

STUDIES IN  
CLASSIFICATION, DATA ANALYSIS, AND KNOWLEDGE ORGANIZATION

M. Vichi · O. Opitz

Editors

# Classification and Data Analysis

Theory  
and  
Application



Springer

# Studies in Classification, Data Analysis, and Knowledge Organization

---

11

## *Managing Editors*

H.-H. Bock, Aachen  
W. Gaul, Karlsruhe  
M. Schader, Mannheim

## *Editorial Board*

F. Bodendorf, Nürnberg  
P. G. Bryant, Denver  
F. Critchley, Birmingham  
E. Diday, Paris  
P. Ihm, Marburg  
J. Meulmann, Leiden  
S. Nishisato, Toronto  
N. Ohsumi, Tokyo  
O. Opitz, Augsburg  
F.J. Radermacher, Ulm  
R. Wille, Darmstadt

**Springer**

*Berlin*

*Heidelberg*

*New York*

*Barcelona*

*Hong Kong*

*London*

*Milan*

*Paris*

*Singapore*

*Tokyo*

## **Titles in the Series**

- H.-H. Bock and P. Ihm (Eds.)  
Classification, Data Analysis, and Knowledge Organization. 1991  
(out of print)
- M. Schader (Ed.)  
Analyzing and Modeling Data and Knowledge. 1992
- O. Opitz, B. Lausen, and R. Klar (Eds.)  
Information and Classification. 1993  
(out of print)
- H.-H. Bock, W. Lenski, and M.M. Richter (Eds.)  
Information Systems and Data Analysis. 1994  
(out of print)
- E. Diday, Y. Lechevallier, M. Schader, P. Bertrand,  
and B. Burtschy (Eds.)  
New Approaches in Classification and Data Analysis. 1994  
(out of print)
- W. Gaul and D. Pfeifer (Eds.)  
From Data to Knowledge. 1995
- H.-H. Bock and W. Polasek (Eds.)  
Data Analysis and Information Systems. 1996
- E. Diday, Y. Lechevallier and O. Opitz (Eds.)  
Ordinal and Symbolic Data Analysis. 1996
- R. Klar and O. Opitz (Eds.)  
Classification and Knowledge Organization. 1997
- C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H.-H. Bock,  
and Y. Baba (Eds.)  
Data Science, Classification, and Related Methods. 1998
- I. Balderjahn, R. Mathar, and M. Schader (Eds.)  
Classification, Data Analysis, and Data Highways. 1998
- A. Rizzi, M. Vichi, and H.-H. Bock (Eds.)  
Advances in Data Science and Classification. 1998

Maurizio Vichi · Otto Opitz  
Editors

# Classification and Data Analysis

Theory and Application

Proceedings of the Biannual Meeting of the  
Classification Group of Società Italiana di Statistica (SIS)  
Pescara, July 3-4, 1997

With 97 Figures  
and 78 Tables



Springer

Prof. Maurizio Vichi  
University "G. D'Annunzio" di Chieti  
Dipartimento Metodi Quantitativi e  
Teoria Economica  
Viale Pindaro 42  
I-65127 Pescara, Italy

Prof. Dr. Otto Opitz  
University of Augsburg  
Lehrstuhl für Mathematische Methoden  
der Wirtschaftswissenschaften  
D-86135 Augsburg  
Germany

ISBN 3-540-65633-2 Springer-Verlag Berlin Heidelberg New York

Cataloging-in-Publication Data applied for -  
Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Classification and data analysis: theory and application; proceedings of the biannual meeting of the Classification Group of Società Italiana di Statistica (SIS), Pescara, July 3-4, 1997; with 78 tables/Maurizio Vichi; Otto Opitz ed. - Berlin; Heidelberg; New York; Barcelona; Hong Kong; London; Milan; Paris; Singapore; Tokyo: Springer, 1999

(Studies in classification, data analysis, and knowledge organization)

ISBN 3-540-65633-2

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

© Springer-Verlag Berlin · Heidelberg 1999  
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Softcover-Design: Erich Kirchner, Heidelberg

SPIN 10700416

32/2202-5 4 3 2 1 0 - Printed on acid-free paper

# PREFACE

## International Federation of Classification Societies

The International Federation of Classification Societies (IFCS) is an agency for the dissemination of technical and scientific information concerning classification and data analysis in the broad sense and in as wide a range of applications as possible; founded in 1985 in Cambridge (UK) from the following Scientific Societies and Groups: British Classification Society - BCS; Classification Society of North America -CSNA; Gesellschaft für Klassifikation - GfKI; Japanese Classification Society - JCS; Classification Group of Italian Statistical Society - CGSIS; Société Francophone de Classification - SFC. Now the IFCS includes the following Societies: Dutch-Belgian Classification Society - VOC; Polish Classification Section - SKAD; Portuguese Classification Association - CLAD; Group-at-Large; Korean Classification Society - KCS.

## Biannual Meeting of the Classification and Data Analysis Group of SIS

The biannual meeting of the Classification and Data Analysis Group of Società Italiana di Statistica (SIS) was held in Pescara, July 3 - 4, 1997.

The 69 papers presented were divided in 17 sessions. Each session was organized by a chairperson with two invited speakers and two contributed papers from a call for papers. All the works were referred. Furthermore, during the meeting a discussant was provided for each session. A short version of the papers (4 pages) was published before the conference.

The scientific program covered the following topics:

- *Classification Theory*

Fuzzy Methods - Hierarchical Classification - Non Hierarchical Classification - Optimisation approach in Classification. - Classification of Multiway Data - Probabilistic Methods for Clustering - Consensus and Comparison Theories in Classification - Spatial data and Clustering - Validity of Clustering - Neural Networks and Classification - Genetic Algorithms - Classification with Constraints

- *Multivariate Data Analysis*

Categorical Data Analysis - Factor Analysis and Related Methods - Discrimination and Classification - Visual Treatment in Data Analysis Symbolic Data Analysis - Non Linear Data Analysis

- *Multiway Data Analysis*

Analysis of Multiway Data - Panel Data Analysis

- *Proximity Structure Analysis*

Multidimensional Scaling - Similarities and Dissimilarities -

- *Software Developments for Classification and Data Analysis*

Algorithms for Hierarchical and Non Hierarchical Classification - Computer Data Visualization. Statistical Algorithms for Multivariate Analysis

- *Applied Classification and Data Analysis in Social, Economic, Medical, and other Sciences*

Classification and Data Analysis of Textual Data - Data Analysis in Economics - Classification and Discrimination Approaches in Medical Science

The present volume contains 45 referred papers presented in four chapters as follows:

### **Classification**

- Methodologies in Classification
- Fuzzy clustering and fuzzy methods

### **Other Approaches for Classification**

- Discrimination and Classification
- Regression Tree and Neural Networks

### **Multivariate and Multidimensional Data Analysis**

- Proximity Methodologies in Classification
- Factorial methods
- Spatial Analysis
- Multiway Data Analysis
- Multivariate analysis

### **Case Studies**

## **Acknowledgements**

The Editors would like to extend their sincere thanks to the authors whose enthusiastic participation made the present meeting possible. We are very grateful to the reviewers of the short papers and the time spent in their professional reviewing work. We are also grateful to the chairpersons and discussants of the sessions that also provided very useful suggestions to the authors.

Special thanks are due to the Local Organizing Committee of Pescara: Mauro Coli, Tonio Di Battista, Stella Iezzi, Eugenia Nissi, Antonio Pacinelli, Tonino Sclocco, Maurizio Vichi (*chairman*)

Finally, thanks are extended to Springer-Verlag, Heidelberg.

Maurizio Vichi

# TABLE OF CONTENTS

Preface.....V

## PART I: Classification

### Methodologies in Classification

*A. Cerioli (Università di Parma)*  
Measuring the influence of individual observations and variables  
in cluster analysis..... 3  
*P. D'Urso, M. G. Pittau (Università "La Sapienza" di Roma)*  
Consensus classification for a set of multiple time series ..... 11  
*T. Di Battista, D. Di Spalatro (Università di Chieti)*  
A bootstrap method for adaptive cluster sampling ..... 19  
*D. Iezzi, M. Vichi (Università di Chieti)*  
Forecasting a classification ..... 27

### Fuzzy Clustering and Fuzzy Methods

*A. Bellacicco (Università di Teramo)*  
Neural networks as a fuzzy semantic network of events ..... 35  
*L. Cerbara (IRP-CNR)*  
Hierarchical fuzzy clustering: an example of spatio-temporal analysis ..... 43  
*G. Iacovacci (ISTAT)*  
A new algorithm for semi-fuzzy clustering ..... 49  
*A. Maturo, B. Ferri (Università di Chieti)*  
Fuzzy classification and hyperstructures: an application to  
evaluation of urban projects..... 55  
*M. A. Milioli (Università di Parma)*  
Variable selection in fuzzy clustering..... 63

## PART II : Other Approaches for Classification

### Discrimination and Classification

*M. Alfò, P. Postiglione (Università di Chieti)*  
Discriminant analysis using markovian automodels ..... 73  
*F. Esposito, D. Malerba, G. Semeraro, S. Caggese (Università di Bari)*  
Discretization of continuous-valued data in symbolic  
classification learning ..... 81



<i>S. Ingrassia (Università di Catania)</i>	
Logistic discrimination by Kullback-Leibler type distance measures.....	89
<i>A. Montanari, D. Calò (Università di Bologna)</i>	
An empirical discrimination algorithm based on projection pursuit density estimation.....	97

### Regression Tree and Neural Networks

<i>R. Miglio, M. Pillati (Università di Bologna)</i>	
Notes on methods for improving unstable classifiers.....	105
<i>F. Mola (Università "Federico II" di Napoli)</i>	
Selection of cut points in generalized additive models .....	113
<i>R. Siciliano (Università "Federico II", Napoli)</i>	
Latent budget trees for multiple classification.....	121

## PART III: Multivariate and Multidimensional Data Analysis

### Proximity Analysis and Multidimensional Scaling

<i>G. Bove, R. Rocci (Università di Roma)</i>	
Methods for asymmetric three-way scaling.....	131
<i>S. Camiz (Università "La Sapienza" di Roma)</i>	
Comparison of Euclidean approximations of non-Euclidean distances .....	139
<i>A. Montanari, G. Soffritti (Università di Bologna)</i>	
Analysing dissimilarities through multigraphs.....	147
<i>C. Quintano (Istituto Universitario Navale di Napoli)</i>	
Professional positioning based on dominant eigenvalue scores (DES), dimensional scaling (DS) and multidimensional scaling (MDS) synthesis of binary evaluations matrix of experts .....	155
<i>M. Vichi (Università di Chieti)</i>	
Non-metric full-multidimensional scaling.....	163

### Factorial Methods

<i>I. Corazziari (Università "Federico II" di Napoli)</i>	
Dynamic factor analysis .....	171
<i>V. Esposito, G. Scepi (Università "Federico II" di Napoli)</i>	
A non symmetrical generalised co-structure analysis for inspecting quality control data .....	179
<i>R. Lombardo, G. Tessitore (II Università di Napoli - Università "Federico II" di Napoli)</i>	
Principal surfaces constrained analysis .....	187

<i>R. Verde (Università "Federico II" di Napoli)</i>	
Generalised canonical analysis on symbolic objects .....	195
<i>G. Vittadini (Università di Milano)</i>	
Analysis of qualitative variables in structural models with unique solutions. ....	203

### Spatial Analysis

<i>A. Capobianchi, G. Jona-Lasinio (Università di Roma)</i>	
Exploring multivariate spatial data: line transect data.....	211
<i>A. Giusti, A. Petrucci (Università di Firenze)</i>	
On the assessment of geographical survey units using constrained classification .....	221
<i>L. Romagnoli (Università di Chieti)</i>	
Kalman filter applied to non-causal models for spatial data .....	229

### Multiway Data Analysis

<i>S. Bolasco, A. Morrone, F. Baiocchi (Università "La Sapienza" di Roma)</i>	
A paradigmatic path for statistical content analysis using an integrated package of textual data treatment .....	237
<i>M. Chiodi, A.M. Mineo (Università di Palermo)</i>	
The analysis of auxological data by means of nonlinear multivariate growth curves .....	247
<i>M. Coli, L. Ippoliti, E. Nissi (Università di Chieti)</i>	
The Kalman filter on three-way data matrix for missing data: a case study on sea water pollution.....	255
<i>P. A. Cornillon, P. Amenta, R. Sabatier, (Laboratoire de physique moléculaire et structurale, Università "Federico II" di Napoli)</i>	
Three-way data arrays with double neighborhood relations as a tool to analyze a contiguity structure.....	263
<i>A. Lemmi, D. Stefano Gazzei (Università di Siena, Università di Firenze)</i>	
Firm performance analysis with panel data .....	271

### Multivariate Data Analysis

<i>M. R. D'Esposito, G. Ragozini (Università di Salerno, Università "Federico II" di Napoli)</i>	
Detection of multivariate outliers by convex hulls.....	279
<i>M. Di Marzio, G. Lafratta (Università di Chieti)</i>	
Reducing dimensionality effects on kernel density estimation: the bivariate gaussian case.....	287
<i>M. Giacalone (Università di Napoli)</i>	
Shewhart's control chart: some observations .....	295

<i>A. Laghi, L. Lizzani (Università di Bologna)</i> Projection pursuit regression with mixed variables.....	303
<i>P. Mantovan, A. Pastore, S. Tonellato (Università di Venezia)</i> Recursive estimation of system parameter in environmental time series models.....	311
<i>A. Pallini (Università di Bologna)</i> Kernel methods for estimating covariance functions from curves.....	319
<i>G. Porzio (Università "Federico II", Napoli)</i> Detection of subsamples in link-free regression analysis .....	327
<i>A. Roverato (Università di Modena)</i> Asymptotic prior to posterior analysis for graphical gaussian models.....	335

## PART IV: Case Studies

### Applied Classification and Data Analysis

<i>S. Borra, A. Di Ciaccio (Università di Urbino)</i> Using qualitative information and neural networks for forecasting purposes in financial time series.....	345
<i>A. M. Mineo, A. Plaia (Università di Palermo)</i> A new approach to the stock location assignment problem by multidimensional scaling and seriation .....	353
<i>A. Turrini (Istituto Nazionale della Nutrizione)</i> Food coding in nutritional surveys.....	361
<i>C. Capiluppi, L. Fabbris, M. Scarabello (Università di Padova)</i> UNAIDED: a PC system for binary and ternary segmentation analysis.....	367
Author Index.....	375
Key Words Index.....	377

# Shewhart's Control Chart: Some Observations

Massimiliano Giacalone

Department of Mathematics and Statistics, University of Naples "Federico II".  
Complesso Monte S. Angelo, Via Cinthia, 80126 Naples.  
e-mail: max@matstat.dms.unina.it

**Abstract:** Data Analysis in Shewhart's Control Chart, to use the original  $m$  samples  $n$  sized intensities, is the main subject of this paper. Given  $m \times n$  intensities we examine three alternatives to sintetize the variability: a) arithmetic mean of  $m$  standard deviations ( $\bar{S}$ ); b) root mean square of  $m$  variances ( $\sqrt{S}$ ); c) global dispersion ( $\bar{\bar{S}}$ ). We prefer the global dispersion to estimate parent population  $\sigma^2$ .

As an alternative we suggest to analyze all the items of an unique random sample dimensioned in such a manner to have an efficient  $\sigma^2$  estimate. A second introduced proposal is to use the Factory's needs: ( $P_0, P_1, \alpha, \beta, L$  and  $U$ ). Some examples are given in the last session of the paper.

**Keywords:** Shewart's Control Chart, Sigma's Estimate, Data Analysis.

## 1. Introduction

Using S.C.C. (Shewhart's Control Chart) it is customary to operate 2 stages; a first stage devoted to data collection and limits  $LCL_{\bar{x}}, UCL_{\bar{x}}, LCL_s, UCL_s$  (Lower Control Limit and Upper Control Limit for mean and dispersion) computation. The second stage is devoted to chart's use.

In the first stage it is customary to produce  $K = m \cdot N$  items, (in other words we have  $m$  lots  $N$  sized), to draw  $m$  single random samples  $n$  size from each lot  $N$  sized. The population is given by all items *produced and to be produced*, its mean is  $\mu$  and its variance is  $\sigma^2$ ;  $\mu$  and  $\sigma^2$  supposed stable in the first stage (*items produced*).

Let us call  $x_{ij}$  the  $i$ th intensity of the  $j$ th sample, so the  $j$ th sample mean is:

$$\bar{x}_j = \sum_i x_{ij} / n \quad (i = 1, 2, \dots, n)$$

$$s_j^2 = \sum_i (x_{ij} - \bar{x}_j)^2 / (n-1) \quad (1)$$

is the  $j$ th sample variance estimate;

$$\bar{X} = \sum_j \bar{x}_j / m,$$

is the mean of the sample means.

Sample mean synthesis create no problem, not the same happens for  $s^2$  or  $s$ . Indeed some authors (W.A. Shewhart, 1931); (A. J. Duncan, 1965); (P.L. Piccari, 1974); (D.C. Montgomery, 1991) propose to compute:

$$S = \sum s_j / m \quad (2)$$

Some other authors (Mittag-Rinne, 1993) propose to compute:

$$S = \left\{ \sum s_j^2 / m \right\}^{1/2} \quad (3)$$

finally one may also compute:

$$S = \left\{ \sum_i \sum_j (X_{ij} - \bar{X})^2 / (m \cdot n - 1) \right\}^{1/2}; \quad (4)$$

In this paper we study the rationale of each solution and we suggest an alternative proposal.

## 2. Synthesis analysis

Since root mean square is greater than or equal to arithmetic mean, we may write:

$$S \leq \bar{s},$$

and declare that one of the introduced formulae can't be correct. Relation (2) is the main suspect because since:

$$E(s) \neq \sigma$$

The same may be said for (3), and this means  $S$  to be a biased  $\sigma$  estimate. Someone notes that, if the underlying population is normal,  $S$  actually estimates  $\sigma \cdot c_2$ ; this is statistically correct but a little cumbersome. We remember that  $c_2$  is a constant depending on the sample size  $n$ :

$$c_2 = \left\{ 2/(n-1) \right\}^{1/2} \cdot \Gamma \left\{ (n-1)/2 \right\};$$

tabulated values are presented in Duncan (1965).

Let us now consider the synthesis of sample variances (3). Kenney and Keeping

(1956), showed that:

$$E(s^2) = \sigma^2,$$

not only for simple samples, but also in presence of  $m$  simple samples. In case  $h$  independent samples are available from the universe, they suggest to use:

$$\hat{\sigma}^2 = Q/(U-h);$$

where

$$Q = n_1 s_1^2 + n_2 s_2^2 + \dots + n_h s_h^2;$$

$$U = n_1 + n_2 + \dots + n_h;$$

and  $s_i^2$  is the variance in the  $i$ th sample consisting of  $n_i$  variates.

If  $n_i = n$  is the same for every sample, we have:

$$\hat{\sigma}^2 = n(s_1^2 + s_2^2 + \dots + s_h^2)/(U-h);$$

where  $U = n \cdot h$ . Clearly the last relation may be written in the form:

$$(n-1)/n \cdot \hat{\sigma}^2 = (s_1^2 + s_2^2 + \dots + s_h^2)/h$$

The constant  $(n-1)/n$  is present because the authors started with  $s_j^2 = \sum_i (X_{ij} - \bar{X})^2 / n$  instead of  $s_j^2$ , but if the degrees of freedom are used, the result is correct and consistent with:  $E(s^2) = \sigma^2$ . This solution records time variations. In other words we have a trace of variability changes during data collection period.

Finally relation (4) is based on the whole group. It may be seen as the *total variance*, while  $\bar{S}^2$  may be seen as *within variance*. Deviances are the same if *between variance* is equal to zero.

There is someone discouraging its use. For instance D.C. Montgomery (1991), affirms that the estimate of the process standard deviation  $\sigma$  used in constructing the control limits is calculated from the variability within each sample. Consequently, the estimate of  $\sigma$  reflects within-sample variability only. It is not correct the estimate of  $\sigma$  based on the usual quadratic estimator, say  $\bar{S}$ , because if the sample means differ, then this will cause  $\bar{S}$  to be too large. Consequently, in this way,  $\sigma$  could be overestimated.

A. J. Duncan (1965) shares the same opinion, and retains that is not correct to estimate the process standard deviation from all the data (e. g.  $\bar{S}$ ) and use this in setting up limits for the  $\bar{X}$ -chart. The estimate of the process standard deviation to be used in setting up limits for the  $\bar{X}$ -chart must be computed from the within-sample variation to the exclusion of the between-sample variation.

Let us remember that if a production process presents stable between-sample variation it could be a good rule to look for the trouble and to remove it if possible. If the problem persists we do not see why to ignore it, computing the so called within variation. Another important remark is the difference between "first stage" and "second stage". In the second stage production must be monitored so that it is very useful to divide output into lots, let us say  $N$  sized, and investigate every single lot produced. If no trouble appears production can continue; on the contrary, if a trouble comes out it is much better to stop production and to look for happenings. In the second stage, points are regarded as independent events and O.C.C. (Operating Characteristic Curve) is computed under this assumption (G. Rouzet, 1957). In short, the division of production into lots  $N$  sized is a suitable procedure for the second stage as we said before.

The first stage problem is a different one. to estimate  $\mu$  and  $\sigma^2$  related to the character of interest. The subject involved is the *parent population* and its parameters. The division of items into lots  $N$  sized is not an essential operation. Perhaps the sample repetition is a mechanical consequence of the second stage technique, to some extent necessary if  $n=5$ , because  $\mu$  and  $\sigma^2$  estimates based on so a little sample should be extremely poor ones, so to have both ways saved some authors suggested to repeat the sample (and the lot)  $m$  times (Mittag-Rinne, 1993). It seemed therefore a natural consequence to compute  $\bar{X}_j$ ,  $s_j$  and  $S$ ,  $S$  and  $S$ .

### 3. Simulation

In order to emphasize our opinion we consider a simulation. We shall use Wold's Random Normal Deviates divided into lots  $N=50$  sized, one numbers column for lot. From each column we draw one sample  $n$  sized and this operation will be repeated  $m$  ( $=20$ ) times as in the first stage practice. We compute  $m$   $\bar{x}$  and  $m$   $\sigma^2$ , and the synthesis  $S^2$  is compared with  $S^2$ . We define:  $\text{DifTot} = \sigma^2 - S^2$  and  $\text{DifUni} = \sigma^2 - S^2$ . We also noted that here  $\sigma^2$  is the population variance computed on  $N \cdot m$  data = 1000 considering series of 100 samples. If  $\text{DifTot} < \text{DifUni}$  one point is given to  $S^2$ , but if  $\text{DifUni} < \text{DifTot}$  then one point is given to  $S^2$ .

For series of samples  $n=5$  sized we found more than 75% points for  $\text{DifUni}$ , then for  $S^2$ .

### 4. Alternative proposals

The first stage procedure is a very expensive one. Infact after  $m$  samples we must revise the production process, therefore to save time and money we

suggest to analyze all the items produced within the first stage and dimension this sample according to wanted protection.

Our suggestion seems particularly useful for destructive control analysis because with customary procedure if not analyzed items are out of tolerance, production-control costs increases.

Calling  $N$  the first stage lot size, we shall have:  $\bar{X} = \sum_i X_i / N$ , and  $\bar{S}^2 = \sum_i (X_i - \bar{X})^2 / (N-1)$ , as an unbiased  $\sigma^2$  estimate.

A different suggestion is based on the introduction of Factory's needs ( $P_0, P_1, \alpha, \beta, L$  and  $U$ ). Many authors, use symbol  $L$  for *Lower specification limit* and symbol  $U$  for *Upper specification limit*.

Now let us call  $P_0$  the well known *Acceptable Quality Level* and we underline that it seems suitable subdivide  $P_0$  into to parts, the one on the left  ${}_L P_0$  (fraction of too small items) and the other on the right  ${}_U P_0$  (fraction of too large items), of course  ${}_L P_0 + {}_U P_0 = P_0$ . This is enough for the computation of:

$$\bar{X}_0 = (Lz_U - Uz_L) / (Z_U - Z_L); \quad (Z_L < 0)$$

$$\sigma_0 = (U - L) / (Z_U - Z_L);$$

where  $Z_L$  is the normal standardized fractile given  ${}_L P_0$ , and  $Z_U$  the one given  ${}_U P_0$ .  $\bar{X}_0$  and  $\sigma_0$  are the parameters to be used for Shewhart's variables Control Chart computation.

The SCC so obtained is a very different tool because it privileges Factory's needs, whereas customary procedure privileges process capability. Therefore, once obtained the new SCC ( $UCL_{\bar{x}}, LCL_{\bar{x}}, UCL_s$ ) we must look if production process is able to output material just as designer wants ( $L$  and  $U$ ).

For this test we must collect  $N$  data related to the character of interest and compute  $S^2 = \sum (X_i - \bar{X})^2 / (N-1)$ , the variance of the last  $N$  items produced and compare  $S^2$  with  $\sigma_0^2$ . If  $S^2 < \sigma_0^2$  the process is capable.

According to capability studies experience it is better to accept the process if  $S^2 / \sigma_0^2 < 1.23$ . We did not use here the so called natural tolerance concept because it is enough to compare directly variances in order to have the test accomplished.

If production process is not able we suggest an innovative maintenance keeping into account cost embroiled with this operation.

To complete Factory needs list we remember  $P_1$  known as *Lot Tolerance Percent Defective*;  $\alpha$ , the producer's risk or first type error probability;  $\beta$ , the consumer's risk or second type error probability. All these values must be contractually chosen.



## 5. An example

Let us take some data based on the: "Inside diameter for automobile engine piston rings" (Montgomery, 1991, pag. 234).

We have:  $\bar{X} = 74.001$ ;  $'S = 0.0090$ ;  $''S = 0.0099$ ;  $'''S = 0.0101$ . The limits for the  $\bar{x}$  chart are:

$$UCL_{\bar{x}} = \bar{X} + A_3 'S = 74.001 + (1.427) (0.009) = 74.014;$$

$$LCL_{\bar{x}} = \bar{X} - A_3 'S = 74.001 - (1.427) (0.009) = 73.988;$$

and for the  $S$  chart:

$$UCL_S = B_4 'S = (2.089)(0.009) = 0.019.$$

If we introduce :  $L = 73.981$ ;  $U = 74.021$ ;  ${}_L P_o = 1\%$   ${}_U P_o = 1\%$ , ( $P = 2\%$ ) we can compute the new parameters according to our proposal. We consider:

$$z_L = -2.32635; \quad z_U = 2.32635; \quad \chi_{0.002}^2 = 16.92386;$$

$$\bar{X}_0 = (L \cdot z_U - U \cdot z_L) / (z_U - z_L) = 74.001;$$

$$\sigma_0 = (U - L) / (z_U - z_L) = 0.0086.$$

Therefore:

$$UCL_{\bar{x}} = 74.001 + 3 (0.0086) / \sqrt{5} = 74.01254;$$

$$LCL_{\bar{x}} = 74.001 - 3 (0.0086) / \sqrt{5} = 73.98946;$$

$$UCL_s = \sigma_0 \sqrt{\chi^2 / (n-1)} = 0.0086 \sqrt{16.92386} = 0.0177.$$

We note that our  $UCL_s$  is based on the  $\chi^2$ -distribution as suggested by Duncan (1965). Our limits are slightly narrower than Montgomery's ones, but if factory needs are the declared ones ( $L$  and  $U$ ) we have a production process not capable. Here is very important the designer responsibility because a little larger tolerances would change the situation.

We repeat the observation outlined above. Customary control charts privileges process capability. In presence of a chart ( $UCL$  and  $LCL$  for  $\bar{x}$  and  $s$ ) we

must verify if designer's needs (L and U) are satisfied. On the contrary with our control chart designer needs are privileged but we do not know if production process is capable. In order to get this last piece of information we must compare  $\bar{S}$  with  $\sigma_0$ . Of course if  $\bar{S} < \sigma_0$  the process can satisfy designer's needs; on the contrary we must solve the trouble. For the process capability analysis many references are given by Montgomery (1991).

## 6. Use of the chart

In order to use the chart we draw a sample with  $n = 5$  and compute  $\bar{x}$  and  $s$ . With the following data ; 74.002; 73.990; 73.997; 74.003, 74.001, we obtain:  $\bar{x} = 74.002$   $s = 0.002588$ . The points are within limits so that production is good. Let us now try to use the  $s^2$  chart as proposed by Duncan. We have:  $s^2 = 0.0000067$  and  $(UCL_s)^2 = 0.000079 \cdot 4.230965 = 0.000313$ ; the point is within limits as before. If we suppose to have a point very near the limit e.g.  $s = 0.0176$ , squaring it we get:  $\sigma^2 = 0.00030976$  and it is also within limits. If the point is just out of control e.g.  $s = 0.0178$ , we get  $s^2 = 0.00031684$  and the point is just out of control also in the new chart.

## 7. Conclusions

SCC (Shewart's Control Chart) is based on process ability to produce wanted items. Indeed, if control limits ( $UCL_{\bar{x}}$   $LCL_{\bar{x}}$   $UCL_s$ ) are computed on either  $\bar{S}^2$  or  $\bar{\bar{S}}^2$ , it is not worthy to insist on process ability. Clearly there is also the designer and their needs (L,U) to be considered so we must consider a capability study to test if they are in accordance.

We have seen how it is possible to get new control limits based on  $\bar{\bar{X}}_0$  and  $\sigma_0$  keeping into account designer's needs (L and U). Items must be output by production process so that now must look if it is able to do its work.

A different subject is the dispersion estimate related to SCC taking into account the presence of  $m$  lots. We have seen that  $s$  is a biased statistic very cumbersome to be adjusted. A simulation leads to prefer  $\bar{\bar{S}}^2$  but this means to use a  $\sigma^2$  chart instead of a  $\sigma$  chart. Nevertheless in our example we used  $\bar{S}$  and  $\bar{S}^2$  in order to simplify the discussion.

It seems worthy to remember that customary control chart construction privileges production process capability whereas our suggestion privileges designer needs. In every case we must verify the second coin's face to compare  $\bar{S}$  with  $\sigma_0$  or better  $\bar{\bar{S}}$  with  $\sigma_0$ . Better else to compare  $\bar{\bar{S}}^2$  with  $\sigma_0^2$  because  $\bar{\bar{S}}^2$  is a  $\sigma^2$  unbiased estimate.

## References

- Duncan, A.J. (1965). *Quality Control and Industrial Statistics*. Richard D. Irwing, Inc. Homewood, Ill.
- Kenney, A.F. & Keeping, E.S. (1956). *Mathematics of Statistics*. D. Van Nostrand Co. Inc., Princeton.
- Mittag, H.G. & Rinne, H. (1993) *Statistical Methods of Quality Assurance*. Chapman & Hall, London.
- Montgomery, D.C. (1991). *Introduction to Statistical Quality Control*. John Wiley & Sons. New York.
- Piccari, P.L. (1974). *Manuale di Controllo di Qualità e Affidabilità*. ISEDI, Milano.
- Rouzet, G. (1957). *Courbes d'Efficacité de la Carte de Contrôle pae Mesures en Fonction du Pourcentage de Pièces Défectueuses*. Revue de Statistique Appliquée, vol. V, 2, 19-32.
- Shewart, W.A. (1931). *Economic Control of Quality of Manufactured Product*. D. Van Nostrand Co., Inc. New York.
- Wold H., (1955) *Random Normal Deviates*. Department of Statistics University of London. Tracts for Computers Edited by E. S. Pearson D. Sc. n.XXV.

**AUTHOR INDEX**

**A**

Alfò .....73  
 Amenta .....263

**B**

Baiocchi .....237  
 Bellacicco .....35  
 Bolasco .....237  
 Borra .....345  
 Bove .....131

**C**

Caggese .....81  
 Calò .....97  
 Camiz .....139  
 Capiluppi .....367  
 Capobianchi .....211  
 Cerbara .....43  
 Cerioli .....3  
 Chiodi .....247  
 Coli .....255  
 Corazzari .....171  
 Cornillon .....263

**D**

D'Esposito .....279  
 D'Urso .....11  
 Di Battista .....19  
 Di Ciaccio .....345  
 Di Marzio .....287  
 Di Spalatro .....19

**E**

Esposito F. ....81  
 Esposito V. ....179

**F**

Fabbris .....367  
 Ferri .....55

**G**

Gazzei .....271  
 Giacalone .....295  
 Giusti .....221

**I**

Iacovacci .....49

Iezzi .....27  
 Ingrassia .....89  
 Ippoliti .....255

**J**

Jona-Lasinio .....211

**L**

Lafratta .....287  
 Laghi .....303  
 Lemmi .....271  
 Lizzani .....303  
 Lombardo .....187

**M**

Malerba .....81  
 Mantovan .....311  
 Maturo .....55  
 Miglio .....105  
 Millioli .....63  
 Mineo .....247;353  
 Mola .....113  
 Montanari .....97; 147  
 Morrone .....237

**N**

Nissi .....255

**P**

Pallini .....319  
 Pastore .....311  
 Petrucci .....221  
 Pillati .....105  
 Pittau .....11  
 Plaia .....353  
 Porzio .....327  
 Postiglione .....73

**Q**

Quintano .....155

**R**

Ragozini .....279  
 Rocci .....131  
 Romagnoli .....229  
 Roverato .....335

*S*

Sabatier .....	263
Scarabello .....	367
Scepi .....	179
Semeraro.....	81
Siciliano.....	121
Soffritti.....	147

*T*

Tessitore.....	187
Tonnellato.....	311
Turrini .....	361

*V*

Verde.....	195
Vichi.....	27;163
Vittadini.....	203

**KEY WORDS**

**A**

adaptive cluster sampling .....19  
 Akaike criterion .....121  
 array of data or cubic matrices .....171  
 asymmetry .....131  
 automatic interaction detector .....367

**B**

Bayes rule .....121  
 bias .....19  
 bivariate time series .....105  
 bootstrap .....19  
 B-spline function .....195

**C**

classification trees .....367  
 cluster .....35  
 cluster analysis .....27; 147  
 cluster stability .....3  
 cluster validity .....63  
 clusters analysis .....221  
 co-inertia .....179  
 combining .....105  
 conditional independence .....335  
 consensus classification .....11  
 constrained principal component analysis .....187  
 contiguity constraints .....221  
 convex hull .....279  
 covariance .....319  
 cross-validation .....319  
 crossvariogram .....211  
 curse of dimensionality .....287

**D**

data analysis .....295  
 deletion diagnostics .....3  
 dimensional scaling (DS) .....155  
 discrimination and classification .....81  
 dissimilarities .....147  
 distance between time series .....11  
 distance-based regression model .....303  
 distances .....139  
 dominant eigenvalue scores (DES) .....155  
 dynamic linear models .....311

**E**

eigenanalysis .....139  
 evaluation urban projects .....55  
 event .....35  
 expected proportion of samples in kernel supports .....287

**F**

factorial analysis .....171  
 financial time series .....345  
 firm performance .....271  
 forecast .....345

functional data analysis .....319  
 fuzzy .....35  
 fuzzy average linkage .....43  
 fuzzy clustering .....63  
 fuzzy c-means algorithm .....49  
 fuzzy methods .....55

**G**

Gauss-Markov .....229  
 Geary coefficient .....263  
 generalised canonical analysis .....195  
 generalised additive models .....113  
 graph .....263  
 graphical displays .....179  
 graphical methods .....131  
 graphical models .....335

**H**

hierarchical cluster analysis .....43  
 hierarchical clustering time series .....11  
 hyperstructures .....55

**I**

identification .....203  
 independence graph .....335  
 indeterminacy .....203  
 interpretation .....35

**K**

Kalman filter .....229; 255  
 kernel density estimation .....319  
 Kronecker product .....263

**L**

latent variable .....121  
 local influence .....327  
 loess .....113  
 logistic discrimination .....89

**M**

marginal model plots .....327  
 matrix completion .....335  
 MDS .....353  
 measures .....353  
 measures of fuzziness .....63  
 method of moments estimators .....311  
 missing data .....255  
 mixed predictors .....303  
 mixing parameter .....121  
 MORALS .....303  
 multidimensional scaling .....147  
 multidimensional scaling(MDS) .....155  
 multigraphs .....147  
 multiple qualitative response variable .....121  
 multiple sets .....179  
 multivariate adaptive splines .....187

*N*

Neighbourhood operator.....	263
network.....	35
neural.....	35
neural networks.....	105; 345
nonhierachical clustering.....	3
nonparametric discriminant analysis.....	97
non-parametric models.....	345

*O*

orthogonal projections.....	179
outliers.....	3; 279

*P*

panel data.....	27; 271
parallel Kalman filters.....	311
parameter estimation.....	89
partitioning.....	27
pattern recognition.....	81
posterior distribution.....	335
pre-processing.....	237
principal component analysis.....	211
principal co-ordinate analysis.....	303
principal surface.....	187
profession market.....	155
projection pursuit.....	97
projection pursuit density estimation.....	97
projection pursuit regression.....	303
qualitative variables.....	203

*R*

rand index.....	3
random fiels.....	229
regression and autoregressive models.....	171
regression tree.....	113; 367
regressogram.....	113
resampling.....	19
restricted regression component decomposition method.....	203

*S*

segmentation analysis.....	367
----------------------------	-----

semantic.....	35
semi-fuzzy classification.....	49
separability measures.....	89
Shewart's control chart.....	295
sigma's estimate.....	295
similarity.....	43; 353
singular value decomposition.....	211
smoothing.....	319
soft clustering.....	49
software.....	237
spatial processes.....	211
spline approximation.....	187
spline smoothers.....	113
stalactite plot.....	3
state-space model.....	255
STATIS method.....	263
statistical data analysis.....	367
stock location assignment.....	353
structural model.....	203
subpopulations.....	327
supervised classifiers.....	105
symbolic data analysis.....	81
symbolic objects.....	195
system parameter estimation.....	311

*T*

technical efficiency.....	271
telephone surveys.....	221
textual variables.....	237
three way data.....	263
three way environmental data matrix.....	255
time series.....	171
three-way Asymmetric scaling.....	131

*U*

unilateral representation.....	229
--------------------------------	-----

*V*

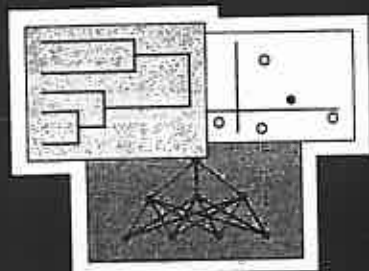
variable selection.....	63
-------------------------	----

*W*

Wilks statistic.....	279
----------------------	-----

## Classification and Data Analysis

The book provides new developments in classification, data analysis and multidimensional methods, topics which are of central interest to modern statistics. A wide range of topics is considered including methodologies in classification, fuzzy clustering, discrimination, regression tree, neural networks, proximity methodologies, factorial methods, spatial analysis, multiway and multivariate analysis.



ISBN 3-540-65633-2



9 783540 656333

<http://www.springer.de>