual geometry and dimensional deviations. Thereby their further implementation into the production will have both scientific and practical interest.

The ideas presented in this PhD thesis could contribute to the solution of some of the problems in geometrical inspection in the industry. However, there is still a large potential for further improvements and research. The solutions related to the outlier detection procedure, the optimal sample size and the MVBB problem can be integrated into the existing software for CMM such as PCDMIS, QUINDOS and similar. The ANN model presented in Paper 4 can be further utilized for simulation of various sample strategies to clarify their relationship with workpiece profiles derived from various machining processes.

The final purpose of all these efforts is unambiguous inspection results, reduction of the measurement uncertainty, and improvement of their reliability with minimized time costs.

6 References

1. Henzold Georg (2006) Geometrical Dimensioning and Tolerancing for Design, Manufacturing and Inspection : A Handbook for Geometrical Product Specification using ISO and ASME standards. Geometrical dimensioning and tolerancing for design, manufacturing, and inspection, 2nd ed. edn. Elsevier Science, Burlington.

2. Henrik S. Nielsen (2012) The ISO Geometrical Product Specifications Handbook. Find your way in GPS. 1 edn. ISO/Danish Standards 2012, Denmark.

3. FMCTechnologies (2012) UCON Connection Systems. - White paper.

4. Colosimo Bianca M., Senin Nicola (2010) Geometric Tolerances : Impact on Product Design, Quality Inspection and Statistical Process Monitoring. Springer London, London.

5. Bjørke Øyvind (1989) Computer-Aided tolerancing. 2nd ed. edn. ASME Press, New York.

6. Zhang Hong-Chao, Lin E.E. (2001) Theoretical Tolerance Stack-up Analysis Based on Tolerance Zone Analysis. The International Journal of Advanced Manufacturing Technology 17(4) (Texas Tech University, Lubbock, Texas, USA):257-262.

7. Kandikjana T. , Shahb J.J. , Davidsonb J.K. (2001) A mechanism for validating dimensioning and tolerancing schemes in CAD systems. Computer-Aided Design:721-737.

8. Antonio Armillotta (2013) A method for computer-aided specification of geometric tolerances. Computer-Aided Design 45:1604-1616.

9. De Chiffre L., Hansen Hans Nørgaard , Andreasen Jan Lasson , Savio Enrico, Carmignato Simone (2015) Geometrical Metrology and Machine Testing. DTU Mechanical Engineering.

10. Chelishchev Petr, Sørby Knut Optimization of sample size for two-point diameter verification in coordinate measurements, Advanced Manufacturing and Automation VIII. In: International workshop of Advanced Manufacturing and Automation (IWAMA2018), Changzhou, China, 2019. Springer Nature Singapore Pte Ltd., pp 313-321.

11. Chelishchev Petr, Popov Aleksandr, Sørby Knut (2018) Robust estimation of optimal sample size for CMM measurements with statistical tolerance limits. Paper presented at the ICMSC 2018 The 2nd International Conference on Mechanical, System and Control Engineering,

12. Chelishchev Petr, Sørby Knut Simulation algorithm of sample strategy for CMM based on Neural Network Approach. In: International workshop of Advanced Manufacturing and Automation (IWAMA2019), 2020. Springer Nature Singapore Pte Ltd.

13. Chelishchev Petr, Popov Aleksandr, Sørby Knut (2018) An investigation of outlier detection procedures for CMM measurement data. Paper presented at the ICMSC 2018 The 2nd International Conference on Mechanical, System and Control Engineering,

14. Smith Graham T. (2016) Machine Tool Metrology: An Industrial Handbook. Cham: Springer International Publishing, Cham.

15. Documents concerning the new definition of the metre (1984). Metrologia 19 (4):163-178.

16. Privalov V.E. (1987) Quantum Electronics and the New Defenition of Meter / Квантовая электроника и новое определение метра. Znanie, Leningrad.

17. Privalov V.E. (1989) Discharge Lasers in Measuring Complexes, Leningrad: Shipbuilding / Газоразрядные лазеры в измерительных комплексах.

18. Ivanov V.A., Privalov V.E. (1993) Laser Applications in Precise Mechanical Devises / Применение лазеров в приборах точной механики. Polytechnics, Saint Petersburg.

19. Dolgikh G.I., Privalov V.E. (2016) Laser Physics. Basic and Applied Research / Лазерная физика. Фундаментальные и прикладные исследования. Reya, Vladivostok.

20. Mian Syed Hammad, Al-Ahmari, Abdulrahman M. (2017) Application of the sampling strategies in the inspection process. Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture 231 (4):565-575.

21. ISO16269-4 (2010) Detection and treatment of outliers. Part 4: Statistical interpretation of data.

22. Chelishchev Petr, Sørby Knut, Privalov Vadim (2019) Perspectives for appliance and accuracy improvement of coordinate measurements with laser technique Paper presented at the 2019 IEEE International Conference on Electrical Engineering and Photonics, St.Petersburg Polytechnic University, Russia,

23. Chelishchev Petr, Sørby Knut (2019) Estimation of Minimum Volume of Bounding Box for Geometrical Metrology. unpublished.

24. BS7172 (1989) Assessment of position, size and departure from nominal form of geometric features. BSI, London

25. Moroni Giovanni, Petrò Stefano, Colosimo Bianca M, Senin Nicola (2010) Coordinate Measuring Machine Measurement Planning. London: Springer London, London.

26. ISO12181-2 (2011) (GPS) Specification operator.

27. Mesay T. Desta, Feng Hsi-Yung , Daoshan OuYang (2003) Characterization of general systematic form errors for circular features. International Journal of Machine Tools and Manufacture 43 (11):1069-1078.

28. Jiang Qimi, Feng Hsi-Yung, Ouyang Daoshan , Desta Mesay T. (2006) A roundness evaluation algorithm with reduced fitting uncertainty of CMM measurement data. Journal of Manufacturing Systems 25 (3):184-195.

29. Cho N., Tu J. (2001) Roundness modeling of machined parts for tolerance analysis. Precision Engineering 25 (1):35-47.

30. Ruffa S., Panciani G.D. , Ricci F., Vicario G. (2013) Assessing measurement uncertainty in CMM measurements: comparison of different approaches. IntJMetrol Qual Eng 4:163-168.

31. Bernard C.J., Chiu Sheng-Dian (2002) Form tolerance-based measurement points determination with CMM. Journal of Intelligent Manufacturing 13 (2):101-108.

32. Yau Hong-Tznong , Menq Chia-Hsiang (1992) An automated dimensional inspection environment for manufactured parts using coordinate measuring machines. International Journal of Production Research 31 (11):1517-1536.

33. Menq C.-H., Yau H.-T., Lai G.-Y. (1992) Automated precision measurement of surface profile in CAD-directed inspection. Robotics and Automation, IEEE Transactions on 8 (2).

34. Cappetti N, Naddeo A, Villecco F (2016) Fuzzy approach to measures correction on Coordinate Measuring Machines: The case of hole-diameter verification. Measurement 93 (Elsevier Ltd.):41-47.

35. Papananias Moschos, Fletcher Simon, Andrew Peter Longstaff, Mengot Azibananye A novel method based on Bayesian regularized artificial neural networks for measurement uncertainty evaluation. In: EUSPEN, Proceedings of the 16th International Conference of the European Society for Precision Engineering and Nanotechnology, Nottingham, UK, 2016. EUSPEN, pp 97-98.

36. Zhang Y. F., Nee A. Y. C., Fuh J. Y. H. , Neo K. S. , Loy H. K. (1996) A neural network approach to determinig optimal inspection sampling size for CMM. Computer Integrated Manufacturing Systems 9:161-169.

37. Minyang Yang, Eungki Lee (2000) Improved neural network model for reverse engineering. International Journal of Production Research 38 (9):2067-2078.

38. Chang-Xue Feng, Xianfeng Wang (2002) Digitizing uncertainty modeling for reverse engineering applications: regression versus neural networks. Journal of Intelligent Manufacturing 13 (3):189-199.

39. Sładek Jerzy A. (2016) Coordinate Metrology: Accuracy of Systems and Measurements. Springer Berlin Heidelberg, Berlin, Heidelberg, Berlin, Heidelberg.

40. Summerhays K. D., Henke R. P., Baldwin J. M. , Cassou R. M. , Brown C. W. (2002) Optimizing discrete point sample patterns and measurement data analysis on internal cylindrical surfaces with systematic form deviations. Precision Engineering 26 (1):105-121.

41. Cui Changcai, Fu Shiwei, Fugui Huang (2009) Research on the uncertainties from different form error evaluation methods by CMM sampling. The International Journal of Advanced Manufacturing Technology 43 (1):136-145.

42. Walpole Ronald E. (2007) Probability & statistics for engineers & scientists. 8th ed. edn. Pearson Prentice Hall, Upper Saddle River, N.J.

43. Alexandre B.T. (2009) Introduction to Nonparametric Estimation. Spring-Verlag, New York.

44. Jones M. C. , Kappenman R. F. (1991) On a class of kernel density estimate bandwidth selectors. Scandinavian Journal of Statistics 19:337-349.

45. Karunamuni R., Mehra K. (1991) Optimal convergence properties of kernel density estimators without differentiability conditions. Annals of the Institute of Statistical Mathematics 43 (2):327-346.

46. Wand M.P., Jones M.C. (1995) Kernel Smoothing Monographs on statistics and applied probability Chapman&Hall/CRC, Boca Raton, Fla.

47. Grubbs Frank (1969) Procedures for Detecting Outlying Observations in Samples. Technometrics 11:1-21.

48. Rosner B. (1983) Percentage Points for a Generalized ESD Many-Outlier Procedure. Technometrics 25:165-172.

49. Scheffe H., Tukey J. W. (1945) Non-Parametric Estimation. I. Validation of Order Statistics. Ann Math Statist 16 (2):187-192.

50. ISO16269-6 (2014) Determination of statistical tolerance intervals. Statistical interpretation of data.

51. Wilks S. S. (1941) Determination of Sample Sizes for Setting Tolerance Limits. AMS Ann Math Statist 12:91-96.

52. Wilks S. S. (1942) Statistical Prediction with Special Reference to the Problem of Tolerance Limits. . AMS Ann Math Statist 13:400-409.

53. Grossberg S. (1982) Studies of the Mind and Brain. Reidel Press, Drodrecht, Holland.

54. Wang Kesheng (2007) Applied computational intelligence in intelligent manufacturing systems, vol vol. 2. International series on natural and artificial intelligence, 2nd ed. edn. Advanced Knowledge International, Adelaide, S.Aust.

55. Vogl T., Mangis J., Rigler A., Zink W., Alkon D. (1988) Accelerating the convergence of the back-propagation method. Advances in Computational Neuroscience 59 (4):257-263.

56. Beale Mark Hudson, Hagan Martin T., Demuth Howard B. (2017) Neural Network Toolbox, User guide.

57. Warren S. Sarle (2002) Neural Networks, FAQ.

58. Jones William P., Hoskins Josiah (1987) Back-propagation: a generalized delta learning rule. Byte 12 (11):155-162.

59. Vogl T.P., Mangis J.K. , Rigler A.K. , Zink W.T. , Alkon D.L. (1988) Accelerating the convergence of the backpropagation method. Biological Cybernetics Vol. 59:257–263.

60. Marquardt D. (1963) An Algorithm for Least-Squares Estimation of Nonlinear Parameters. SIAM Journal on Applied Mathematics 11:431–441.

61. Hagan M.T., Menhaj M.B. (1994) Training feedforward networks with the Marquardt algorithm. Neural Networks, IEEE Transactions on 5 (6):989-993.

62. ISO14405-1 (2016) Geometrical product specifications (GPS) — Dimensional tolerancing. Linear sizes.

63. Barnett V., Lewis T. (1994) Outliers in Statistical data. Wiley, New York.

64. Chambers John, William Cleveland, Beat Kleiner, Paul Tukey (1983) Graphical Methods for Data Analysis. Wadsworth.

65. Sim C.H., Gan F.F., Chang T.C. (2005) Outlier Labeling with Boxplot Procedures. Journal of the American Statistical Association 100:642-652.

66. ISO5725-2 (1994) Accuracy of measurement methods and results. Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method.

67. Langford E. (2006) Quartiles in Elementary Statistics. Journal of Statistics Education 14(3).

68. Anderson T.W., Darling D.A. (1954) A Test of Goodness-of-Fit. Journal of the American Statistical Association 49:765–769.

69. Royston Patrick (1982) An extension of Shapiro and Wilk's W test for normality to large samples. Applied Statistics 31:151-124.

70. Dupuis Nathan Fellowes (1893) Elements of Synthetic Solid Geometry. Macmillan.

71. Moulai-Khatir Djezouli, Pairel Eric, Favreliere Hugues (2018) Influence of the probing definition on the flatness measurement. International Journal of Metrology and Quality Engineering 9:15.

72. Shamos Michael (1978) Computational Geometry. Ph.D. thesis, Yale University,

73. Freeman H., Shapira R. (1975) Determining the minimum-area encasing rectangle for an arbitrary closed curve. Communications of the ACM 18 (7):409-413.

74. Toussaint Godfried Solving geometric problems with the rotating calipers. In: IEEE MELECON, Greece, 1983.

75. Timos Sellis, Nick Roussopoulos, Christos Faloutsos (2018) The R+-Tree: A Dynamic Index for Multi-Dimensional Objects. Figshare. doi:10.1184/R1/6610748.V1

76. Beckmann N., Kriegel H. P., Schneider R., Seeger B. (1990) The R-tree: An Efficient and Robust Access Method for Points and Rectangles. ACM SIGMOD Record 19 (2):322-331.

77. Roussopoulos Nick, Leifker Daniel (1985) Direct spatial search on pictorial databases using packed R-trees. doi:10.1145/318898.318900

78. Gottschalk S., Lin M. C., Manocha D. (1996) OBB tree: A hierarchical structure for rapid interference detection. doi:10.1145/237170.237244

79. Dimitrov Darko, Knauer Christian, Kriegel Klaus, Rote G. (2007) New upper bounds on the quality of the PCA bounding boxes in r2 and r3. doi:10.1145/1247069.1247119

80. O'Rourke Joseph (1985) Finding minimal enclosing boxes. International Journal of Computer & Information Sciences 14 (3):183-199.

81. Bespamyatnikh Sergei, Segal Michael (2000) Covering a set of points by two axis-parallel boxes. Information Processing Letters 75 (3):95-100.

82. Lahanas M., Kemmerer T., Milickovic N., Karouzakis K., Baltas D., Zamboglou N. Optimized bounding boxes for three-dimensional treatment planning in brachytherapy. Medical Physics 27 (10):2333-2342.

83. Barequet Gill, Har-Peled Sariel (2001) Efficiently Approximating the Minimum-Volume Bounding Box of a Point Set in Three Dimensions. Journal of Algorithms 38 (1):91-109.

84. Dimitrov D., Holst M., Knauer C. , Kriegel K. Experimental study of bounding box algorithms. In: The Third International Conference on Computer Graphics Theory and Applications, 2008. pp 15-22.

85. Barber C., Dobkin David, Huhdanpaa Hannu (1996) The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software (TOMS) 22 (4):469-483.

86. Александров П. С., (Aleksandrov P. S.) (1968) Лекции по аналитической геометрии (Lectures of Analitical Geometry). Наука (Science).

87. Leon Steven J. (2010) Linear algebra with applications. 8th ed. edn. Pearson, Upper Saddle River, N.J.

88. Groen F. C. , Verbeek P. W., de Jong N., Klumper J. W. (1981) The smallest box around a package. Pattern Recognition 14 (1):173-178.

Part II

Papers

Paper 1

Robust estimation of optimal sample size for CMM measurements with statistical tolerance limits

Proceedings of the ICMSC 2018. The 2nd International Conference on Mechanical, System and Control Engineering, MATEC Web of Conferences, Volume 220

Robust Estimation of Optimal Sample Size for CMM Measurements with Statistical Tolerance Limits

Petr Chelishchev[1], Aleksandr Popov[2] and Knut Sørby[1]

[1]*Department of Mechanical and Industrial Engineering, NTNU, NO-7491 Trondheim, Norway*
[2]*Department of Mathematics, Baltic State Technical University, St. Petersburg, Russia*

Abstract. The paper proposes the kernel probability density function approach to estimate the distribution of measurements on a part which is measured in a coordinate measuring machine (CMM). The study is based on the experimental data derived from internal cylinder measurements. The distribution free model suggested by Wilks was used as a reference for the selection of the sample size. Three cross sections of a cylinder were measured regarding to this reference. The work defines the minimum required sample size for obtaining at least 0.95 proportion of radius variation for particular studied cylindrical part with 95% confidence level.

1 Introduction

The main goal of Geometric Dimensioning and Tolerancing (GD&T) inspection of a part is to assess if the geometry and dimensions of the part are inside of the specified tolerance limits to verify that an assembly fits, or that the intended functionality of the part is guaranteed. This paper deals with verification of radius size variation of internal cylinder.

Coordinate measuring machines (CMMs) are universal and widely employed automated measuring systems in industry [1]. One of the most critical parameter of measuring strategy with CMM is the number of measuring points that is used to extract data from the part features. Obviously, a greater number of measuring points provides a better accuracy, however it leads to higher time consumption and costs. Since the accuracy requirement in a design specification is defined by the tolerance interval, within which the part dimension or geometry may vary, then evidently, it should exist a certain number of points sufficient enough to confirm with some given probability if the size is inside of the tolerance limits or not.

The influence of sample size on the measurement result has been widely discussed, and several different approaches has been used to estimate the contribution to the measurement uncertainty. Approaches such as statistical methods [2,3] (for normal distribution), fuzzy logic [4], genetic algorithm [5], extended zone model optimization [6], adaptive sample strategy with use of Kriging models [7] and analytical methods with implementation uncertainty simulations [8, 9] have been suggested. However, a standard guide or criterion for sample strategy with CMM GD&T inspection has not been established yet.

According to [10] "*a statistical tolerance interval is an estimated interval, based on a sample, which can be asserted with confidence level* $1-\alpha$ *to contain at least a specified proportion p of the items in the population. The limits of a statistical tolerance interval are called statistical tolerance limits.*" Theoretically, the statistical tolerance interval for the case of normal distributed data set is the most well developed method [11, 12]. There are tabulated data in international standard ISO16269-6 for calculating both one-side and two-sided statistical tolerance intervals for sample size $n \geq 2$ and the at least population proportion p for such confidence levels $100(1-\alpha)\%$ as 90%, 95%, 99%, 99.9%. However, if other tolerance intervals (in form of $\pm k\sigma$) need to be found, which are not provided by the standard, for example other p and/or not typical $1-\alpha$ value, or other sample size, then the *K.factor* function from the "*tolerance*" package in *R* programming language may be employed to calculate the factor k .

This paper provides an approach for estimation of an optimal number of the measuring points for a two-sided statistical tolerance interval based on a distribution-free model. A continuous probability density function (pdf) from measurements from a workpiece was approximated by kernel density estimator (KDE). The estimated continuous pdf was further used to simulate different sample strategies and evaluation of the confidence level for detecting at least 0.95 content of total radius variation range.

2 Data and experimental study

An internal cylindrical hole of an aluminum workpiece produced by a turning operation with nominal diameter 60 mm and length 130 mm was inspected in a CMM (Leitz PMM-C-600) with an analogue probe to detect a largest possible deviation of radius variable. The cylinder axis was aligned with the vertical axis (z axis) of the CMM. Three cross sections of the cylinder are measured: the first close to the top (Section A), the second in the middle (Section B), and the third on the bottom (Section C).

Table 1. Shapiro-Wilk normality test for 473 points sample

Sections	***Section A***	***Section B***	***Section C***
P-value	5.392e-07	0.007842	0.01435

There are uniformly distributed 473 points coordinates (x_i, y_i) measured around the circle in each section, and a least squares circle (LSC) method was utilized to calculate the circle centre. The radius variable r_i, for each measured point was calculated by:

$$r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \qquad (1)$$

The uncertainty of the CMM itself is about 10 times less than inspected radius variance range, and thus it is not considered in the analysis of sample size.

Intuitively it is clear that less number of points may provide lower measurement accuracy due to the probability that extreme points on the feature are missing in the extracted data set. We have used MATLAB source for our simulation approach to investigate the degree of influence of sample size on the inspection confidence level and the detected radius variation range.

It is always advisable to evaluate the normality of a distribution of the original data set in the very beginning. The Shapiro-Wilk normality test [13] was applied to the measured data sets by use of the *shapiro.test* function in the R programming language. The results are shown in Table 1. The extremely lower p-value (especially Section A and Section B) yields us a reason to reject the assumption about normal distribution of the measurements.

2.1 Distribution-free model

The distribution of the radius variable of the part is not known before we start the measurements. We will therefore suggest to use the Wilks criterion [14, 15] to define the minimum sample size. The criterion is based on the order statistic. It postulates the following: *if an investigated random characteristic belongs to a population of any unknown continuous distribution function, then at least a content p of the population included between the smallest observation* $r_{\min}$ *and the largest observation* $r_{\max}$ *of the data sample with confidence level* $(1-\alpha)$, *and a required minimum sample size* $n_{\min}$. For the two-sided tolerance interval with the conditions determined above, can be expressed by following [10]:

$$n_{min} \cdot p^{n_{min}-1} - (n_{min} - 1) \cdot p^{n_{min}} \le \alpha \,. \qquad (2)$$

Results computed by (2) of minimal sample size for the two-sided statistical tolerance limits (between the first and *n*-th order of sample order statistic) with unknown continuous distribution, and predefined $(1-\alpha)$ and p, are shown in Table 2. As long as the number of measuring points supposed to be the natural numbers ($n_{\min} \in ¥$), negative solutions were not considered, and all results of Table 2 were rounded to the nearest upper integer. This particular fact together with the distribution independency of (2) provides a robust property of the method, which is further going to be confirmed by the expirement data and a simulation model.

Table 2. Minimal sample size $n_{\min}$ for proportion p and confidence level $1-\alpha$

Confidence level , $100(1-\alpha)\%$	**Proportion of population, *p***				
	0.500	0.750	0.900	0.950	0.990
50	3	7	17	34	168
75	5	10	27	53	269
90	7	15	38	77	388
95	8	18	46	93	473
99	11	24	64	130	662

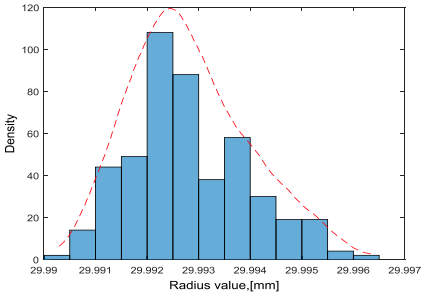


Figure 1. A histogram and kernel density estmate of *f(r)* for Section A with sample size 473 points

Regarding to the above description for the two-sided statistical tolerance interval with $p = 0.99$ and $100 \cdot (1-\alpha)\% = 95\,\%$ (the actual confidence level is 95.02%), the minimum sample size is 473 (Table 2). That is the total number of measuring points we used in our inspection of the cylinder sections.

2.2 Kernel density estimation

In practice, the data distribution is often unknown and/or may contain outliers. Hence, it is reasonable to estimate the tolerance intervals based on more general assumptions when it is impossible to describe sample data with any known standard distribution functions. A possible solution in such case is an estimation of the pdf directly from the measured data sample. A non-parametric statistic may be used in this way. One possibility to do that is the well-known histogram technique. However, the histogram suggests a distribution

interpretation only in form of bins and it is less useful for further appliance due to lack of continuity. Meanwhile a limited number of known pdf $f(r)$ are available to describe a continuous-valued random variable (logarithmic, exponential and so on). To avoid such restrictive assumptions about the form of $f(r)$ the KDE may be applied [16]. Further, using the kernel estimator based on original measured data, an opportunity appears to generate any different random data samples regarding to the initial data distribution.

2.3 Kernels and weighting function

Similar to the histogram we need an estimator of $f(r)$. The probability that random variable is within of the interval $r \pm b$ can be written as following:

$$P(r-b<R<r+b)=\int_{r-b}^{r+b} f(\delta)d\delta \approx 2bf(r), \quad (3)$$

and hence

$$f(r) \approx \frac{1}{2b} P(r-b<R<r+b). \quad (4)$$

Alternatively, the frequency for the given interval could be estimated by the equation:

$$\hat{f}(r)=\frac{1}{n}\sum_{i=1}^{n} w(r-r_i,b), \quad -\infty<r<\infty \quad (5)$$

where the estimator $\hat{f}(r)$ has the properties of a pdf, i.e. positive for any r and an integral area equal to 1. Then a weighting function $w(\delta,b)$ can be generalized in this way:

$$w(\delta,b)=\frac{1}{b}K\left(\frac{\delta}{b}\right), \quad (6)$$

where b is the *bandwidth* or *smoothing* constant of weighting function and K is the standardized weighting function (with $b=1$), which is the *kernel*.

2.4 Kernels' parameters

The degree of smoothing of $\hat{f}(r)$ depends on parameters of $w(\delta,b)$ such as the kernel K and the bandwidth b, which determine a shape and a width of the weighting function respectively. The proper choice of K and b is a subject of an optimization problem.

The accuracy of the kernel density estimator can be evaluated by mean squared error (MSE), mean integrated squared error (MISE) and asymptotic mean integrated squared error (AMISE).

According to previous research [17] *Epanechnikov* function was defined as the optimal kernel in respect to $MISE(\hat{f})$:

$$K(\delta)=\begin{cases}\frac{3}{4\sqrt{5}}\left(1-\frac{1}{5}\delta^2\right) & for\ |\delta|<\sqrt{5}\\ 0 & otherwise.\end{cases} \quad (7)$$

The bandwidth b depends on different factors e.g. unknown pdf $f(r)$, kernel type, number of observations in the sample and so on. There are a number of methods available to optimize the bandwidth parameter such like bias cross-validation (BCV), unbiased cross-validation (UCV), direct plug-in rule (DPI) and others. The methods can have a different performance dependently on the estimation function $\hat{f}(r)$ used and the pdf $f(r)$ estimated. Thus, we use *Epanechnikov* kernel and default MATLAB bandwidth estimation in this study.

2.5 Date estimation by kernel function

The radius variable r_i used in simulation was computed by (1) with assumption of a unique circle centre, which was obtained from LSC based on 473 measurements points. The rounding of the values by one decimal place ($1\cdot10^{-4+1}$ mm), on the one hand allows considering the cylinder form tendency and possible outliers, and on the other hand do not take into account unnecessary accuracy requirements to the data estimated circle centre coordinates.

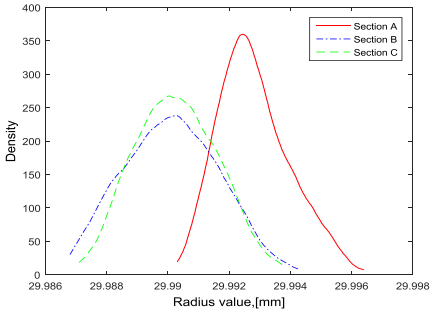


Figure 2. Kernel estimates $\hat{f}_A(r)$, $\hat{f}_B(r)$, $\hat{f}_C(r)$ for three sections based on 473 points sample size machine.

Estimates of pdf $f(r)$ for the three cylinder sections based on Epanechnikov kernel and the sample size 473 measuring points are shown on Fig. 2. Observing the curves one can notice a distinction of the distribution parameters such as the mean values, the variations and the data spread for the cross-sections, which belong to the same cylinder.

All these mentioned facts together with rounding of the sample point numbers give us the reason to presume that the small centre coordinate offsets can be neglected. The robustness of the simulation model is discussed in the next sections.

2.6 Estimation of an optimal sample size

In order to discover an optimal sample strategy for the inspection of the part, a statistical simulation was carried out in MATLAB, by using the KDE $\hat{f}_A(r)$, $\hat{f}_B(r)$, $\hat{f}_C(r)$, estimated from the CMM measurements of workpiece, see Fig. 2.

The eight initially predefined different sample sizes $n=\{5; 10; 15; 30; 60; 90; 93; 95\}$ were simulated with 10^5 iterations for each sample size n_i. The maximum r_{max} and the minimum r_{min} values were detected for every new generated sample. The population content p for the each iteration was evaluated as a difference of the

Table 3. The statistical test simulation with 10^5 iterations for each n_i sample size

Section A								
Sample size n	5	10	15	30	60	90	93	95
Probability P%, (p≥0.95)[a]	2.2	8.5	16.8	44.9	81.2	94.5	95.3	95.5
Section B								
Sample size n	5	10	15	30	60	90	93	95
Probability P%, (p≥0.95)	2.1	8.5	16.4	44.3	80.9	94	95	95.4
Section C								
Sample size n	5	10	15	30	60	90	93	95
Probability P%, (p≥0.95)	1.9	8.5	18	45.1	81.3	93.6	94.9	95.5

a. Probability P of covering a range with population content p

cumulative distribution functions (cdf) of maximum and minimum random variable r_i, based on KDE. Then conditions of equality/exciding 0.95 of total radius variation range was tested by:

$$p(r_{\min} < r < r_{\max}) = F_R(r_{\max}) - F_R(r_{\min}) \geq 0.95 \ , \quad (8)$$

where $F_R(r)$ is cdf of a real-valued (either maximum or minimum) random variable r_i, calculated with the kernel pdf estimator $\hat{f}(r)$. The number of successful iterations was assigned as 1 (or 0 – otherwise) and summed up as Sum_N. The final probability $P\%$ for each sample size n_i was calculated as a rate of Sum_N / M, where M is the total iteration number.

In spite of the predefined initial sample size (473 points) taken from the Table 2 with given $p = 0.99$ and $100(1-\alpha)\% = 95\%$, the total area under the estimated kernel function is equal to 1 (from pdf properties). That gives as an opportunity to generate any size of data sample even larger than the initial sample size. This in turn leads to independency of our simulated results presented in Table 3 on the initial parameters of the model (2).

3 Discussion of results

Analysis of measurements obtained from cylinder sections shows that parts produced by a turning operation has unknown non-normal distribution of radius variables. The parameters (e.g. mean, variance, skewness) of the distributions in difference sections of the cylinder are apart from each other. However, the sample strategy according to table 3 for the sections is finally the same. Thus, the applied simulation model based on the experimental data confirms the robustness property of the method proposed by (2) in section 1. For example, Table 3 shows that the optimal sample size is about 93 points for all sections, which agrees with the data in Table 2 (for $100(1-\alpha)\% = 95\%$, $p = 0.950$). Thereby the simulation based on distributions estimated by kernel function confirms Wilks model given by equation (2). We can also notice that the probability to estimate at least 0.95 fraction of radius variation range is only about 2% in the case of 5 points sample.

This fact gives us the reason to expect that further research of cylindrical parts with larger radius values, from other machine operations and different materials of workpiece, most likely will provide the similar results.

In the simulated model, the circle centre is assumed the same for any data samples. For different sample points the centre point would vary, but maximum possible range between $r_{\min}$ and $r_{\max}$ remains a similar. In addition, the minimum number of points $n_{\min}$ (Table 2) was rounded to the nearest upper integer, thus it makes negligible the influence of the centre coordinates. Again, a good compliance of the simulation results with the distribution free model given in formula (2) demonstrate the insignificant influence of the assumptions about the centre of coordinates. That also confirm that the simulation model itself employs a robust principle. Namely, the identical optimal number of points (Table 3) for different sections with observably diverse distributions improves this statement.

4 Conclusion

The innovation of this work is to show the possibility to use the distribution free model (2) to predict the sample size, its least content and confidence level for GD&T inspection with CMM before any measurements are performed.

In addition, the simulation model for robust estimation of the optimal sample size based on the experimental measuring date has been developed. The provided simulation procedure allows evaluating the sample sizes and their confidence levels for real cylindrical components in industry independently of their dimensions and machining process accuracy. Moreover, the finite sample sizes, which often used in industry, were evaluated. The obtained results demonstrate the particular low confidence level especially for the sample sizes from 5 to 30 measuring points.

The inspection sample size in production is often defined with cost and time-consumption in mind, and thereby it could be too small. The applied technique provide the demanded guidance criteria based on the confidence level and the real data distribution for

choosing the proper measuring sample strategy for GD&T inspection with CMM in manufacture.

In solving of a practical problem, it is recommended to evaluate the distribution of the initial data in the very beginning. If the data distribution is close to the normal distribution then the standardized procedure (ISO16269-6) can be used to estimate the tolerance interval limits. Otherwise, the original distribution based on the CMM measurements with predefined confidence level $1-\alpha$, the variation proportion p and the minimum sample size $n_{\min}$ should be estimated by (2). Then the smallest and the largest order statistics of the sample should be used as the tolerance limits.

References

1. Chiffre, L.D., *Geometrical Metrology and Machine Testing*, DTU Mechanical Engineering, (2011)
2. Jiang, B. and S.-D. Chiu, *Form tolerance-based measurement points determination with CMM*, Journal of Intelligent Manufacturing, **13**(2): p. 101-108, (2002)
3. Hong-Tznong Yau, C.-H.M., *An automated dimensional inspection environment,* (1992)
4. N. Cappetti, A.N., F. Villecco, *Fuzzy approach to measures correction on Coordinate Measuring Machines: The case of hole-diameter verification*, Measurement, **93** (Elsevier Ltd.): p. 41-47, (2016)
5. Changcai Cui, S.F., Fugui Huang, *Research on the uncertainties from different form error evaluation methods by CMM sampling,* Int J Adv Manuf Technol, **43**: p. 136-145, (2008)
6. K. D. Summerhays, R.P. Henke, J. M. Baldwin, R. M. Cassou and C. W. Brown, *Optimizing discrete point sample patterns and measurement data analysis on internal cylindrical surfaces with systematic form deviations*, Precision Engineering, **26**(1): p. 105-121, (2002)
7. G. Barbato, E.M.B., P. Pedone, D. Romano, G. Vicario, *Sampling point sequential determination by kriging for tolerance verification with CMM*, in Proceedings of the 9th Biennial ASME Conference on Engineering Systems Design and Analysis, ESDA08, ASME: Israel. p. 10, (2008)
8. S. Ruffa, G.D. Panciani, F. Ricci, and G. Vicario, *Assessing measurement uncertainty in CMM measurements: comparison of different approaches*, Int.J.Metrol. Qual. Eng. 4, p. 163-168, (2013)
9. A.Weckenmann, M. Knauer, H. Kunzmann, *The influence of measurement strategy on the uncertainty of CMM measurements*, **47**: p. 451-454, (1998)
10. ISO16269-6, *Determination of statistical tolerance intervals*, in Statistical interpretation of data, (2014)
11. I. Janiga, I. Garaj, V. Witkovský, *On Exact Two-Sided Statistical Tolerance Intervals for normal distributions with unknown means and unknown common variability*, Journal of Mathematics and Technology, **3**: p. 25-32, (2012)
12. Janiga, I., Garaj, I., *On exact two-sided statistical tolerance intervals for normal distributions with unknown means and unknow common variability*, In 2009 Quality and Productivity Research Conference, Yorktown Heights, Nerw York: IBM Thomas J. Watson Research Center, (2009)
13. Royston, P., *An extension of Shapiro and Wilk's W test for normality to large samples*, Applied Statistics, **31**: p. 151-124, (1982)
14. Wilks S.S., *Determination of Sample Sizes for Setting Tolerance Limits,* AMS Ann. Math. Statist., **12**: p. 91-96, (1941)
15. Wilks S.S., *Statistical Prediction with Special Reference to the Problem of Tolerance Limits*, AMS Ann. Math. Statist., **13**: p. 400-409, (1942)
16. Alexandre B.T., *Introduction to Nonparametric Estimation,* New York: Spring-Verlag, (2009)
17. Wand M.P., Jones M.C., *Kernel Smoothing,* Monographs on statistics and applied probability, Boca Raton, FlaChapman&Hall/CRC, (1995)

Paper 2

An investigation of outlier detection procedures for CMM measurement data

Proceedings of the ICMSC 2018. The 2nd International Conference on Mechanical, System and Control Engineering, MATEC Web of Conferences, Volume 220

MATEC Web of Conferences **220**, 04002 (2018) https://doi.org/10.1051/matecconf/201822004002
ICMSC 2018

An investigation of Outlier Detection Procedures for CMM Measurement Data

Petr Chelishchev[1], Aleksandr Popov[2] and Knut Sørby[1]

[1]*Department of Mechanical and Industrial Engineering, NTNU, NO-7491 Trondheim, Norway*
[2]*Department of Mathematics, Baltic State Technical University, St. Petersburg, Russia*

Abstract. The paper analyses methods for outlier detection in dimensional measurement. The cross sections of an internal cylinder were inspected by CMM (coordinate measuring machine), and received data sets were employed for further investigation. The efficiency of Rosner's and Grubbs' methods for excluding outliers from the measuring data had been estimated. The method of Rosner had been defined as the most effective for this case study. The simulation results were confirmed by experimental verification.

1 Introduction

The purpose of this work is to analyze the efficiency of outlier test procedures for particular type of data sets received from inspection with CMM (coordinate measuring machine). The following inspection conditions are considered:

- Varying sample size of measurements;
- Unknown number of outliers presented in the sample;
- A spectrum of different distributions of original data sets with unknown dispersion.

In the CMM inspection of the geometrical characteristics of components, the outliers are not necessarily incorrect measurements. The existence of an outlier could indicate that a further investigation of manufacturing processes, measurements procedure, or data analysis methods themselves is required.

The estimation of different statistical parameters (e.g. sample standard deviation, sample mean and so on) may be affected by outlier presence in the measuring data. As a result, it can lead to the invalid estimation of a confidence interval and inflate the random uncertainty estimates as well, thus a good component may be erroneously rejected. That especially yields the particular case, when contact fit methods such as MIC (maximum inscribed circle) or MZ (minimum-zone) are utilized, which are based on the most extreme points and hence very sensitive to outliers.

Outliers are extreme observations, which stay apart from the majority of other measurements. In a simple case, when only one outlier is presented, its inconsistence can be easily observed with respect to the rest of the data. However, when a group of outliers is present, it is difficult to detect them because of the masking effect, which will be described below. At the same time, an incorrect assumption about the original data distribution may lead to confusion of valid observations with outliers. According to ISO 16269-4 [1], the main causes for outliers are the following:

- a measurement or recording error (imprecise or/and incorrect);
- a distribution contamination (one or more contaminating distributions);
- an incorrect distributional assumption;
- rare observations (extreme observations from heavy-tailed original distribution).

In the particular case of measuring in manufacturing conditions, a contamination of a part surface is a frequent cause of outliers, even after attempt of surface cleaning.

In addition, a masking and a swamping effect can occur during the data analysis with parametric statistical test. The masking effect can happen when too few outliers are specified in the outlier detection procedure. Then the test performance can be influenced by the other outliers and as result, no outliers will be detected. On the other hand, if too many outliers are specified in the parameters of outlier test, then some valid observations can be incorrectly labeled as outliers, which is the so-called swamping effect. Therefore, to make a correct decision whether suspected observations are outliers or not can be a complicated task.

2 Methods

2.1 Graphical methods

The first step, before any analytical outlier detection algorithms are applied it is a visual analysis of measurement data. There are a number of graphical methods available such as histogram, scatter diagram, dot

plot and so on [2]. The pox plot become a very popular descriptive tool to reveal the most suspected measurements [3]. In fact, the box plot is a hybrid based on both a model and the graphical method. The graphical interpretation of data helps to choose the most appropriate analytical algorithm i.e. identify whether a single outlier or a group of outliers are present in order to prevent an influence of the masking or swamping effects, as described in previous section.

There are six data sets (with 475 observations in each) comparing with each other on Fig.1. The data sets denoted by A1, B1, C1 represent the first measurement results with outliers. After the contamination was physically removed from the workpiece surface, the measurements at the same sections and with the same point distribution were repeated with CMM. These data sets are denoted as A2, B2, C2 on Fig.1. The box plot gives a good demonstration of the influence on statistical parameters such as the sample mean, median, skewness, data spread, IQR (interquartile range). The relative displacement of these parameters can be easily observed.

The lower and upper fences (lower and upper outlier cut-off) is expressed by following:

$$LF = q_1 - w(q_3 - q_1) \tag{1.1}$$

$$UF = q_3 + w(q_3 - q_1), \tag{1.2}$$

where q_1, q_3 are the first (lower) and the second (upper) quartiles of data sample, and w is the significant factor [4]. The extreme points, which are outside of these fences, are indicated by red dots. For example, with significant factor $w = 1.5$ red dots can be classified as suspected outliers (Fig.1, left), with $w = 3$ as extreme outliers (Fig.1, right) [5]. The vertical box represents IQR of the data, the different between the lower and upper quartiles. Thus, we do not have extreme outliers in the studied case (Fig. 1, right), but there are some suspected observations in all sections. The section A represents the case with multiple potential outliers, section B with two, and section C with a single potential outlier (Fig.1, left). From now on, we can precede with selection of the most suitable outlier detection analytical algorithm for our particular problem.

2.2 Analytical algorithms

There are many outlier methods proposed in the last decade [6]. The difference between them can shortly formulated by following.

- What a sample size can method be applied for (only for small, only for large, or both)?
- How strict is a requirement to the distribution of data set?
- Can method be exploited whether for a single or multiple outliers?
- In a case of multiple outliers method, it is either necessary to provide exact number of outliers or only an upper amount.

Two suitable outlier methods according to these conditions are considered in this research. These methods are Grubbs and Rosner/GESD (Generalized Extreme Studentized Deviate) tests, which are recommended by ISO [1, 7]. Both methods are based on an estimation of a distance deviation from the sample mean with assumption about an approximately normal distribution. The strictness of this normality assumption is examined in this paper.

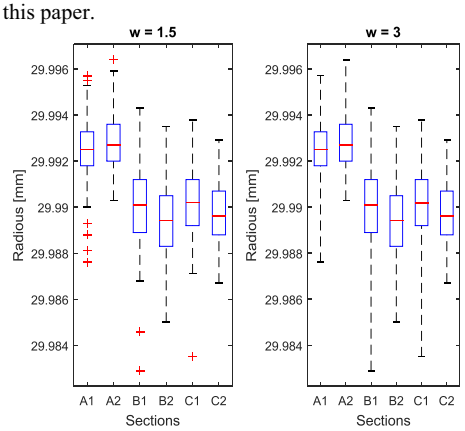


Figure 1. Boxplot of CMM measurements with and without suspected observations.

2.2.1 Grubbs method

The Grubbs method is used to determine a single outlier in a normally distributed data set and can be utilized as sequentially outlier detection procedure for multiple outliers [7, 8]. It tests two types of hypothesis: null hypothesis H_0 – no outliers in the sample, alternative H_1 – the sample has a one outlier. The test statistic for two-sided case computed by (2):

$$G = \frac{\max|\rho_i - \bar{\rho}|}{s} \tag{2}$$

where $\bar{\rho}$ is the sample mean and s is the sample standard deviation. The G statistic shows how many standard deviations are in an absolute distance of an individual observation from the sample mean. The null hypothesis must be rejected with a significance level α, if the following condition is met:

$$G > \frac{n-1}{\sqrt{n}} \sqrt{\frac{\left(t_{\left(\frac{\alpha}{2n}, n-2\right)}\right)^2}{n-2+\left(t_{\left(\frac{\alpha}{2n}, n-2\right)}\right)^2}} \tag{3}$$

where $t_{\left(\frac{\alpha}{2n}, n-2\right)}$ - the Student's quantile given at probability $\frac{\alpha}{2n}$ and $n-2$ degrees of freedom in the data set with number of observations n. One of the weaknesses of the method is the influence by the masking effect.

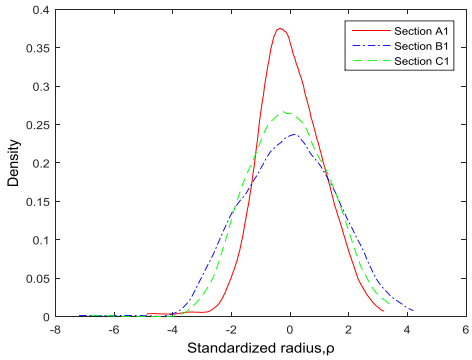


Figure 2. Kernel estimates $\hat{f}_{A1}(\rho), \hat{f}_{B1}(\rho), \hat{f}_{C1}(\rho)$ based on 475 measuring points with outliers.

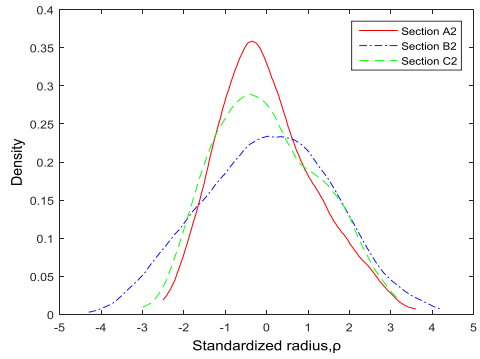


Figure 3. Kernel estimates of $\hat{f}_{A2}(\rho), \hat{f}_{B2}(\rho), \hat{f}_{C2}(\rho)$ based on the repeated 475 measuring points after surface cleaning.

2.2.2 Rosner method

The Rosner (GESD) method is exploiting for detection of single and multiple outliers in nearly normal distributed data, when exact number of outliers are unknown. The only upper limit m of expecting outliers is required to indicate.

In order to avoid the masking effect m should not be chosen too small. There are two hypothesis types: null hypothesis H_0 – no outliers in the sample, alternative hypothesis H_1 – the sample has up to m outliers. For two sided case, the ESD test statistic is computed as a following [9]:

$$R_i = \frac{\max_i |\rho_i - \bar{\rho}|}{s}, \qquad (4)$$

where $\bar{\rho}$ and s are the sample mean and the standard deviation, respectively. Excluding one observation, which maximized $|\rho_i - \bar{\rho}|$, the test statistic in (4) is recalculated again for $n-1$ data sample. This procedure repeats m times until all extreme measurements are removed from the data set. The output of this computation will be the array of $R_1, R_2, \ldots R_m$. Then the critical value k_i for each single element of the vector R_i is calculated:

$$k_i = \frac{t_{(p,n-i-)}(n-i)}{\sqrt{\left(n-i-1+\left(t_{(p,n-i-1)}\right)^2\right)(n-i+1)}}, i = 1,2,\ldots m \quad (5)$$

where $t_{(p,\nu)}$ is quantile of Student's distribution with degrees ν of freedom and probability $p = 1 - \frac{\alpha/2}{n-i+1}$. Thus the total number of outliers is the largest i such that $R_i > k_i$. Opposite to Grubbs test, GESD can be influenced by the swamping effect (described above), but the influence of the masking effect is relatively neglected.

Table 1. Efficiency rate of outlier detection (randomly located outliers, medium values, 100 observations)

Number of outliers	Method	Section A $\bar{\rho}_m^A = 4.49$	Section B $\bar{\rho}_m^B = 6.00$	Section C $\bar{\rho}_m^C = 4.94$
1	Grubbs	0.66	0.65	0.67
	Rosner	0.66	0.65	0.67
2	Grubbs	0.36	0.33	0.34
	Rosner	0.60	0.58	0.61
3	Grubbs	0.12	0.09	0.09
	Rosner	0.58	0.57	0.58
4	Grubbs	0.02	0.01	0.01
	Rosner	0.57	0.55	0.57

Table 2. Efficiency rate of outlier detection (randomly located outliers, large values, 100 observations)

Number of outliers	Method	Section A $\bar{\rho}_l^A = 5.2$	Section B $\bar{\rho}_l^B = 6.9$	Section C $\bar{\rho}_l^C = 5.7$
1	Grubbs	0.99	0.99	1.00
	Rosner	0.99	0.99	1.00
2	Grubbs	0.95	0.94	0.96
	Rosner	0.99	0.99	1.00
3	Grubbs	0.77	0.72	0.76
	Rosner	0.99	0.99	0.99
4	Grubbs	0.37	0.32	0.34
	Rosner	0.99	0.99	0.99

3 Data simulation and case study

The data sets used in this study were derived from CMM (Leitz PMM-C-600) measurements (x_i, y_i) taken from three cross-sections (A, B, C) of an internal cylindrical surface. The cylinder axis was aligned with z axis. The circle center coordinates (x_c, y_c) for each section were estimated with LSC (least squired circle) method by PC-DMIS software based on 475 measured points. Then the radius variable for each measured point r_i was calculated by:

$$r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \quad . \qquad (6)$$

For practical convenience, the result data arrays of r_i were standardized by $\rho_i = (\bar{r} - r_i) \times 1000$, where $\bar{r}$ is

the average radius in the cross section. The standardized radius variable ρ_i were further used in simulation tests.

The data sets of repeated measurements A2, B2, and C2 were tested for normality distribution with Anderson-Darling method [10]. This method is more sensitive to outliers and especially effective for detecting any departure in the tails of data distribution. Only one of three data sets (p-values: 0.001, 0.138, 0.001 for A2, B2, C2 correspondently) had p-value over specified significance level 0.05, thus the data are quite unlikely from a population with normal distribution.

In order to obtain a form distribution of $f(\rho)$ for standardized variable ρ_i, the kernel density estimator $\hat{f}(\rho)$ was applied [11]:

$$\hat{f}(\rho) = \frac{1}{bn}\sum_{i=1}^{n} K\left(\frac{\rho-\rho_i}{b}\right), \quad -\infty < \rho < \infty \qquad (7)$$

where K is the kernel smoothing function, b is a bandwidth and n is the sample size. Epanechnikov kernel was used as the smoothing function K with default MATLAB bandwidth b and the sample size n with 475 observations. The estimates of pdf (probability density function) for sections A1, B1, and C1 with outliers are illustrated on Fig. 3, and Fig. 4 for repeated measurements of the same sections A2, B2, and C2, after the outlier issue was physically eliminated (no analytical algorithm were used so far).

The estimated pdf objects $\hat{f}_{A2}(\rho), \hat{f}_{B2}(\rho), \hat{f}_{C2}(\rho)$ were further used to generate random data samples to simulate workpiece measurements without outliers (Fig. 4). In addition, some of the data points were replaced by simulated outliers. The simulation of outliers was based on a uniform distribution around a specified deviation from the mean value of the random data sample. The effectiveness of outlier detection of the Grubbs and Rosner methods with different combination of correlated factors were estimated from 10^5 iterations with summing two possible results: 0 – failure; 1 – success. In order to meet success requirements the same number of outliers with identical indexes must be detected (e.g. if only three outliers from four detected correctly then result is considered as a failure). The efficiencies e_G, e_R (Grubbs and Rosner method, respectively) were estimated simultaneously as a rate of number of success iteration Sum_G, Sum_R to total iteration number M, then $e_G = Sum_G / M$ and $e_R = Sum_R / M$.

The following factors were considered in the test of the efficiency of outlier detection procedures:

- non-normal distribution of random data samples;
- size variation of the random data samples;
- outliers randomly distributed around a mean value of the random data sample with specified deviation values;
- a defined number of outliers in each data set (from 1 to 4);
- outliers as randomly distributed data points or as a block of data points.

The more detail description of these factors is given in the next section.

Table 3. Efficiency rate of outlier detection (located as a block, large values, 100 observations)

Number of outliers	Method	Section A $\bar{\rho}_l^A = 5.2$	Section B $\bar{\rho}_l^B = 6.9$	Section C $\bar{\rho}_l^C = 5.7$
2	Grubbs	0.95	0.92	0.96
	Rosner	1.00	1	1
3	Grubbs	0.6	0.52	0.55
	Rosner	1.00	0.99	1
4	Grubbs	0.07	0.07	0.06
	Rosner	1.00	0.99	1

Table 4. Efficiency Rate of Outlier Detection for Various Sample Sizes (2 Outliers with Random Locations, Large Values)

Sample size	Method	Section A $\bar{\rho}_l^A = 5.2$	Section B $\bar{\rho}_l^B = 6.9$	Section C $\bar{\rho}_l^C = 5.7$
15	Grubbs	0.07	0.06	0.06
	Rosner	0.75	0.74	0.76
30	Grubbs	0.44	0.42	0.42
	Rosner	0.92	0.92	0.94
60	Grubbs	0.84	0.83	0.85
	Rosner	0.98	0.98	0.99
100	Grubbs	0.95	0.94	0.96
	Rosner	0.99	0.99	1.00

4 Simulation and experiment results

The distribution of outliers in the simulations was based on a *medium* and a *large* deviation from the mean value of the simulated measurements. The medium value for the outliers are generated in the interval $\rho_m \in 3.90s \pm 0.01$ (Table 1) and the large value of the outliers from an interval $\rho_l \in 4.5s \pm 0.1$ (Table 2, 3), where s is the estimated standard deviation of the simulated measurements ($s_A = 1.15$, $s_B = 1.54$, $s_C = 1.27$ for measurement sets A2, B2, and C2, correspondingly). There are different numbers of outliers tested both with randomly distributed locations (Table 1, 2) and with location as a block (Table 3). For the random location, two discrete values, $\pm\rho_m$ and $\pm\rho_l$, were used, while for the block location only negative $-\rho_l$ values were integrated to meet the most typical conditions (associated with contamination). Due to low skewness, the simulation results for $+\rho_l$ were very similar thus, they are not shown here. The simulation results of the influence of the sample size on an outlier detection performance were specified in Table 4. The significance level $\alpha = 0.05$ was applied in all tests. All simulation tests were carried out in MATLAB and results are tabulated in Table 1, 2, 3, and 4. The simulation results were rounded up to the second digit from decimal point.

Both methods were also applied with the experimental measurements. The detected outliers were tabulated in Table 5. The comparing boxplot of data set after removing of outliers (A1*, B1*, C1*) with data samples of repeated measurements (A2, B2, C2) are illustrated on Fig. 4. There are two large outliers were removed in

Table 5. Appliance of the outlier methods with real measurement data (475 observations)

Outliers no	Outlier parameters	Section A1*		Section B1*		Section C1*	
		Rosner	Grubbs	Rosner	Grubbs	Rosner	Grubbs
1	**Index**	60	60	459	459	2	2
	Values [mm]	29.9881	29.9881	29.9829	29.9829	29.9835	29.9835
2	**Index**	61	61	-	-	-	-
	Values [mm]	29.9876	29.9876	-	-	-	-

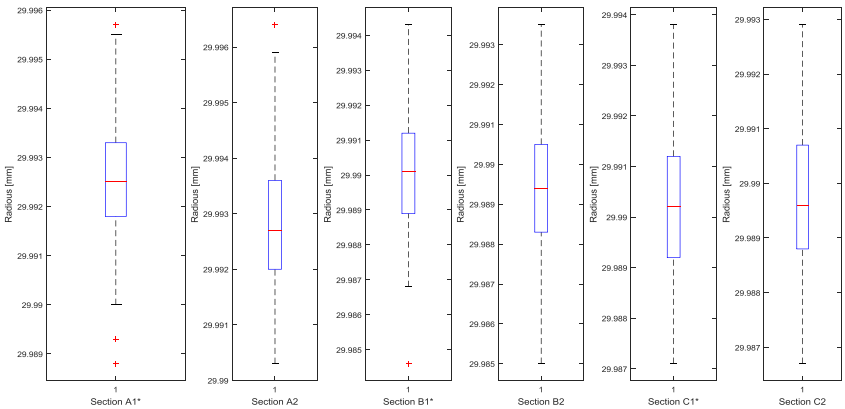


Figure 4. Boxplot of the measuring data: A1*, B1*, C1* - after removing of outliers; the repeated measurements A2, B2, C2 – after cleaning of workpiece.

section A1 with both Rosner and Grubbs tests and by one outlier in section B1 and C1. There were number of medium outliers which not detected by any of the methods (Sections: A1* and B1*) though some of these suspected points disappeared after measurement were repeated (Section A2). This fact confirm the simulation results, which were obtained in Table 1, for medium values of outliers.

5 Discussion

In the case of a single outlier, the Grubbs and Rosner tests have similar performance. For more than two outliers cases there was significant difference in the outlier detection efficiency. Both procedure had a lower efficiency rate for the medium outlier values (Table 1). Therefore, the parametric outlier tests must be used very carefully when small value of outliers are presented. However, Rosner method had at least 0.98 efficiency in whole range 1 – 4 of outliers in case of large outlier value, while the Grubbs method has 0.95 efficiency for two outliers, but even lower for larger number of outliers (Table 2). This is a good demonstration of the influence of the masking effect on the Grubbs procedure and the very low influence on Rosner method.

According to the Table 3, the additional simulation test showed that outliers distribution either as the block or random had no notable influence on Rosner method performance what is opposite to the Grubbs method, which efficiency was fairly lower for the block location of outliers than for random distributed among the sample set. Meanwhile, there was a great influence of the sample size on efficiency rate observed for both of the methods as shown in Table 4. The consecutive outlier detection procedure (Grubbs) had efficiency below 0.5 for sample size with 30 observations or lower, while Rosner's test could provide at least 0.75 efficiency rate even for 15 observations sample. In spite of some differences in distribution, form and variation range between all three data sets the test performance did not distinguished so much within each individual method. That leads us to a conclusion that both methods have no any strict requirements to the normal distribution.

There were no additional tests of masking or swamping effects presented for Rosner method in this paper. It is, however, a well-known fact that too small number of outliers initially applied in the Rosner test (relatively to the actual number of outliers in the sample) can lead to the masking effect. However, when the extra two outliers had been initialized with Rosner test (additionally to the actual number of outliers) during of simulation the swamping effect was not observed.

6 Conclusion

There are many different outlier procedures available for data analysis, but it is a difficult task for an unexperienced operator to choose the most suitable test for a particular problem. The following specific conditions were met for this study with considered methods:

- the ability of the methods work with various sample sizes;

- the ability to detect multi outliers when maximum outlier number is unknown;
- the stable efficiency (over 0.9) to detect the large outliers, which bring the most significant influence;
- the applicability for data from unknown non-normal distributions;
- the stability to the masking and swamping effects.

The outlier detection procedures such as Grubbs and Rosner can be successfully applied even with real workpiece measurements, which are difference from the normal distribution. However, the Rosner method is more reliable and hence preferable. Meanwhile the medium outliers should be double-checked before removing/accepting for further analysis. It is not recommended to use samples below 30 measuring points to avoid the low efficiency outlier detection procedure. The measurement tests conducted with CMM confirm the simulation results and all conclusions above. The research of experimental measurements also revealed that multiple outliers groups can be expected with CMM measurements. Therefore, the automated outlier detection procedure based on the Rosner / GESD method can be effectively applied with a geometry inspection routine.

References

1. ISO16269-4, Detection and treatment of outliers, in Part 4: *Statistical interpretation of data*, (2010)
2. Chambers John, William Cleveland, Beat Kleiner, and Paul Tukey, *Graphical Methods for Data Analysis*, Wadsworth, (1983)
3. Sim C.H., Gan F.F. and Chang T.C., *Outlier Labeling with Boxplot Procedures*, Journal of the American Statistical Association, **100**: p. 642-652, (2005)
4. Langford E., *Quartiles in Elementary Statistics*, Journal of Statistics Education, **14**, (2006)
5. Tukey J.W., *Exploratory data analysis*, Massachusetts: Addison-Wesley, (1977)
6. Barnett V., Lewis T., *Outliers in Statistical data*, ed. 3rd. 1994, New York: Wiley.
7. ISO5725-2, Accuracy of measurement methods and results, in Part 2: *Basic method for the determination of repeatability and reproducibility of a standard measurement method*, (1994)
8. Grubbs F., *Procedures for Detecting Outlying Observations in Samples*. Technometrics, **11**: p. 1-21, (1969)
9. Rosner B., *Percentage Points for a Generalized ESD Many-Outlier Procedure.* Technometrics, **25**: p. 165-172, (1983)
10. Anderson T.W., Darling D.A., *A Test of Goodness-of-Fit*. Journal of the American Statistical Association, **49**: p. 765–769, (1954)
11. Alexandre B.T., *Introduction to Nonparametric Estimation*, New York: Spring-Verlag, (2009)

Paper 3

Optimization of sample size for two-point diameter verification in coordinate measurements

Proceedings of the IWAMA2018. International workshop of Advanced Manufacturing and Automation , Advanced Manufacturing and Automation VIII, Volume 484, p.313-321

Paper 4

Simulation algorithm of sample strategy for CMM based on Neural Network Ap-proach

Proceedings of the IWAMA2019. International workshop of Advanced Manufacturing and Automation , Advanced Manufacturing and Automation IX, Volume 634

Paper 5

Perspectives for appliance and accuracy improvement of coordinate measurements with laser technique

Proceedings of the 2019 IEEE International Conference on Electrical Engineering and Photonics, EExPolytech, p.282-284

Paper 6

Estimation of Minimum Volume of Bounding Box for Geometrical Metrology

Submitted to International Journal of Metrology and Quality Engineering (IJMQE)

Estimation of Minimum Volume of Bounding Box for Geometrical Metrology

Petr Chelishchev,
Knut Sørby
Department of Mechanical and Industrial Engineering, NTNU, NO-7491 Trondheim, Norway
petr.chelishchev@ntnu.no

Abstract. This paper presents algorithms for estimating the smallest volume of a bounding box based on a three-dimensional point set for metrological application. The aim of this work is to investigate method accuracy in order to reduce an unnecessary computation. The algorithms are demonstrated on an artefact measured by a coordinate measurement machine (CMM). The estimation results of reference objects can be used for calibration of dimensional measuring systems. The principles proposed in this paper may also be utilized to improve a software functionality for the measuring systems.

Keywords: minimum volume bounding box / minimum bounding rectangle / convex polyhedron / CMM

1 Introduction

In various applications, it can be useful to circumscribe a given set of three-dimension coordinate points by an ideal shape *rectangular parallelepiped*. It was suggested by Dupuis [1] to use the term *cuboid* when referring to a rectangular parallelepiped. However, in the literature of the computational geometry, the term *box* is commonly associated with the rectangular parallelepiped. In this text, we use the term *side* for the bounding box face. This term may be also used while referring to the physical cuboid object side. The term *face* is mainly used for the inscribed convex polyhedron faces, which are the product of 3D convex hull operation. All six sides (faces) of the box are rectangles and each side is parallel with the *opposite side* and orthogonal with the other four *adjacent sides*. These four *adjacent sides* comprise a "*closed loop*". For example, the *Top* side has a "*closed loop*" of *adjacent sides* that consists of: *Front, Left, Right, and Back*. The opposite, *Bottom* side has the same "*closed loop*" of *adjacent sides* as the *Top* side.

An estimation of the minimal volume bounding box (MVBB) often includes an estimation of the minimal area bounding rectangle (MABR). Both problems are commonly used in computer graphics (e.g. collision detection, optimal layout detection etc.), image processing, medicine (e.g. brachytherapy), metrology, automatic tariffing in goods-traffic and many other applications. Together with other association criteria (e.g. minimum zone, least squares), the minimum volume criterion can be applied for estimation of the flatness deviation of mechanical parts in industry [2]. Depending on an application, the MVBB algorithm may be optimized either for computation time or for measurement accuracy.

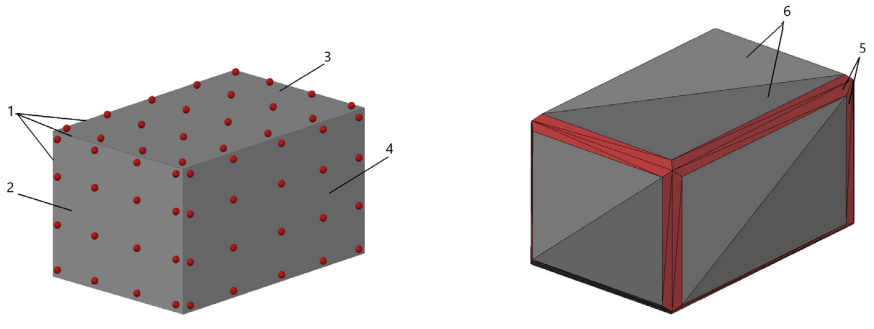
Based on the proposals of Shamos [3], Freeman and Shapira [4], Toussaint presented an elegant unambiguous MABR solution in [5]. This exact solution of the MABR problem has $O(n^2)$ computing time with the use of the rotating caliper algorithm for n-point set in $\mathbb{R}^2$, and $O(n)$ time with the use of two pairs of rotating calipers orthogonal to each other. A number of approximation algorithms and heuristic alternatives are suggested to solve the two-dimension

problem. Among them, the searching algorithms based on the R-tree data structures [6-8] and the principle components [9, 10].

The most exact solution of the MVBB problem for n-point set in $\mathbb{R}^3$ with computation time $O(n^3)$ was provided by O'Rourke [11], which remains the state-of-the-art so far. Alternative approximation algorithms have been developed to reduce the computation time. Bespamyatnikh and Segal [12] suggested an efficient $O(n^2)$ approximation algorithm. A search based on Powell's quadratic convergent method was proposed by Lahanas et al. [13]. Later, Barequet and Har-Peled [14] presented an approximating algorithm with $O(n+1/\varepsilon^{4.5})$ computation time, and a simplified version with $O(n\log n+n/\varepsilon^3)$, where $0<\varepsilon\le 1$. Recently, Dimitrov et al. developed a faster algorithm based on the discrete and the continuous versions of principal component analysis (PCA) [10, 15]. The continuous version guarantees a constant approximation factor but it is still limited by $O(n\log n)$ – time required for computation of a convex hull. The commonly used solutions for MABR and MVBB are based on the convex hull operation [3, 16], in order to reduce the number of considered points and avoid redundant computation.

Some approximation algorithms may provide a large systematic error. The majority of approaches presented above are mainly focused on reducing the computation time, but at the expense of accuracy. In this paper, we consider calculation of the minimum bounding box on reference standards used for calibration of dimensional measuring systems; hence, the accuracy must be ensured. The elegant approach provided by O'Rourke is the accurate solution, but it does not take into account some metrological issues related to the discrete point measurement with CMM, which are discussed below.

The physical edges (denoted by 1 in Figure 1a) of the cuboid object are typically not measured and there is always a distance between the edges and the measured points. As a result, there is an intermediate space between the measured points on all pairs of the *adjacent sides* (e.g. side 2 and side 3, side 3 and side 4 in Figure 1a) of the cuboid object. This intermediate space is transformed into a large number of the convex polyhedron faces after appliance of the convex hull operation. Such newly constructed faces provide acute angles and look similar to "chamfer" faces (denoted by 5 in Figure 1b). These faces cut off the physical cuboid object and they will lead to unnecessary computation in the O'Rourke algorithm. Obviously, these "chamfer" faces cannot be a part of the minimum bounding box solution and these faces should be excluded from the algorithm.



(a) (b)

Figure 1. An example of the metrological issue: a) a cuboid object with CMM measured points; b) an example of the convex polyhedron with the chamfer faces after convex-hull operation; 1 – edges; 2 – left side; 3 – top side; 4 – front side; 5 – "chamfer" polyhedron faces; 6 – ordinary polyhedron faces.

In this paper, we deal with the minimum bounding box problem for physical objects with an actual shape close to the perfectly rectangular bounding box. The proposed algorithms for estimation of MVBB take into account the effect of the "chamfer" faces. The most accurate algorithm searches for the minimum solution according to the conditions defined by two theorems related to the MABR and the MVBB problems presented in section 2. A detailed overview of the three conventional geometrical algorithms suggested by the author are given in section 3. Implementation of the methods is presented in section 4 with description of the experimental setup and computational results.

2 Theoretical Background

The solution of the three-dimension MVBB problem involves the two-dimension case. After the orientation of one side of the bounding box is locked in the MVBB algorithm, all points are projected onto the **xy**-plane, and the orientations of other adjacent sides of the bounding box can be found by the MABR algorithm as the two-dimension problem.

2.1 Minimum-Area Bounding Rectangle

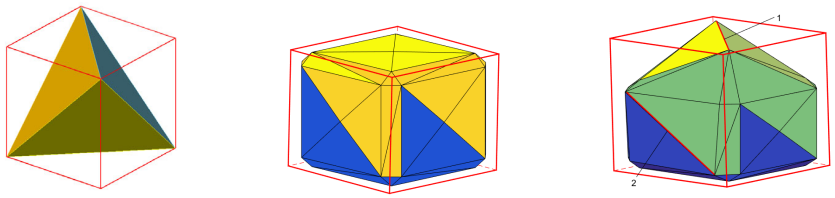
The earliest known solution of the MABR problem was presented by Freeman and Shapira [4]. They presented the following theorem, which is the basis for minimum bounding rectangle algorithms: *The rectangle of minimum area enclosing a convex polygon has a side collinear with one of the edges of the polygon.*

The MABR solution is based on the 2D convex hull operation [3], which is applied as the first step. In the second step, we search for the minimum-area bounding rectangle circumscribing the convex polygon constructed by the convex hull algorithm in the first step. The theorem mentioned above limits the number of bounding rectangles that are candidates for the minimum-area bounding rectangle.

2.2 Minimum-Volume Bounding Box

The second theorem presented here was formulated and proved for the MVBB problem by O'Rourke [11]: *A box of minimal volume circumscribing a convex polyhedron must have at least two adjacent sides flush with edges of the polyhedron.*

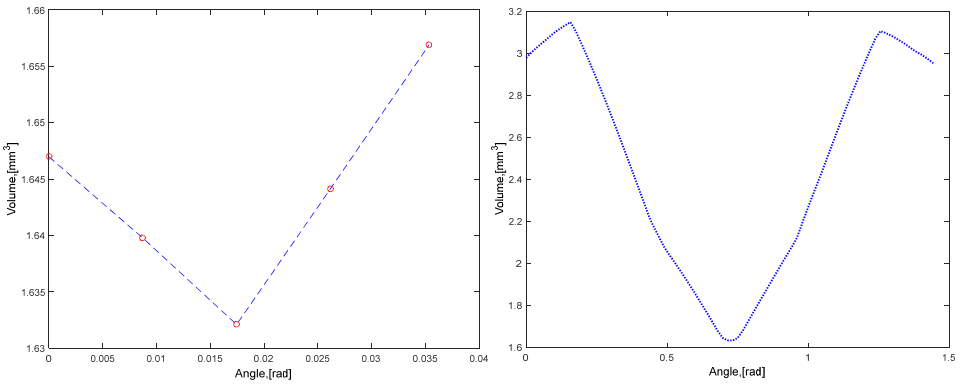
It is not necessary that one of the sides of the bounding box is coplanar with one of the faces of the convex polyhedron. In fact, the bounding box with minimal volume circumscribing a regular tetrahedron has all six sides coplanar with the tetrahedron edges without flushing with any tetrahedron faces (Figure 2a).



(a) (b) (c)

Figure 2. Examples of convex polyhedrons: (a) a regular tetrahedron with edge length $\sqrt{2}$ circumscribed by minimal box with edge length 1 (conventional units); (b) convex polyhedron related to Model B; (c) convex polyhedron related to Model C; 1 – the edge is flush with the *Top* side; 2 – the edge is flush with the *Left* side

However, in practise (to be shown in the experimental part, section 4.2), the minimal solution may also correspond to the case when one or more sides of the bounding box are coplanar with faces of the convex polyhedron. An example, where each side of the bounding box is coplanar with face of the polyhedron (Model B) is shown in Figure 2b. The vertex coordinates of this convex polyhedron are given as the Model B in Table 1. The Model B was derived from the reference Model A. The Model A is based on a regular cube with edge length 1 and chamfers with distances 0.1x0.1 (conventional units). So that each side of the Model A is given by five points. There is one point in the middle of a face, and there are four points in the corners of the face. The modified coordinates of Model B and Model C relative to the Model A are marked by **bold text** in Table 1. Figure 2c shows the other example, where two adjacent sides of MVBB are coplanar only with two edges 1 and 2 of the convex polyhedron. Optimization curves for the minimum volume versus an orientation angle between a bounding box side and a face of the Model C are illustrated in Figure 3. The relationship of the minimal volume versus the orientation angle of the Models may appear either linear (Figure 3a) or nonlinear (Figure 3b). The beginning of both curves corresponds to the volume where one side of the bounding box is coplanar with the polyhedron face. The end of the curves corresponds to the volume where the same side of the bounding box is coplanar with its adjacent polyhedron face. The other points on the curves correspond to the volume for orientation angles where one side of the bounding box coincides with the polyhedron edge.



(a) (b)

Figure 3. The optimization functions of volume versus orientation angle between two faces (Model C): (a) around the edge 2 on the *Left* side (small angle); (b) around the edge 1 on the *Top* side (large angle).

Table 1. The coordinates of points for the theoretical models (in convention units)

Side	***Model A***			***Model B*** (Figure 2b)			***Model C*** (Figure 2c)		
	X	***Y***	***Z***	***X***	***Y***	***Z***	***X***	***Y***	***Z***
	0,90	0,40	0,5	**0,91**	0,40	0,5	**0,91**	0,40	0,5
	0,90	0	0,1	0,90	0	0,1	0,90	0	0,1
Front	0,9	0,8	0,9	0,9	0,8	0,9	0,9	0,8	0,9
	0,9	0,8	0,1	0,9	0,8	0,1	0,9	0,8	0,1
	0,9	0	0,9	0,9	0	0,9	0,9	0	0,9
	-0,1	0,4	0,5	**-0,11**	0,4	0,5	**-0,11**	0,4	0,5
	-0,1	0	0,1	-0,1	0	0,1	-0,1	0	0,1
Back	-0,1	0,8	0,9	-0,1	0,8	0,9	-0,1	0,8	0,9
	-0,1	0,8	0,1	-0,1	0,8	0,1	-0,1	0,8	0,1
	-0,1	0	0,9	-0,1	0	0,9	-0,1	0	0,9
	0,4	-0,1	0,5	0,4	**-0,11**	0,5	0,4	**-0,11**	0,5
	0	-0,1	0,1	0	-0,1	0,1	0	-0,1	0,1
Left	0,8	-0,1	0,9	0,8	-0,1	0,9	0,8	-0,1	0,9
	0,8	-0,1	0,1	0,8	-0,1	0,1	0,8	-0,1	0,1
	0	-0,1	0,9	0	-0,1	0,9	0	-0,1	0,9
	0,4	0,9	0,5	0,4	**0,91**	0,5	0,4	0,91	0,5
	0	0,9	0,1	0	0,9	0,1	0	**0,95**	0,1
Right	0,8	0,9	0,9	0,8	0,9	0,9	0,8	**0,95**	0,9
	0,8	0,9	0,1	0,8	0,9	0,1	0,8	0,9	0,1
	0	0,9	0,9	0	0,9	0,9	0	0,9	0,9
	0,4	0,4	1	0,4	0,4	**1,01**	0,4	0,4	**1,01**
	0,8	0,8	1	0,8	0,8	1	0,8	0,8	1
Top	0	0	1	0	0	1	0	0	1
	0,8	0	1	0,8	0	1	0,8	0	**1,5**
	0	0,8	1	0	0,8	1	0	0,8	**1,5**
	0,4	0,4	0	0,4	0,4	**-0,01**	0,4	0,4	**-0,01**
	0	0,8	0	0	0,8	0	0	0,8	0
Bottom	0,8	0	0	0,8	0	0	0,8	0	0
	0	0	0	0	0	0	0	0	0
	0,8	0,8	0	0,8	0,8	0	0,8	0,8	0

3 Computation methods

In the following sections, three methods for finding the Minimum Volume Bounding Box (MVBB) are considered. The methods are denoted as the "*side-"* method (MVBBS), the "*face-"* method (MVBBF) and the "*edges-"* method (MVBBE). All three methods differ from each other by accuracy, complexity and hence the computation time.

All the three methods utilizes the MABR algorithm [4]. Two of the methods ("*face-"*, "*edges-"*) include the specific *data pre-processing algorithm* (section 3.3), which distinguishes these methods from other known methods. Only the *MVBBE method* completely satisfies to both theorems given in section 2.1 and 2.2, and therefore it can be used as the *reference* for the other alternative methods.

3.1 The Minimum Area Bounding Rectangle (MABR) algorithm

The MABR algorithm is based on 2D convex hull operation [3]. After a convex polygon P is constructed, the angles θ_i between the polygon edges and the **x**-axis are calculated as follows:

$$\theta_i = \text{atan}\,2(y_{i+1} - y_i, x_{i+1} - x_i), \qquad -\pi \le \theta \le \pi \tag{1}$$

where atan2 is the four-quadrant tangent inverse function. The polygon vertices $(p_1, p_2, \ldots p_n)$ are rotated in such way that the first convex polygon edge e_1 is parallel with the **x**-axis. Then at least three other points p_i with extreme (x, y) coordinates are defined – two in orthogonal direction to the **x**-axis ($y_{\max}$, $y_{\min}$), and another two coordinates in orthogonal direction to the **y**-axis ($x_{\min}$, $x_{\max}$). The polygon vertices continues rotating with angle $-\theta_i$ in clockwise direction from one edge e_i to another e_{i+1} until all polygon edges are checked. The two-dimension rotation matrix is a follows:

$$R(\theta_i) = \begin{bmatrix} \cos(\theta_i) & \sin(\theta_i) \\ -\sin(\theta_i) & \cos(\theta_i) \end{bmatrix} \tag{2}$$

A new rectangle area A_i is calculated for each rotation. The corresponding rectangle length $L_i = x_{\max} - x_{\min}$ and width $W_i = y_{\max} - y_{\min}$ are updated, when a new minimum area $A_{\min} = L_i \cdot W_i$ is obtained. An example of the MABR is shown in Figure 4. One of the edges of the polygon is collinear with one of the sides of the bounding rectangle. The algorithm also checks whether the solution is unique or not.

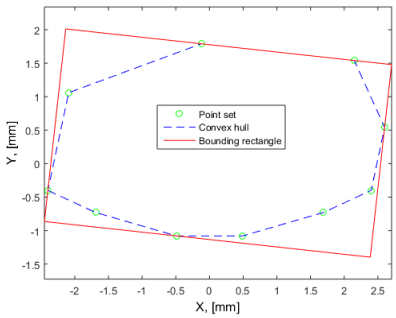


Figure 4. The MABR of a convex polygon based on 2D point set

3.2 The Minimum Volume Bounding Box Side (MVBBS) Method

This MVBBS approximation method is well known and often used in practice. It is fast and straightforward, based on an assumption that the test object has one perfectly flat side e.g. *Bottom*, which is aligned with the support surface ($Z_{\min}$). Such assumption allows a substantial simplification, both the measurement procedure and the computation procedure. However, because of the assumption of one perfectly flat side, the estimated minimal volume by this method can be not accurate. Groen et al. [17] developed an operational automatic system for measurement of parcels and suitcases on a conveyor belt based on this principle. The flowchart of the MVBBS method is illustrated in Figure 5.

The principle of this method is to define the height as $H_{\min} = Z_{\max} - Z_{\min}$ and the smallest area $A_{\min}$ of the bounding rectangle for the **xy**-projection of all measured points. As long as we

consider a single 2D projection of the convex polyhedron, then the MABR algorithm (section 3.1) is applied only once.

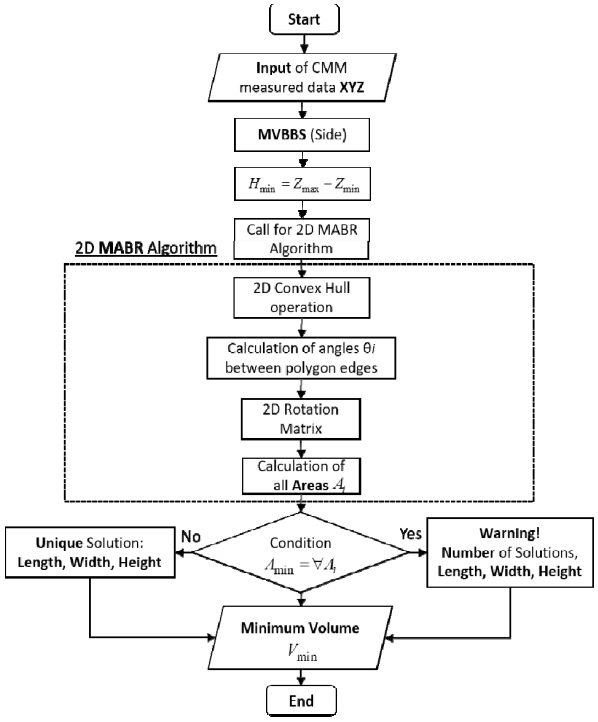


Figure 5. The flowchart of the MVBBS method with the MABR algorithm

3.3 Data pre-processing

In this paper, we focus on solving the minimum bounding box problem for physical objects that are rectangular objects close to the perfectly shaped bounding box. The measurement points of each side of the objects are given as six sets of points: *Front, Back, Right, Left, Top* and *Bottom*. The data set of each side is a $n \times 3$ matrix containing **x**- **y**- and **z**-coordinates for the n number of points.

In order to reduce the number of points for further computation, the 3D convex hull operation is applied. The input of the convex hull operation are the point coordinates from the six sets of points jointed together as illustrated in Figure 6. The output from the convex hull operation is a matrix $S_{m,3}^{Pol}$ with m rows. Each row of the matrix is a convex polyhedron face φ_i defined by its vertices. The vertices are given as indices that refer to the input data to the convex hull operation.

Some of the faces of the polyhedron described by $S_{m,3}^{Pol}$ will have vertices from two or three sides of the physical object. For example, the measured points from the *Top* side may be combined with measured points from the *Front* side into common faces, or "chamfer" faces between the sides. When defining the minimum bounding box in measurement and calibration of rectangular objects, these combined faces and their edges will not contribute to the solution, and they should not be used in the permutation part of the algorithm.

Two data structures are constructed from the output matrix $S_{m,3}^{Pol}$ by the pre-processing algorithm. The first structure represents six matrices $S^F, S^B, S^R, S^L, S^T, S^M$ of face vertices $v_{i,j}$ separated according to the reference object sides (*Front, Back,...Bottom*) without common faces (Figure 6, denoted by I); the second is a data matrix $P_{k,4}$ with **x**- **y**- and **z**-coordinates for face vertices (Figure 6, denoted by II).

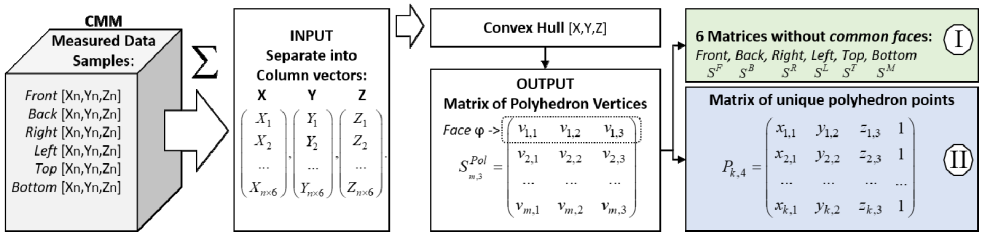


Figure 6. The flowchart of the data pre-processing with the output of two data structures I and II

3.4 The Minimum Volume Bounding Box Face (MVBBF) Method

The MVBBF method developed by the author is more accurate than the MVBBS method, but it is still an approximation version. The flowchart of the algorithm of MVBBF method is shown in Figure 7.

The theorem presented in section 2.2 does not provide an upper limit for how many edges that can be coplanar with one side of the bounding box, and then we may assume that one side is coplanar with more than one edge. It is a well-known fact that two distinct but intersecting lines uniquely determine a plane. Hereby, if a side is coplanar with two edges then it is coplanar with a face of the convex polyhedron. Obviously, one side of the bounding box cannot be coplanar with more than one face of the convex polyhedron. The second adjacent bounding box side must flush with at least one edge or face of the convex polyhedron.

In order to compensate the computation complexity of the MVBBF method, first we apply the *data pre-processing* (section 3.3). Then, the MVBBF algorithm searches through the six matrices $S^F, S^B, ...S^M$ associated with sides of the measured object and checks all faces within each sample. When a side and the first face φ of the polyhedron are chosen, three vertices $v_{1,1}, v_{1,2}, v_{1,3}$ of the face are defined. Two vectors $\boldsymbol{e}_1\{a_1, b_1, c_1\}$, $\boldsymbol{e}_2\{a_2, b_2, c_2\}$ are constructed based on the three given points $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$, $P_3(x_3, y_3, z_3)$, Figure 8. The cross product of the two vectors in $\mathbb{R}^3$ is a new vector $\boldsymbol{n}$, which is perpendicular to both given vectors [18], and this vector $\boldsymbol{n}$ is a normal vector to the face φ:

$$\boldsymbol{n} = \boldsymbol{e}_1 \times \boldsymbol{e}_2 = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & y_3 - y_1 \end{vmatrix} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{vmatrix}. \tag{3}$$

Then coordinates of the normal vector $\boldsymbol{n}\{A, B, C\}$ can be found as the minors of the matrix in (3) as following:

$$A = \begin{vmatrix} b_2 & c_2 \\ b_1 & c_1 \end{vmatrix}, \quad B = \begin{vmatrix} c_2 & a_2 \\ c_1 & a_1 \end{vmatrix}, \quad C = \begin{vmatrix} a_2 & b_2 \\ a_1 & b_1 \end{vmatrix}. \tag{4}$$

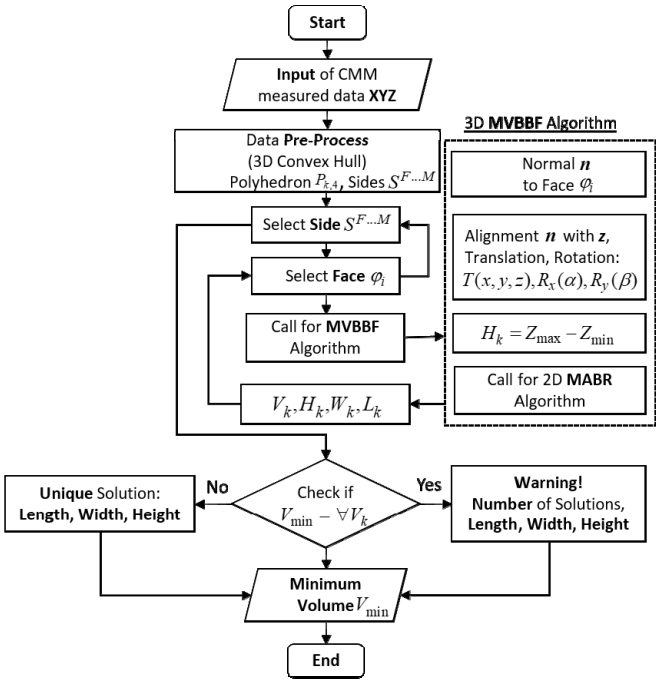


Figure 7. The flowchart of the MVBBF method

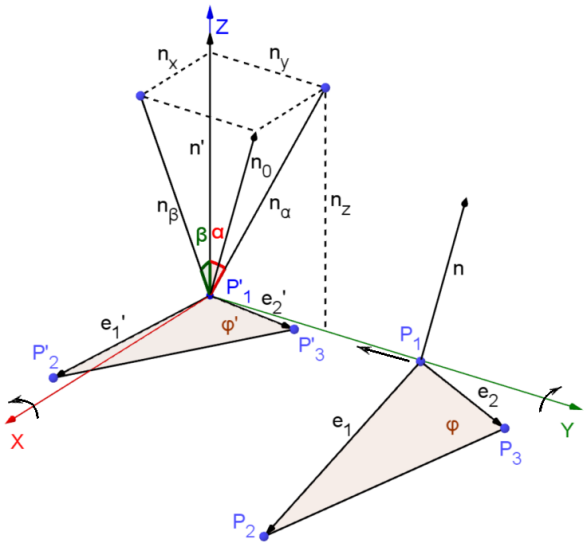


Figure 8. Coordinate transformation of an arbitrary polyhedron face

In order to combine the polyhedron face φ with an associated bounding box side (e.g. **xy** plane), we need to align the normal vector $\boldsymbol{n}$ with positive **z**-axis Figure 8. The first step is to move the vector $\boldsymbol{n}$ to the origin by using a translation matrix M with the row vector coordinates of the point $P_1(x_1, y_1, z_1)$, [19]:

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -x_1 & -y_1 & -z_1 & 1 \end{bmatrix}, \tag{5}$$

which gives us a vector $\boldsymbol{n}_0$. The projections n_α, n_β of the vector $\boldsymbol{n}_0$ on planes **yz** ($x=0$) and **zx** ($y=0$) respectively (Figure 8), give us two angles α and β:

$$\alpha = \arcsin\left(\frac{n_y}{n_\alpha}\right) = \arcsin\left(\frac{n_y}{\sqrt{{n_z}^2 + {n_y}^2}}\right) \tag{6}$$

$$\beta = \arcsin\left(\frac{n_x}{n_\beta}\right) = \arcsin\left(\frac{n_x}{\sqrt{{n_z}^2 + {n_x}^2}}\right). \tag{7}$$

Then a rotation matrix $R_x(\alpha)$ with the angle α for rotating counterclockwise around **x**-axis can be written:

$$R_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos(\alpha) & \sin(\alpha) & 0 \\ 0 & -\sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{8}$$

A rotation matrix $R_y(\beta)$ with the angle β for rotating clockwise around **y**-axis can be expressed in the following way:

$$R_y(\beta) = \begin{bmatrix} \cos(\beta) & 0 & \sin(\beta) & 0 \\ 0 & 1 & 0 & 0 \\ -\sin(\beta) & 0 & \cos(\beta) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{9}$$

The final transformation matrix T_φ to combine the polyhedron face φ with **xy**-plane as the bounding box side will be as follows (equivalent to alignment of $\boldsymbol{n}$ with **z**-axis):

$$T_\varphi = [M][R_x(\alpha)][R_y(\beta)]. \tag{10}$$

The transformed face φ and normal vector $\boldsymbol{n}$ are denoted as φ' and $\boldsymbol{n}'$ in Figure 8. In order to rotate the convex polyhedron, the transformation matrix T_φ is applied to the matrix $P_{k,4}$ (Figure 6, denoted by II) of the unique polyhedron vertices.

After the coordinate transformation is completed, all newly transformed points are projected into the **xy**-plane. Then the MABR algorithm (section 3.1) is applied for these projected points. It defines an orientation of the "close-loop" of *adjacent sides* (section 1) and, hence the estimation of width W_k and length L_k of the minimum bounding box. The height H_k is defined

as a difference between maximum and minimum **z**-values: $Z_{\max} - Z_{\min}$. Thus, the volume is: $V_k = H_k \cdot W_k \cdot L_k$.

The described procedure is repeated for each face of the chosen matrix and for all six matrices ($S^F, S^B, ...S^M$). The minimum volume V_k is calculated in each iteration. After all iterations are completed, the smallest value $V_{\min}$ is chosen as the solution.

3.5 The Minimum Volume Bounding Box Edge (MVBBE) Method

The third method corresponds to the conditions of the theorems presented in section 2.2 and therefore this is the most accurate method, which guarantees the global minimum solution. However, the algorithm is more complex and hence slower than two previous methods. In this case, the data pre-process (section 3.3) becomes the crucial part of the algorithm due to a significant reduction of unnecessary computation of the "chamfer" faces and corresponding edges of the convex polyhedron. The MVVBE algorithm is shown in Figure 9.

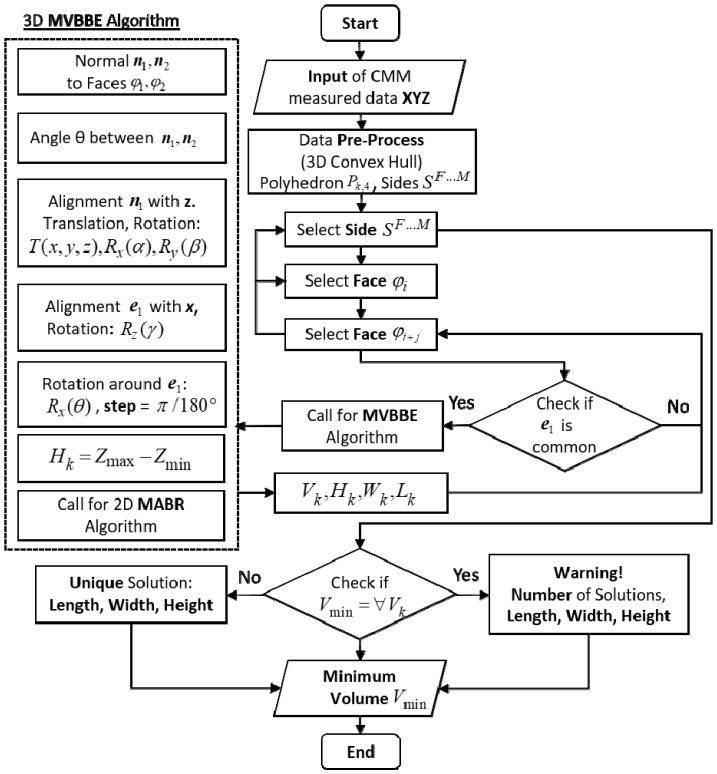


Figure 9. The flowchart of the MVBBE method

The MVBBE method is applied after the 3D convex hull operation and the data pre-process are completed. As before, we use six matrices (S^F, $S^B, ...S^M$) associated with the cuboid reference object sides as the output of the data pre-processing algorithm (Figure 6, denoted by I). The algorithm checks for each pair of faces with common edges. Since such pair of two faces

with their vertices $\varphi_1\left[v_{1,1}, v_{1,2}, v_{1,3}\right]$, $\varphi_2\left[v_{2,1}, v_{2,2}, v_{2,3}\right]$ are found, it gives us the four non-collinear points $P_0(x_0, y_0, z_0)$, $P_1(x_1, y_1, z_1)$, $P_2(x_2, y_2, z_2)$, $P_3(x_3, y_3, z_3)$ and three non-collinear vectors corresponding to the polyhedron edges $\boldsymbol{e}_1\{x_1 - x_0, y_1 - y_0, z_1 - z_0\}$, $\boldsymbol{e}_2\{x_2 - x_0, y_2 - y_0, z_2 - z_0\}$ and $\boldsymbol{e}_3\{x_3 - x_0, y_3 - y_0, z_3 - z_0\}$ or as a simplified form $\boldsymbol{e}_1\{a_1, b_1, c_1\}$, $\boldsymbol{e}_2\{a_2, b_2, c_2\}$ and $\boldsymbol{e}_3\{a_3, b_3, c_3\}$ respectively, (Figure 10).

Thus, the plane corresponding to the face φ_1 can be defined by the two non-collinear vectors $\boldsymbol{e}_2, \boldsymbol{e}_1$ in the following parametric form:

$$\begin{vmatrix} x - x_0 & y - y_0 & z - z_0 \\ a_2 & b_2 & c_2 \\ a_1 & b_1 & c_1 \end{vmatrix} = 0 \,, \tag{11}$$

and similarly by $\boldsymbol{e}_1, \boldsymbol{e}_3$ for the face φ_2:

$$\begin{vmatrix} x - x_0 & y - y_0 & z - z_0 \\ a_1 & b_1 & c_1 \\ a_3 & b_3 & c_3 \end{vmatrix} = 0 \,. \tag{12}$$

The angle θ between the faces φ_1 and φ_2 is the angle between the normal vectors $\boldsymbol{n}\{A, B, C\}, \boldsymbol{n}_2\{A_2, B_2, C_2\}$ [18]:

$$\theta = \pi - \arccos\left(\frac{AA_2 + BB_2 + CC_2}{\sqrt{A^2 + B^2 + C^2}\sqrt{A_2^2 + B_2^2 + C_2^2}}\right) \tag{13}$$

where A, B, C, A_2, B_2, C_2 are three corresponding minors of matrix (11) and (12), which can be calculated by using of equation (4).

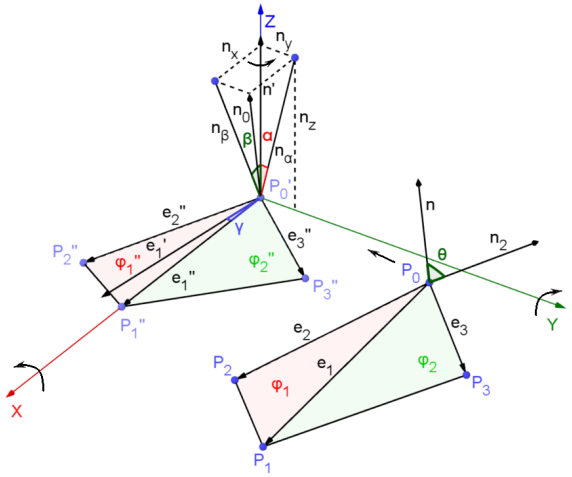


Figure 10. Coordinate transformation of two arbitrary polyhedron faces φ_1, φ_2 with the common edge $\boldsymbol{e}_1$

In order to find the minimum volume solution, the polyhedron points $P_{k,4}$ need to be rotated around the common edge $\boldsymbol{e}_1$ to the angle θ, from the face φ_1 to the face φ_2. A one-degree step is used for each iteration, but at least three iterations are applied if the angle θ is less than $2°$.

A certain alignment may be done to simplify the rotation of the polyhedron around the edge. First, **xy**-plane is made flush with the face φ_1 by alignment of the normal vector $\boldsymbol{n}$ with **z**-axis. The same technique is applied as it was described in section 3.4. The vector $\boldsymbol{n}$ moves to the origin by the translation matrix $M(x_0, y_0, z_0)$ in (5), rotate counterclockwise with angle α around **x**-axis by using the rotation matrix $R_x(\alpha)$ in (8), and clockwise with angle β around **y**-axis by using the rotation matrix $R_y(\beta)$ in (9). The next, the common edge $\boldsymbol{e}_1$ is aligned with **x**-axis by following rotation matrix $R_z(\gamma)$ around **z**-axis with angle, for clockwise (the example in Figure 10 has positive γ - counterclockwise):

$$R_z(\gamma) = \begin{bmatrix} \cos(\gamma) & -\sin(\gamma) & 0 & 0 \\ \sin(\gamma) & \cos(\gamma) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{14}$$

Thus, a full transformation matrix T_e for alignment of normal vector $\boldsymbol{n}$ with **z**-axis and the polyhedron edge $\boldsymbol{e}_1$ with **x**-axis, can be written in this way:

$$T_e = [M][R_x(\alpha)][R_y(\beta)][R_z(\gamma)] \tag{15}$$

A result of transformation of the two faces φ_1, φ_2 into φ_1", φ_2" and the edge $\boldsymbol{e}_1$ into $\boldsymbol{e}'_1$ is shown in Figure 10.

The alignments of the face φ_1" with **xy**-plane, and the edge $\boldsymbol{e}'_1$ with **x**-axis provide a transformed edge denoted as $\boldsymbol{e}''_1$ (Figure 10). Then, the rotation of all points $P_{k,4}$ around the edge $\boldsymbol{e}''_1$ with one-degree step angle $d\theta = \pi/180°$ can be proceeded by using the rotation matrix $R_x(d\theta)$ given earlier in eq. (8). After each rotation step, newly transformed points are projected into **xy**-plane and the MABR algorithm (section 3.1) is applied for the projected points to estimate the width W_k and the length L_k of the minimum bounding box. The height H_k is defined as a difference between maximum and minimum $Z_{\max} - Z_{\min}$ values. Finally, the volume of the bounding box is: $V_k = H_k \cdot W_k \cdot L_k$.

The above procedure is carried out for each common edge of all six matrices $S^F, S^B, ...S^M$ (Figure 6, denoted by I). The volume V_k is calculated in each rotation step, and the smallest volume $V_{\min}$ is the solution.

4 Implementation

The algorithms described in the previous sections are developed and implemented in MATLAB® programming environment based on CMM measurement data. The measurements have been performed in a Leitz PMM-C-600 CMM with an analogue probe. The PC-DMIS software was utilized for operation of the CMM.

4.1 Experiment setup

For the experimental tests, we have used a cuboid object with the following nominal dimensions (the true values are unknown): *length* 210 mm, *width* 140 mm, and *height* 120 mm. The test object is shown in Figure 11. The measured data is arranged into separated data samples according to the cuboid sides: *Front, Back, Right, Left, Top* and *Bottom*. Each sample is a $n \times 3$ matrix with three columns and *n*-rows of **xyz**-coordinates corresponding to the *n*-measured points as shown in Figure 6. We have used a uniform distribution of measured points with 15 mm distance between the points. The total number of the measured points is $N = 650$.

In order to get complete measurements of all six sides of the test object in a common coordinate system, we have measured the object in two setups. The measurements of the two setups have been combined by using common alignment points in the two setups.

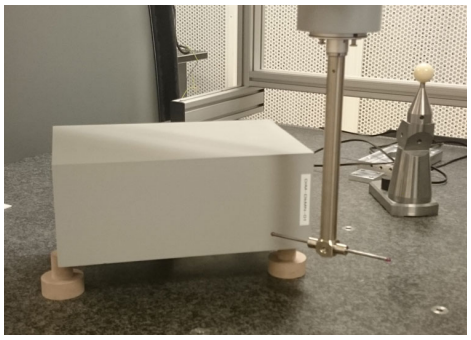


Figure 11. CMM measurement of the test object

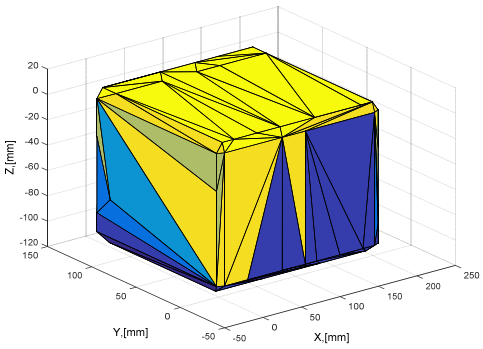


Figure 12. The result of the 3D convex hull operation

4.2 Results and Discussion

The collected date is further exported to a MATLAB code as an input for the developed algorithms. The first MVBBS method can be applied straightforward on the data – no data preprocess is required. We consider only the Bottom side as a support side. For the other two methods, we apply the data pre-process algorithm after the 3D convex hull operation. The result of the convex hull operation for measured data is shown on Figure 12. There are **166** faces combined together into one convex polyhedron and **88** faces after applying of the *data preprocess algorithm* (almost 50% of calculations were reduced). The computation results of all three methods are tabulated in Table 2 (the results are rounded to 1e-3).

The MVBBS method provides a significant overestimation of the volume V_S of the bounding box relative to the other two methods: $\Delta V = V_S - V_E = 951.959$ mm^3. Meanwhile, there is no difference between estimated volumes from the MVBBF and the MVBBE methods. A possible reason for such coincidence may be a small form deviation and as a result, small angles between polyhedron faces.

Table 2. The computation results of the MVBBS, MVBBF, MVBBE methods for the first test

Method	Width, mm	Length, mm	Height, mm	Volume, mm^3	Number of solutions
MVBBS	140.016	210.035	119.980	3528388.312	1
MVBBF	139.997	210.010	119.978	3527436.353	1
MVBBE	139.997	210.010	119.978	3527436.353	1

An extra test was applied for estimation of MVBB for the cuboid object with the same nominal dimensions but with larger flatness deviations. The computation results are given in Table 3.

Table 3. The computation results of the MVBBS, MVBBF, MVBBE methods for the second test

Method	Width, mm	Length, mm	Height, mm	Volume, mm^3	Number of solutions
MVBBS	140.016	210.195	120.068	3533662.007	1
MVBBF	139.995	210.195	120.054	3532743.929	1
MVBBE	139.994	210.195	120.055	3532735.210	1

The second test demonstrates the difference between MVBBF and MVBBE methods. The following difference can be observed from Table 3: $\Delta V = V_F - V_E = 8.675$ mm^3, where V_F is the solution of the MVBBF method and V_E is the solution of the MVBBE method.

In order to verify the proposed approaches, the developed methods were applied on Model B and Model C, which are illustrated in Figure 2(b, c). The vertex coordinates of the theoretical models are given in Table 1. The computation results for estimation of MVBB based on the developed methods for Model B and Model C are given in Table 4 (the results are rounded to 1e-4).

It can be observed a difference between the solutions for MVBBS, MVBBF and MVBBE methods for Model C. The MVBBE method provides the smallest solution for Model C. The results for Model B and for all three methods are equal.

The MVBBE method may provide the minimal solution and yet, it includes all solutions of the MVBBF method and therefor it is more reliable and accurate.

Table 4. The computation results of the developed methods for Model B and Model C

Method	Width, mm	Length, mm	Height, mm	Volume, mm^3	Number of solutions
MVBBS (Model B)	1.0197	1.0197	1.02	1.0605	2
MVBBS (Model C)	1.02	1.06	1.51	1.6326	1
MVBBF (Model B)	1.0197	1.02	1.0197	1.0605	4
MVBBF (Model C)	1.0197	1.0600	1.5195	1.6424	1
MVBBE (Model B)	1.0197	1.02	1.0197	1.0605	4
MVBBE (Model C)	1.0198	1.0598	1.5098	1.6319	1

5 Conclusion

Three methods have been proposed and demonstrated in this work for estimation of the minimum volume of bounding box with the proposed data pre-process algorithm for the metrological applications. The first two methods are based on a number of assumptions allowing decreasing of a computation time but often with overestimated results. The minimal and the most optimal solution is provided by the MVBBE method. Furthermore, the solution of the MVBBE method is based on theorems presented in this paper (sections 2.1 and 2.2) and hence, its estimation is the most accurate. Relying on type of dimensional measurement system, different methods may be applicable while the MVBBE method should utilize as the reference.

However, the MVBBE method includes a large number of an additional calculation. The proposed pre-process data algorithm (section 3.3) based on the specific metrological conditions (described in section 1) allows a significant reduction of the computation (about 50 %) preserving the initial accuracy at the same time. Thus, the MVBBE method should be used for those metrological tasks, where the accuracy is the critical factor, particularly when a large geometry form deviation is expected. The principles outlined in this work could also improve the functionality of operation software for the measuring systems.

6 Acknowledgements

The authors wish to thank Dr. Christoph A. Thieme, NTNU, Trondheim, for valuable advices and comments. The authors acknowledge the financial support of the Research Council of Norway, Grant No. 235315.

References

1. Dupuis Nathan Fellowes (1893) Elements of Synthetic Solid Geometry. Macmillan,

2. Moulai-Khatir Djezouli, Pairel Eric, Favreliere Hugues (2018) Influence of the probing definition on the flatness measurement. International Journal of Metrology and Quality Engineering 9:15. doi:10.1051/ijmqe/2018011

3. Shamos Michael (1978) Computational Geometry. Ph.D. thesis, Yale University,

4. Freeman H., Shapira R. (1975) Determining the minimum-area encasing rectangle for an arbitrary closed curve. Communications of the ACM 18 (7):409-413. doi:10.1145/360881.360919

5. Toussaint Godfried Solving geometric problems with the rotating calipers. In: IEEE MELECON, Greece, 1983.

6. Timos Sellis, Nick Roussopoulos, Christos Faloutsos (2018) The R+-Tree: A Dynamic Index for Multi-Dimensional Objects. Figshare. doi:10.1184/R1/6610748.V1

7. Beckmann N., Kriegel H. P., Schneider R., Seeger B. (1990) The R-tree: An Efficient and Robust Access Method for Points and Rectangles. ACM SIGMOD Record 19 (2):322-331. doi:10.1145/93605.98741

8. Roussopoulos Nick, Leifker Daniel (1985) Direct spatial search on pictorial databases using packed R-trees. doi:10.1145/318898.318900

9. Gottschalk S., Lin M. C., Manocha D. (1996) OBB tree: A hierarchical structure for rapid interference detection. doi:10.1145/237170.237244

10. Dimitrov Darko, Knauer Christian, Kriegel Klaus, Rote G. (2007) New upper bounds on the quality of the PCA bounding boxes in r2 and r3. doi:10.1145/1247069.1247119

11. O'Rourke Joseph (1985) Finding minimal enclosing boxes. International Journal of Computer & Information Sciences 14 (3):183-199. doi:10.1007/BF00991005

12. Bespamyatnikh Sergei, Segal Michael (2000) Covering a set of points by two axis-parallel boxes. Information Processing Letters 75 (3):95-100. doi:10.1016/S0020-0190(00)00093-4

13. Lahanas M., Kemmerer T., Milickovic N., Karouzakis K., Baltas D., Zamboglou N. Optimized bounding boxes for three-dimensional treatment planning in brachytherapy. Medical Physics 27 (10):2333-2342. doi:10.1118/1.1312808

14. Barequet Gill, Har-Peled Sariel (2001) Efficiently Approximating the Minimum-Volume Bounding Box of a Point Set in Three Dimensions. Journal of Algorithms 38 (1):91-109. doi:10.1006/jagm.2000.1127

15. Dimitrov D., Holst M., Knauer C. , Kriegel K. Experimental study of bounding box algorithms. In: The Third International Conference on Computer Graphics Theory and Applications, 2008. pp 15-22

16. Barber C., Dobkin David, Huhdanpaa Hannu (1996) The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software (TOMS) 22 (4):469-483. doi:10.1145/235815.235821

17. Groen F. C. , Verbeek P. W., de Jong N., Klumper J. W. (1981) The smallest box around a package. Pattern Recognition 14 (1):173-178. doi:10.1016/0031-3203(81)90059-5

18. Aleksandrov P. (1968) Lectures of Analitical Geometry /Лекции по аналитической геометрии. Science/Наука,

19. Leon Steven J. (2010) Linear algebra with applications. 8th ed. edn. Pearson, Upper Saddle River, N.J