CRACOW UNIVERSITY OF ECONOMICS

FOUNDATION OF THE CRACOW UNIVERSITY OF ECONOMICS

KNOWLEDGE ECONOMY SOCIETY

SELECTED CHALLENGES FOR STATISTICS IN CONTEMPORARY MANAGEMENT SCIENES

Edited by
Daniel Kosiorowski, Małgorzata Snarska

KNOWLEDGE – ECONOMY – SOCIETY

SELECTED CHALLENGES FOR STATISTICS IN CONTEMPORARY MANAGEMENT SCIENCES

CRACOW UNIVERSITY OF ECONOMICS
Faculty of Management
FOUNDATION OF THE CRACOW UNIVERSITY OF ECONOMICS

KNOWLEDGE – ECONOMY – SOCIETY

SELECTED CHALLENGES FOR STATISTICS IN CONTEMPORARY MANAGEMENT SCIENCES

Edited by

Daniel Kosiorowski, Małgorzata Snarska

Cracow 2016

Reviewers

Tomasz Górecki
Grzegorz Kończak

All papers have been prepared in English by the Authors

Wydanie publikacji zostało sfinansowane z dotacji na utrzymanie potencjału badawczego przyznanej Uniwersytetowi Ekonomicznemu w Krakowie

The book was financed with subsidies for maintaining the research capacity granted to the Cracow University of Economics

ISBN 978-83-65173-38-6 (printed version)
ISBN 978-83-65173-39-3 (pdf on–line)

Publishing House
Foundation of the Cracow University of Economics
ul. Rakowicka 27, 31–510 Kraków, Poland

Table of contents

Paweł Fiedor
Noise-robustness of Networks in Financial Markets 9

Beata Koń, Michał Jakubczyk
Performance of Random Forests in Explaining the Health-related Utility of Life 25

Ewa Szlachtowska, Dominik Mielczarek, Tomasz Burchard
The Robust Clustering Methods 41

Michał Miłek
Simulation Comparison of Methods for Estimation the Forecasts Intervals for Time Series 55

Agnieszka Gołdyn, Agnieszka Góra, Robert Olechnowicz
Exploration of Internet User Activity with the {sm} Package 65

Dominika Polko, Grzegorz Kończak
On Using Permutation Tests in the Data Homogeneity Analysis 79

Daniel Kosiorowski, Jerzy Rydlewski, Małgorzata Snarska
A Note on the Nonstationary Functional Time Series 89

Sławomir Śmiech
Why Forecasting of Crude Oil Price is Difficult Task? Results from Comparison of Large Set of VAR Models 103

Angelina Rajda-Tasior
The Application of the Matrix Flow Diagram for the Errors in a Quality Control in Production Area 117

Weronika Toszewska-Czerniej
Human Capital Management in Services Providing Enterprise 129

Małgorzata Szerszunowicz
On Non-classical Methods of Design of Experiments Using the R Program 145

Introduction

Contemporary changes of global market structure, introduction of new communication and transaction technologies as well as changes in socio, demographic and cultural aspects of our life force economists and managers to looking for new statistical tools enabling decision makers for dealing with huge data sets appearing in non-equally spaced in time packets in empirical finance, for new decision procedures dedicated for public health management purposes, which take into account pour quality or missing data or small knowledge about considered phenomena.

Modern managerial sciences give incentives for new understanding of causality in a context of labour market regulation or evaluation of prediction quality in energy economics. This part of the publication refers to the problem of using nonparametric and robust statistical methods in contemporary managerial sciences. This book is book presents highlights of the First Cracow Seminar on Application of Robust and Nonparametric Methods in Economics, which took place in May 2015 at campus of Cracow University of Economics. The book has a character of empirical as well as methodological studies whose aim is the presentation and systematization of the scientific and practical outputs from selected analyzed issues.

Handling the discussed work to the readers, we express our belief that the publication in the presented formulation is fully justified both for theoretical as well as practical and methodological reasons. It can be a starting point for new proposals, new scientific meetings, disputes and critical discussions over the presented problems.

The involvement of a large group of participants in the first Cracow seminar on applications of robust and nonparametric methods in economics enabled showing the discussed issues in a broad and many-sided way. As the scientific editors of this study, we would like to thank cordially all the Authors, Speakers and Referees for accepting our invitation to co-create the seminar and the publication.

Daniel Kosiorowski and Małgorzata Snarska

Noise-robustness of Networks in Financial Markets

Paweł Fiedor

Cracow University of Economics, Poland

Abstract

Over the last decade there has been considerable effort put into investigating how network theory may be employed to unravel the complex nature of financial markets. These efforts branch out in a few directions, with the main area studying correlation-based networks. These networks are used to simplify the complexity of financial dependencies, and consequently to better understand the dynamics within financial markets. While ample effort is being put into discussing the applications of these methods to various financial markets, the methods used to form correlation-based networks are still being developed as well. In fact, the name itself is a misnomer. While most such networks are based on Pearson's correlation coefficient, this is by no means a requirement, and the choice of appropriate measure of dependence is one of the most important, yet least investigated, problems in this field of study. There are a few problems with Person's correlation, for example it is susceptible to outliers, and it is not sensitive to non-linear dependencies. With regards to the latter, an approach allowing for the inclusion of non-linear dependencies between financial assets has been recently introduced, based on information-theoretic concept of mutual information. It has been shown that, in practical applications, such an approach has certain advantages over the one based on Pearson's correlation. With regards to the former, while entropy and mutual information are rather insensitive to outliers, a simpler approach may also be employed, based on the biweight mid-correlation. Sensitivity to outliers and non-linear dependencies are not the only issues however. The networks are based on market data which contains noise. Such noise stems both from errors in databases containing financial times series, and from the complex, noisy behaviour of the markets themselves. Thus it would be desirable if the method of constructing

financial networks would be robust to noise. In this paper we investigate how the three mentioned dependence measures used to create financial networks behave when the data is contaminated with both additive and multiplicative noise. We create Minimal Spanning Trees based on time series describing stock log returns on Warsaw's stock exchange between 2005 and 2013, and use them to test the robustness of these methods with regards to noise. We show how both the dependence measures and the structure of the resulting networks are affected by various amounts and types of noise.

Keywords: noise-robustness, network, financial markets

1. Introduction

Due to the involvement of large amounts of people, financial markets are not only complex, but also complex adaptive systems. Over the last decade there has been considerable effort put into investigating how network theory may be employed to unravel the complex nature of financial markets. These efforts branch out in a few directions, with the main area studying correlation-based networks. These networks are used to simplify the complexity of financial dependencies, and consequently to better understand the dynamics within financial markets, particularly for financial instruments traded on stock markets [15, 16]. These studies can unravel the underlying principles guiding financial markets, but are also useful for practical risk and investment assessments. Econophysicists have created a method of analysing financial markets based on single linkage clustering analysis, commonly based on Pearson's correlation coefficients. Such correlation structures have been employed in studying time series describing stock returns [20, 17, 13], market index returns [2, 22] and currency exchange rates [18]. While ample effort is being put into discussing the applications of these methods to various financial markets, the methods used to form correlation-based networks are still being developed as well. In fact, the name itself is a misnomer. While most such networks are based on Pearson's correlation coefficient, this is by no means a requirement, and the choice of appropriate measure of dependence is one of the most important, yet least investigated, problems in this field of study. There are a few problems with Person's correlation, for example it is susceptible to outliers, and it is not sensitive to non-linear dependencies. With regards to the latter, due to overwhelming evidence

of non-linear behaviour in financial markets [3, 21, 24], an approach allowing for the inclusion of non-linear dependencies between financial assets has been recently introduced, based on information-theoretic concepts of mutual information and mutual information rate [8, 9]. It has been shown that, in practical applications, such an approach has certain advantages over the one based on Pearson's correlation. With regards to the former, while entropy and mutual information are rather insensitive to outliers, a simpler approach may also be employed, based on the biweight mid-correlation. Sensitivity to outliers and non-linear dependencies are not the only issues however. The networks are based on market data which contains noise. Such noise stems both from errors in databases containing financial times series, and from the complex, noisy behaviour of the markets themselves [1, 12, 10, 4]. Thus it would be desirable if the method of constructing financial networks were robust to noise. In this paper we investigate how the three mentioned dependence measures used to create financial networks behave when the data is contaminated with both additive and multiplicative noise. We create Minimal Spanning Trees based on time series describing stock log returns on Warsaw's stock exchange between 2005 and 2013, and use them to test the robustness of these methods with regards to noise. We show how both the dependence measures and the structure of the resulting networks are affected by various amounts and types of noise.

This paper is organised as follows. In Section 2 we present the methods used in the analysis. In Section 3 we present the dataset used, obtained results, and the discussion of these. In Section 4 we conclude the study and propose further research.

2. Methods

In this section we present the three mentioned dependence measures (Pearson's correlation, biweight mid-correlation and mutual information), their appropriate distance measures, as well as the method for using them to construct Minimal Spanning Trees. We also discuss the noise we employ in the analysis.

The topological arrangement of stocks (or any other well-defined objects) within hierarchical structures is usually based on the empirical Pearson's correlation coefficient matrix, usually calculated from logarithmic returns for financial data, as the prices themselves are not stationary. Pearson's cor-

relation coefficient is estimated for all pairs within studied dataset, and for a pair of random variables (X,Y) is defined as:

$$\rho_{X,Y} = \frac{E(XY) - E(X)E(Y)}{\sqrt{(E(X^2) - E(X)^2)(E(Y^2) - E(Y)^2)}} \tag{1}$$

where X and Y are, for practical purposes, the time series describing log returns of the two studied financial instruments; and E is the expectation (mean).

Pearson's correlation coefficient is a similarity measure, and as such is not convenient to directly form the topology of a network. For this purpose we need an Euclidean metric. Usually the below metric is being used, based on the above-defined Pearson's correlation coefficient [17]:

$$\delta(X,Y) = \sqrt{2\left(1 - \rho_{X,Y}\right)} \tag{2}$$

For the purpose of creating networks more robust to outliers the Pearson's correlation coefficient may be exchanged for biweight mid-correlation. To our best knowledge, this is the first study of asset networks using this dependence measure.

There are many disadvantages attributed to Pearson's correlation, one of the most prevalent being that it is susceptible to outliers. Several robust alternatives have been proposed, including the Spearman correlation or the biweight mid-correlation [28], the latter of which we use in this study. To define the biweight mid-correlation of two vectors X, Y with components x_a, y_a (a = 1, 2, ..., m); we need to introduce u_a, v_a defined as:

$$u_a = \frac{x_a - med(X)}{9mad(X)} a \tag{3}$$

$$v_a = \frac{y_a - med(Y)}{9mad(Y)} \tag{4}$$

where med(X) is the median of X, and mad(X) is the median absolute deviation of X. mad(X) is the raw median absolute deviation of X, that is without the correction factor for asymptotic consistency of mad and standard deviation. Further, weights $w_a^{(x)}$ for x_a may be defined as:

$$w_a^{(x)} = (1 - u_a^2)^2 I(1 - | u_a |) \tag{5}$$

where the indicator function $I(1 - |u_a|)$ equals 1 if $1 - |u_a| > 0$ and 0 otherwise. As such, the weight $w_a^{(x)}$ is close to 1 if x_a is close to med(X), and approaches 0 when x_a differs from med(X) by nearly 9mad(X), and is zero if x_a differs from med(X) by more than 9mad(X). Weight $w_a^{(y)}$ is analogously defined for each y_a. The biweight mid-correlation of X and Y, bicor(X, Y), is then defined as

$$bicor(X,Y) = \frac{\sum_{a=1}^{m}(x_a - med(X))w_a^{(x)}(y_a - med(Y))w_a^{(y)}}{\sqrt{\sum_{b=1}^{m}[(x_b - med(X))w_b^{(x)}]^2}\sqrt{\sum_{c=1}^{m}[(y_c - med(Y))w_c^{(y)}]^2}} \quad (6)$$

The factor of 9 multiplying mad in the denominator is a standard choice, discussed in Wilcox [28]. Details of mathematical properties of biweight mid-correlation can be found in Langfelder & Horvath [14]. We use the implementation provided in package WGCNA within R. The distance for biweight mid-correlation is defined analogously to the one defined for Pearson's correlation coefficients.

Finally, in order to include non-linear relationships, which are prevalent in financial markets, we proposed to base the topology of financial networks on mutual information between logarithmic returns for the studied financial instruments [9], which procedure is explained below. Mutual information denotes the amount of information two stochastic processes share, or, in other words, by how much information about one stochastic process reduces uncertainty about the other process. Mutual information can be defined for two random variables X and Y as [5]:

$$I(X;Y) = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \quad (7)$$

where p(x,y) denotes joint probability for X and Y, while p(x) and p(y) denote marginal probabilities. Mutual information may also be defined in terms of Shannon entropy H:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \quad (8)$$

where H(X) and H(Y) denote Shannon entropies, and H(X,Y) denotes joint Shannon entropy of X and Y, defined below.

$$H(X) = -\sum_{x\in X} p(x)\log p(x) \quad (9)$$

$$H(X,Y) = -\sum_{y\in Y}\sum_{x\in X} p(x,y)\log p(x,y) \tag{10}$$

Mutual information is non-negative and I(X;X)=H(X).

Having the definition, we also need an estimator of entropy for practical purposes. In this study we use the Schurmann-Grassberger estimate of the entropy of a Dirichlet probability distribution, implemented in infotheo package in R. There are plenty of estimators to choose from, most of which have comparable characteristics, yet the above-mentioned seems to be most suited for general purposes [19]. Dirichlet distribution, upon which the estimator is based, is the generalisation of the Beta distribution. The density of this distribution is described by:

$$p(X;\Theta) = \frac{\prod_{i\in\{1,2,\ldots,|X|\}}\Gamma(\Theta_i)}{\Gamma(\sum_{i\in\{1,2,\ldots,|X|\}}\Theta_i)}\prod_{i\in\{1,2,\ldots,|X|\}} x_i^{\Theta_i-1} \tag{11}$$

where Θ_i is is the prior probability of an event x_i, which is the i-th element in the set , and $\Gamma(.)$ is the gamma function. The entropy can then by estimated by:

$$\hat{H}(X) = \frac{1}{m+ \mid X \mid N}\sum_{x\in X}(\#(x)+N)(\psi(m+ \mid X \mid N+1) - \psi(\#(x)+N+1)) \tag{12}$$

where #(x) is the number of data points with value x, |X| is the number of bins from the discretisation step, m is the sample size, and $\psi(z)=d\ln\Gamma(z)/dz$ is the digamma function. Various choices of weighing factor N have been proposed, but the Schurmann–Grassberger estimator usually assumes $N=1/|X|$ as the prior, and we follow this practice [23]. This estimator gives virtually the same results as empirical distribution estimator and Miller-Madow estimators, thus the analysis is robust with regards to this choice (in particular the prior).

We also need an Euclidean metric based on mutual information. We cannot use the metric based on correlation, as mutual information is bound by 0 and max(H(X);H(Y)), and not by -1 and 1. Fortunately, such metrics are well known in information theory [5]. In particular:

$$d(X,Y) = H(X,Y) - I(X;Y) \tag{13}$$

$$d(X,Y) = H(X) + H(Y) - 2I(X;Y) \tag{14}$$

is an Euclidean metric. There are other metrics proposed in the literature (for example when the time series are of different lengths), but for our analysis this one is sufficient.

Having defined the distance measures for financial instruments based on Pearson's correlation, biweight mid-correlation, and mutual information, we may construct filtered graphs used to analyse financial networks. Matrices containing these distances may be treated as distance matrices of full graphs. Due to the complexity of these graphs (and the financial systems), such graphs need to be filtered to be useful for analysis. Such graphs may be either filtered by using a threshold on the distance metric, or by filtering based on topology. It is very hard to find one appropriate threshold for the whole financial systems, and the graphs resulting from setting a threshold don't have a satisfying structure, thus to analyse synchronous financial networks usually a filtering procedure based on the topology of the graphs is employed [17]. The most popular filtered graph is the Minimal Spanning Tree (MST), which is the smallest consistent subgraph. Having a distance matrix D filled with d(X,Y) or δ(X,Y) for all studied pairs of financial instruments, we may create a sorted list S, in which the distances are listed in increasing order. Then, to create a minimal spanning tree, we go through the list linearly, adding an arc to the graph if and only if the resulting graph is still a tree or a forest. After all connections have been added the resulting graph is guaranteed to reduce to a tree [25]. Such filtering allows analysts to concentrate on the most important information contained in the market data, and on the most important dependencies within the financial networks. Such procedure is not problematic, as the connections which are filtered out are mostly due to statistical noise [11]. For detailed description of these methods see presented references [25-27].

In our analysis we use logarithmic returns, to avoid issues with non-stationarity. If P(t) is the price at time t, then the logarithmic returns are defined as:

$$r(t) = \ln(P(t)) - \ln(P(t-1)) \tag{15}$$

For the purposes of entropy and mutual information, we need a discrete variable. Thus we discretise the log returns by binning them into quantiles. In this study we follow Fiedor [7, 9] in binning the data into quartiles. We note that the number of quantiles is largely irrelevant for this procedure, as discussed in Fiedor [7].

Finally, we add noise to the log returns in order to analyse the changes in the dependence measures stemming from the noise, as well as the changes in the constructed networks themselves. As we expect the noise on financial markets to be stemming from a large number of sources, we expect the noise to be Gaussian (in accordance with CLT). We analyse both additive and multiplicative noise. Thus, for the former, we add to each log return a term drawn from $N(0,\sigma)$. For the latter we multiply each log return by a term drawn from $N(1, \sigma)$. We average the results over 50 realisations (realisations denote adding noise to the matrix containing log returns). We create minimal spanning trees for the original data, and for data contaminated with noise (as above) with varying degrees of parameter σ. In this way we analyse the robustness of the dependence measures and, more importantly, the constructed networks, to various amounts of noise.

3. Results and discussion

For empirical analysis we have obtained end-of-day prices for 53 stocks traded on Warsaw Stock Exchange (chosen due to least amount of missing data). The list of stocks appears in Appendix A. This database has been obtained from DM BOŚ website[1]. The time series contain data for dates between 18th October 2005 and 5th July 2013, thus leaving us with time series of length 1884. We transform these time series describing prices into logarithmic returns, and discretise them into four quantiles for the purpose of estimating mutual information. We then calculate minimal spanning tree based on Pearson's correlation, biweight mid-correlation, and mutual information. We contaminate all log returns with additive and multiplicative (separately) Gaussian noise for various values of σ. For both cases we do this 50 times, and average the results over all realisations. We calculate minimal spanning trees for the noisy matrices as well, and compare them to the original graphs, and each other as well.

[1] http://bossa.pl/notowania/metastock/

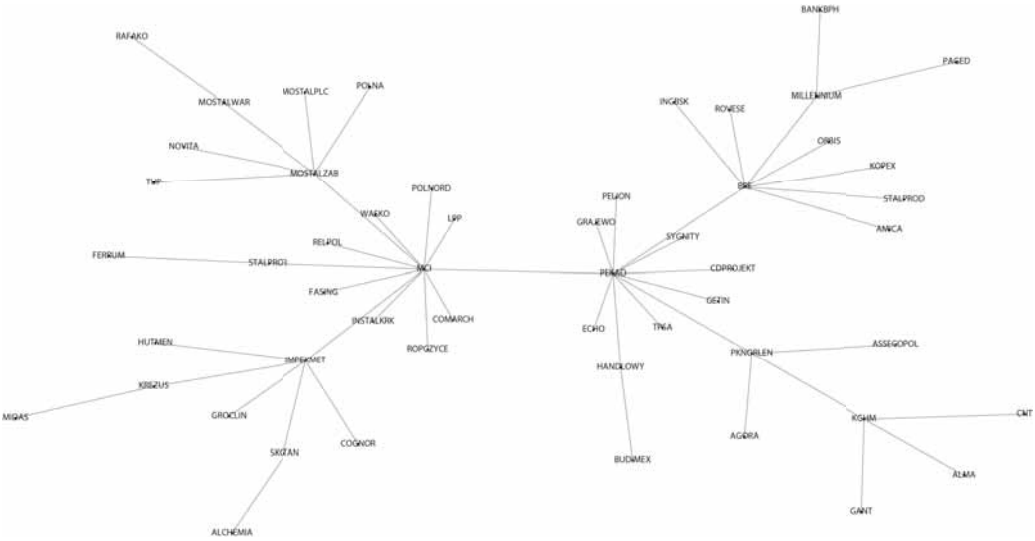


Figure 1: Minimal Spanning Tree based on Pearson's correlation coefficients for Warsaw Stock Exchange data between 2005 and 2013 (with no added noise)

In Figure 1 we show an example of a Minimal Spanning Tree created for the above-mentioned data without adding any noise. The presented network is based on Pearson's correlation. Hubs can be easily spotted, which means these are not random graphs. It is very difficult to approximate which dependency measure creates better networks, as we have no theoretical picture of how such networks should look like for a given market. But we have some characteristics we would like that the created networks have. One of such is an appropriate degree distribution, preferably we would like to obtain a scale-free network. Thus we compare degree distributions for networks created with the three mentioned dependence measures, as shown in Figure 2. We note that all three methods create networks with fat-tailed degree distributions. Power laws within these figures have been fitted using poweRlaw package in R. The p-values of Kolmogorov Smirnoff test are as follows: 0.1163133 for Pearson's correlation, 0.1165243 for bi-weight mid-correlation, and 0.07810688 for mutual information. Thus, we see (albeit accounting for the fact that the networks are relatively small and thus the test is not as convincing as it would be had we performed it on a much larger network) that mutual information works best in this respect, creating networks closest to scale free networks. Further, correlation networks recreate sector structure from prices in a way which cannot

be reproduced by simulating a market, and as such this is an important feature of such networks. Thus we also compare the percentage of intrasector links (links between companies listed as belonging to the same economic sector by the Warsaw's exchange) in all links within the created graphs, and compare these values for the three created networks. The percentage of intrasector links equals 19.23% for Pearson's correlation-based MST, 19.23% for biweight mid-correlation-based MST (the values is the same but the networks are different), and 26.92% for mutual information-based MST. For comparison, it's under 10% for an unfiltered network containing all edges. It is very difficult to assess the statistical significance of these differences, nonetheless these results hint that mutual information is a better tool for creating asset networks than Pearson's correlation or biweight mid-correlation. We have not yet assessed their robustness to noise however.

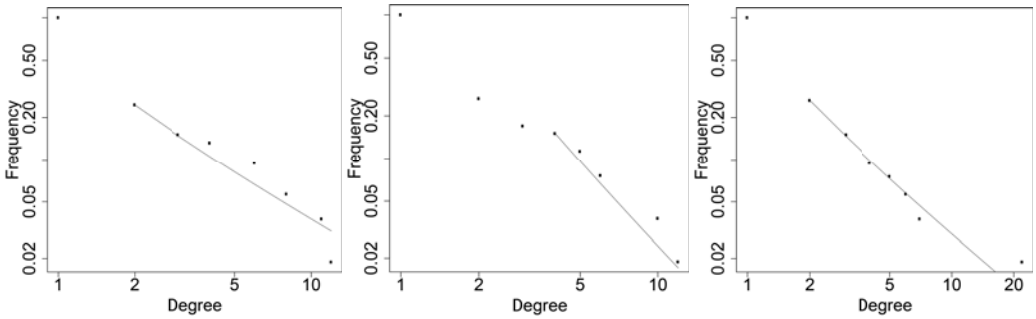


Figure 2: Degree distribution for Minimal Spanning Tree based on Pearson's correlation coefficients (on the left), biweight mid-correlation (in the middle), and mutual information (on the right) for Warsaw Stock Exchange data between 2005 and 2013 (with no added noise), together with fitted power law

First, we investigate the effects of multiplicative noise. In Figure 3 we show Pearson's correlation coefficients between the three dependence measures (Pearson's correlation on the left, biweight mid-correlation in the middle, mutual information on the right) calculated for all pairs of studies stocks, with various amounts of multiplicative noise. Appropriate values of σ are presented on the diagonal, 0 denotes the original time series with no added noise. While the mutual information seems to be more robust with regards to multiplicative noise, all measures of dependence seem very robust indeed, and as such we don't delve into more details, and instead move on to the analysis of additive noise, which may prove more insightful.

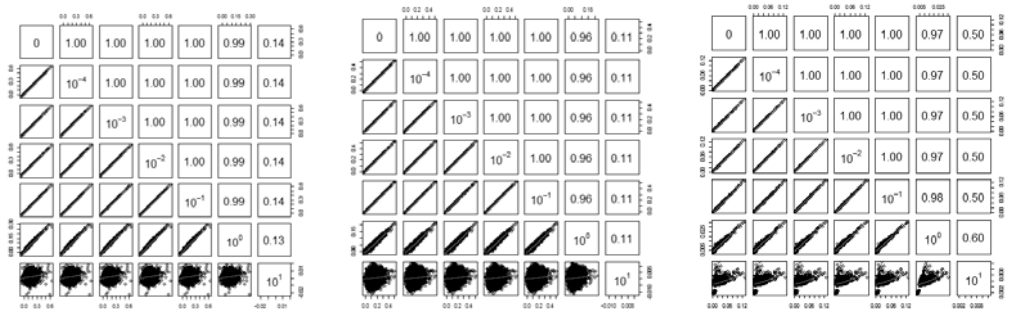


Figure 3: Correlations between dependence measures (Pearson's correlation on the left, biweight mid-correlation in the middle, mutual information on the right) calculated for all pairs of studied stocks for log returns with multiplicative Gaussian noise (N(1, σ)), with various levels of σ shown on the diagonal

In Figure 4 we show Pearson's correlation coefficients between the three dependence measures (Pearson's correlation on the left, biweight mid-correlation in the middle, mutual information on the right) calculated for all pairs of studies stocks, with various amounts of additive noise. Appropriate values of σ are presented on the diagonal, 0 denotes the original time series with no added noise. All measures are less robust with regards to additive noise, which is not surprising. This time mutual information is the most sensitive measure, with both correlations firmly ahead. Though we must note that biweight mid-correlation appears to be more robust, albeit it's hard to address the question of significance of this result. We are more interested in the robustness of the networks rather than the dependence measures, thus we will also look into the robustness of the degrees of nodes within the minimal spanning trees with regards to additive noise. In Figure 5 we show Pearson's correlation coefficients between the node degrees within Minimal Spanning Trees based on the three dependence measures (Pearson's correlation on the left, biweight mid-correlation in the middle, mutual information on the right) calculated with various amounts of additive noise. Appropriate values of σ are presented on the diagonal, 0 denotes the original time series with no added noise. We note that while the results are similar to results for the dependence measures, the difference in robustness between the correlation measures and mutual information is far less pronounced for degrees, and as such it appears that while networks based on mutual information are less robust to additive noise than networks based on Pearson's correlation or biweight mid-correlation, the difference in sen-

sitivity is not as large as one would expect from looking at the dependence measures themselves. We conclude that mutual information may be preferable in all cases except where one expects significant additive noise, then biweigth mid-correlation may be advisable. We further note that mutual information is more sensitive to large amounts of missing data, thus in this cases we would advise using biweight mid-correlation as well.

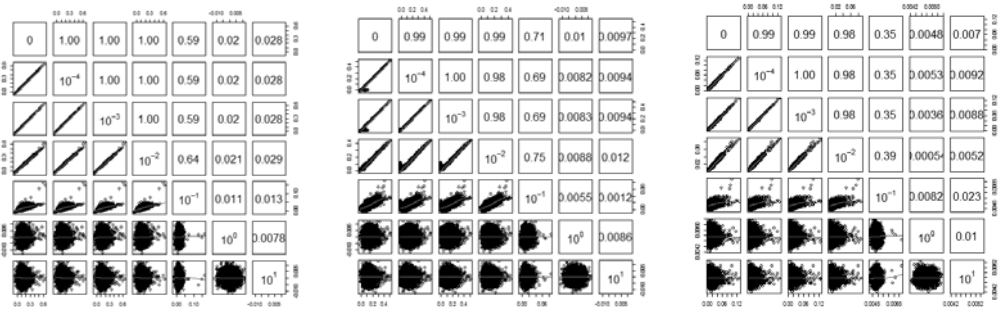


Figure 4: Correlations between dependence measures (Pearson's correlation on the left, biweight mid-correlation in the middle, mutual information on the right) calculated for all pairs of studied stocks for log returns with additive Gaussian noise (N(0, σ)), with various levels of σ shown on the diagonal

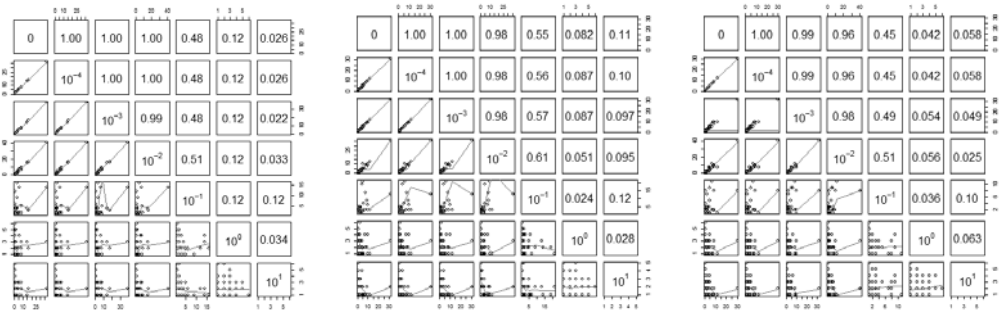


Figure 5: Correlations between node degrees within Minimal Spanning Trees based on Pearson's correlation (on the left), biweight mid-correlation (in the middle), and mutual information (on the right) based on log returns with additive Gaussian noise (N(0, σ)), with various levels of σ shown on the diagonal

4. Conclusion

We have presented the canonical method for creating asset networks by filtering graphs based on Pearson's correlation between financial assets into

Minimal Spanning Trees, as well as the recently introduced method based on mutual information, which accounts for non-linear dependencies, and is more robust with regards to outliers. We have also introduced a method based on biweight mid-correlation, which should be more robust with regards to outliers than the first of the mentioned methods, but doesn't account for non-linear dependencies. We have compared these methods by analysing the networks they produce, particularly concentrating on the robustness with regards to noise. It appears that networks based on mutual information provide better overall results, but are slightly more sensitive to additive noise. If additive noise is suspected to be a problem then an approach based on bigweight mid-correlation is advised, as it's the most robust with regards to additive noise out of the three methods. The most popular approach based on Pearson's correlation seems to perform the worst in all respects.

Appendix A

The time series describing prices of the below stocks have been used in the analysis: AGORA, ALCHEMIA, ALMA, AMICA, ASSECOPOL, BANKBPH, BRE, BUDIMEX, CDPROJEKT, CNT, COGNOR, COMARCH, ECHO, FASING, FERRUM, GANT, GETIN, GRAJEWO, GROCLIN, HANDLOWY, HUTMEN, IMPEXMET, INGBSK, INSTALKRK, KGHM, KOPEX, KREZUS, LPP, MCI, MIDAS, MILLENNIUM, MOSTALPLC, MOSTALWAR, MOSTALZAB, NOVITA, ORBIS, PAGED, PEKAO, PELION, PKNORLEN, POLNA, POLNORD, RAFAKO, RELPOL, ROPCZYCE, ROVESE, SKOTAN, STALPROD, STALPROFI, SYGNITY, TPSA, TUP, WASKO.

References

[1] Black, F. (1986). Noise, *The Journal of Finance*, 41(3), 529-543.

[2] Bonanno, G., Vandewalle, N., & Mantegna, R.N. (2000). Taxonomy of Stock Market Indices, *Physical Review E*, 62(6), 7615-7618.

[3] Brock, W.A., Hsieh, D.A., & LeBaron, B. (1991). *Non-linear Dynamics, Chaos, and Instability. Statistical Theory and Economic Evidence.* Cambridge: MIT Press.

[4] Burda, Z., & Jurkiewicz, J. (2004). Signal and Noise in Financial Correlation Matrices. *Physica A*, 344(1-2), 67-72.

[5] Cover, T., & Thomas, J. (1991). *Elements of Information Theory*. New York: John Wiley & Sons.
[6] Dorogovtsev, S.N., Goltsev, A.V., & Mendes, J.F.F. (2008). Critical Phenomena in Complex Networks. *Review of Modern Physics*, 80, 1275-1335.
[7] Fiedor, P. (2014a). *Frequency Effects on Predictability of Stock Returns*. In Proceedings of the IEEE Computational Intelligence for Financial Engineering & Economics 2014, 247-254, IEEE: London.
[8] Fiedor, P. (2014b). Information-theoretic Approach to Lead-lag Effect on Financial Markets. *European Physical Journal B*, 87(8), 1-9.
[9] Fiedor, P. (2014c). Networks in Financial Markets Based on the Mutual Information Rate, *Physical Review E*, 89(5), 052801.
[10] Guhr, T., & Kalber, B. (2003). A New Method to Estimate the Noise in Financial Correlation Matrices. *Journal of Physics A*, 36(12), 1-25.
[11] Kwapien, J., & Drożdż, S. (2012). Physical Approach to Complex Systems. *Physics Reports*, 515, 115-226.
[12] Laloux, L., Cizeau, P., Bouchaud, J., & Potters, M. (1999). Noise Dressing of Financial Correlation Matrices. *Physical Review Letters*, 83(7), 1-3.
[13] Laloux, L., Cizeau, P., Potters, M., & Bouchaud, J. (2000). Random Matrix Theory and Financial Correlations. *International Journal of Theoretical & Applied Finance*, 3, 391-398.
[14] Langfelder, P., & Horvath, S. (2012). Fast R Functions for Robust Correlations and Hierarchical Clustering. *Journal of Statistical Software*, 46(11).
[15] Mandelbrot, B.B. (1963). The Variation of Certain Speculative Prices. *Journal of Business*, 36(4), 394-419.
[16] Mantegna, R. (1991). Levy Walks and Enhanced Diffusion in Milan Stock Exchange. *Physica A*, 179, 232-242.
[17] Mantegna, R. (1999). Hierarchical Structure in Financial Markets. *European Physical Journal B*, 11, 193-197.
[18] McDonald, M., Suleman, O., Williams, S., Howison, S., & Johnson, N.F. (2005). Detecting a Currency's Dominance or Dependence Using Foreign Exchange Network Trees. *Physical Review E*, 72, 046106.
[19] Paninski, L. (2003). Estimation of Entropy and Mutual Information. *Neural Computation*, 15, 1191-1254.

[20] Plerou, V., Gopikrishnan, P., Rosenow, B., Nunes-Amaral, L.A., & Stanley, H.E. (1999). Universal and Non-universal Properties of Cross-correlations in Financial Time Series. *Physical Review Letters*, 83(7), 1471-1474.

[21] Qi, M. (1999). Nonlinear Predictability of Stock Returns Using Financial and Economic Variables. *Journal of Business & Economic Statistics*, 17(4), 419-429.

[22] Sandoval, L., & Franca, I.D.P. (2012). Correlation of Financial Markets in Times of Crisis. *Physica A*, 391, 187-208.

[23] Schurmann, T., & Grassberger, P. (1996). Entropy Estimation of Symbol Sequences. *Chaos*, 6, 414-427.

[24] Sornette, D., & Andersen, J. (2002). A Nonlinear Super-exponential Rational Model of Speculative Financial Bubbles. *International Journal of Modern Physics C*, 13(2), 171-188.

[25] Tumminello, M., Aste, T., Matteo, T.D., & Mantegna, R. (2005). A Tool for Filtering Information in Complex Systems. *Proceedings of the National Academy of Science USA*, 102(30), 10421-10426.

[26] Tumminello, M., Aste, T., Matteo, T.D., & Mantegna, R. (2007a). Correlation based Networks of Equity Returns Sampled at Different Time Horizons. *European Physical Journal B*, 55(2), 209-217.

[27] Tumminello, M., Coronnello, C., Lillo, F., Micciche, S., & Mantegna, R. (2007b). Spanning Trees and Bootstrap Reliability Estimation in Correlation-based Networks. *International Journal of Bifurcation & Chaos*, 17, 2319-2329.

[28] Wilcox, R.R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*. New York: Academic Press.

Performance of Random Forests in Explaining the Health-related Utility of Life

Beata Koń[a,b], Michał Jakubczyk[a]

[a] *Warsaw School of Economics, Poland*
[b] *Ministry of Health, Poland*

Abstract

Decisions regarding which health technologies should be used and reimbursed are important and require quantifying possible health outcomes due to introducing given technology. This quantification requires, inter alia, assigning utilities to the health states. Typically, information about societal preferences is elicited from some subset of all the possible health states (defined, e.g., with EQ-5D questionnaire), and then the results are extrapolated for other possible health states. Our aim is to verify the properties of such an extrapolation with respect to the available data. Our results show that omitting information about one health state does not influence the overall model performance, and that performance of random forest used for results extrapolation differs for various health states. We conclude, that health state characteristic and especially distribution of health-related utility in the particular health state determine model ability to extrapolate the results. This information is essential in terms of choosing subset of health states to be used in a valuation study.

Keywords: health-related utility, EQ-5D-3L, health-related quality of life, random forest

1. Introduction

The subject of the present paper is the modelling of the preferences towards the health states. The aim – to be made more precise later – is to identify the determinants of the quality of such models, i.e., to identify how

data available for modelling influence the quality of the resulting models. The motivation for the present study comes from that such models are being widely employed in actual decision making to quantify the benefits and cost-effectiveness of various health technologies.

The health care regulator has to decide which of the available treatments to reimburse from the public resources. Its decisions have to be based on the clearly specified criteria, so as to ensure transparency and reliability. They also need to be made for different types of diseases, and so require a unified framework to improve comparability, as money from a single budget is spent. A collection of methods and approaches often used to support such a decision making process is called health technology assessment (HTA). In Poland, HTA process is supervised and managed by the Agency of Health Technology Assessment and Tariff System (AOTMiT)[1], and final recommendations are reported to the Ministry of Health.

Polish law stresses that HTA process should be a multidisciplinary one and should encompass medical, social, and economic aspects [13, 16]. It should inform about the total burden of the disease that analysed technology concerns, include cost–effectiveness analysis describing relation between health outcomes and costs of the technology, and also should describe impact of introducing such technology on the total budget. One of the important elements of HTA is to measure how technology will affect the health-related quality of life (HRQoL). There are several approaches to define HRQoL. The EuroQol Group in 1990 [5] suggested using EQ-5D-3L. It takes into account 5 dimensions of health: mobility (MO), self-care (SC), usual activities (UA), pain/discomfort (PD), and anxiety/depression (AD). Each dimension is valued by selecting one of three levels: 1 – no problems, 2 – some problems, and 3 – severe problems, in a given dimension (in EQ-5D-5L five levels are used [6]). In EQ-5D-3L, the combinations of levels in 5 dimensions define 243 possible health states.

Introducing HRQoL to the cost-effectiveness analysis requires combining it with the longevity of life. That is often done using a concept of a QALY (quality adjusted life years) [2]. Within this model we need to measure the attractiveness of a health state as a single number, interpreted as a health-related utility of a year spent in that state (utility of being dead normalized to 0, utility of full health normalized to 1). Several methods have been proposed

[1] http://www.aotmit.gov.pl/

to elicit this utility, e.g., visual analogue scale (VAS), standard gamble (SG), or time trade–off (TTO) [4]. The VAS is based on a 'thermometer–type' scale with endpoints representing perfect and the worst health state (possibly death). Respondent places the other states and interval interpretation of the scale allows to assign utilities. SG is based on preferences measured under risk. Respondent is asked to choose between living in a given health state with certainty and a risky lottery involving a treatment with two possible outcomes: improving health state to the perfect health or dying. Finding the probability that results in indifference allows to elicit utility. In TTO, similarly, the respondent chooses between two options: i) living in a given state for some predefined time (e.g., 10 years) and then dying, ii) living in the perfect health for some time t and then dying. Finding t that results in indifference allows to assign the utility to the tested health state.

In applications, when eliciting societal preferences towards health state, the data are collected only for a subset of health states (and not 243 of them) and the utility values are extrapolated onto the whole set using some econometric modelling. The choice of the actual health states used in the study may possibly determine the resulting value set and so the decisions that are based thereon. There are several approaches for selecting the number and actual subset of health states to be employed. Dolan (1997) presented British study where each of 3,395 respondents evaluated 13 health states. Respondents had different sets of health states that enabled to collect information about 42 health states [3]. Dutch study provided information about 17 health states, that were evaluated by 298 respondents [11]. Japanese study also included 17 health states that were valued by 543 respondents [15], whereas Italian used 19 health states with a similar number of respondents [14]. In the literature, modelling of health-related utility assigned to particular health states is based mostly on random effects (RE) regression model [3, 11, 14, 15]. In some cases RE approach is compared with fixed effect approach [11, 15].

Being more precise, the aim of the present study is to verify the influence of the health state characteristics on the performance of a model extrapolating the utility from a survey onto the complete set of health states. So as to avoid dilemmas with specifying the econometric model, removing outliers, treating missing data, etc., we decided to use a more flexible, data mining approach. This aim is further split into the following. Firstly, we analysed how removing all the information about the valuation of a single health state influences the overall performance of the model. Secondly, we analysed the ability of the model to predict mean valuations of individual states to see

how well individual health states can be reconstructed using the other states (if poorly, we can conclude that a given state should be kept in the survey). We tried to see any patterns analysing the health state characteristics.

In section 2 we briefly describe the data used in the analysis, present the specific data mining approach used in the study and show how we present the influence of health state characteristic on the model performance. Section 3 presents the main results of our study. Finally, section 4 concludes our study and includes the most important results.

2. Data and methods

In the present study we used data from the only Polish study assigning societal utilities to EQ-5D-3L health states, conducted in 2008. All details and data characteristic were described by Golicki et al. [7]. We used the results of 7,351 TTO experiments performed on 321 respondents (each evaluating 23 health states defined using EQ-5D-3L taken from a wider set of 44 health states). Therefore, due to lack of the information about all possible health states (the study not being saturated), it seems essential to verify the performance of the model extrapolating the utility values to health states exluded from the study. As is usually done, our dependent variable was a loss of utility as compared to the perfect health state (due to given health state and was noted as $(1 - u)$, where u is health-related utility of a given state). Some health states may be valued as worse than death in TTO, and so have the utility lower than zero. Typically this utility is then transformed to guarantee $u > -1$. Thus, the dependent variable took values from interval $[0, 2]$ and the lower the value, the better the assessment of a given health state is.

As mentioned in the introduction, we decided to use data mining approach to modelling data in order to gain flexibility. Specifically, we used random forests—a method that uses modified bagging approach: a collection of trees is built with cross-validation, where each tree is built on different subset of observations and also on different subset of predictors. Each tree calculates prediction of dependant variable for each observation (based on available subset of predictors). After running specified number of trees, we have in the result multiple values of predictions for each observation. To provide one value of prediction, random forest averages the values of predictions (for regression problems) or finds their dominant (for classification problems). Data that are not used in building a given tree (due to using cross-validation) are

called OOB (out-of-bag) data and are used to calculate the prediction error [9]. An advantage of such a method is reduced variance thanks to defining result as average, as well as lack of assumptions on the preference structure (e.g., linearity of relationship, interdependencies between variables) and increased robustness of the results. Based on this, it is also possible to omit data cleaning process prior to modelling. Using random forest increases the accuracy of model, however it is harder to interpret the results and learning procedure [10].

The random forest provides two measures of variable importance: the mean decrease in accuracy and the mean decrease of node impurity. The former is calculated with OOB data: the overall prediction error is measured and then it is compared with the error calculated after omitting a given variable; the difference between these two values is averaged for all the trees and normalized using the standard deviation of the differences. The latter indicates node purity thanks to using a given variable: at each split on the given variable random forest calculates the decrease of residual sum of squares (RSS), i.e., the difference between RSS before and after split; aggregates over all splits for that variable; and then averages that value for all trees [12]. Measuring variable importance with such measures enables to provide conclusions similar to econometric approach and indicates variables having the most significant influence on the results. We employed this general approach in several ways. In the first step, we tried to identify dimensions (of a health state) having the greatest impact on the utility loss. It showed which health dimensions are the most important for Poles and which had the biggest influence on the perception of health-related utility. To do that, we built random forest for all five dimensions and tested variable importance. In this step we focused only on variable importance based on the model accuracy, because our goal is to extrapolate obtained results for health states that were not included in the Polish study, as well as to indicate health states that were the most problematic for assigning the loss of health-related utility.

After showing dimension that had the highest influence on the loss of utility, we checked how omitting one health state affects model performance. To do that, we measured at the beginning overall prediction error as out of sample MAE for 30% randomly chosen observations. Then we built random forest with exclusion of one health state, predicted values of the loss of utility, measured its error for the same out of sample data and recorded prediction error. That showed, how omitting information about given health state influences the model performance and enables to conclude if information about

one health state can contribute to an increased predictive power of the model (Fig. 1).

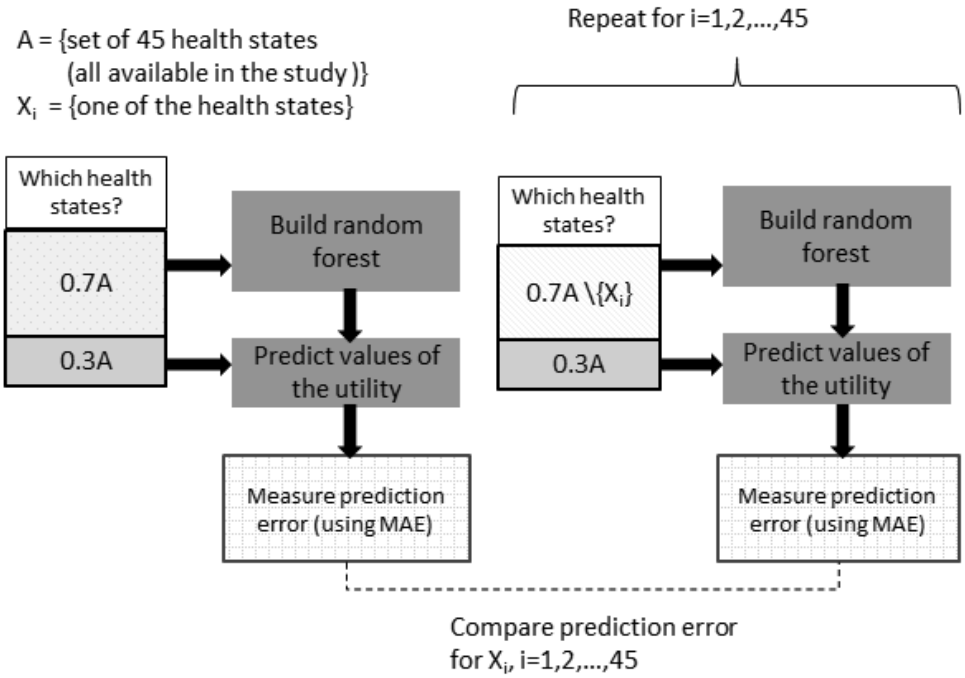


Figure 1: Algorithm 1 describing influence of health state on the overall model performance

In the present study, we also tried to identify health states that are most difficult to extrapolate the results of the model onto. In order to perform such a study, we excluded iteratively a single health state from the dataset, built random forest, used it to predict loss of the utility for omitted state, and then measured model performance (defined as out of sample MAE) (Fig. 2).

We also statistically verified (with t-Student test) if predicted values of the loss of utility (for a given health state, when this health state is not available in the modelling) were equal to empirical valuations (and so whether the extrapolation was correct when confronted with actual valuations). After identifying states onto which values of utilities were hard to predict we tried to identify characteristic of such health state. We checked distribution of the loss of utility assigned to the health states and its relation to prediction

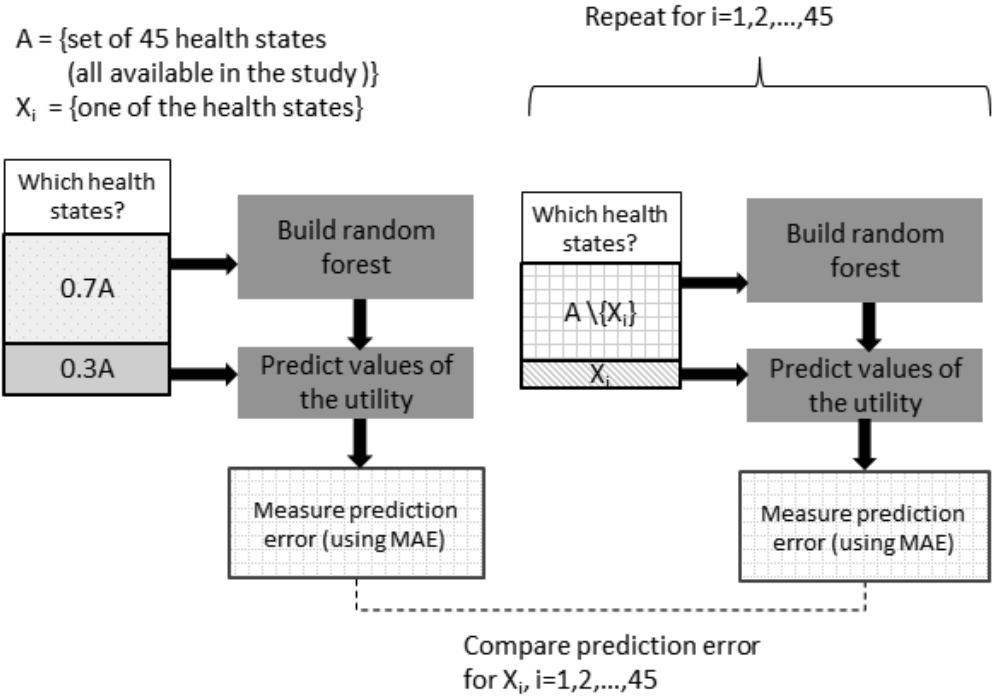


Figure 2: Algorithm 2 describing ability of the model to predict values of utility

error. Such approach showed elements which occurrence hindered model extrapolation to other health states.

Lastly, we did health state clustering according to prediction error. Here, we created several variables providing information about health state and indicated with random forest, which of them were the most relevant. We also provided simple linear model that informed about direction of that relationship.

3. Results

Analysis of health dimensions enabled to conclude that PD was the most important variable; omitting it caused the highest increase of prediction error (Fig. 3). The accuracy of models omitting dimensions AD, MO, SC, and

UA was almost identical; different seed settings changed order of importance of those dimensions suggesting that they had almost the same influence on the health-related utility.

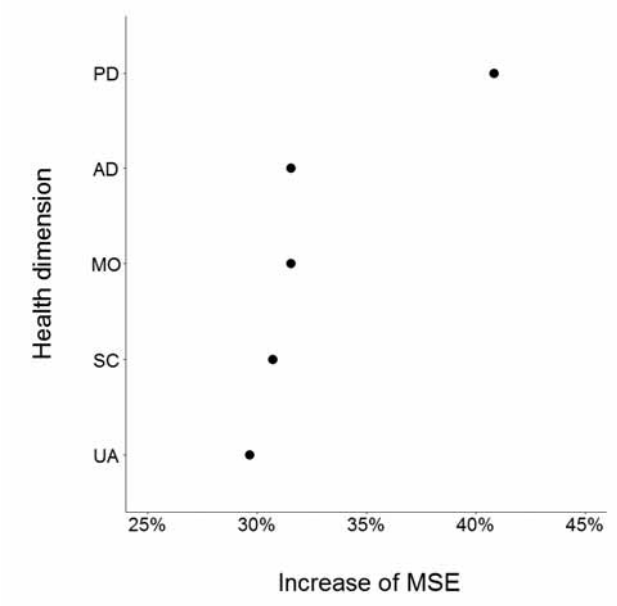


Figure 3: Importance of dimensions

The reduction of the information in the original data sample did not harm substantially the overall model performance. Excluding one health state from the train set did not cause the increase of prediction error (in the comparison to model built with the usage on all available health states). That means that information provided by one health state was not unique in terms of model building, because omitted information was replaced by other health states. Scatter plot showing out-of-sample MAE presents Figure 4. Solid line presents overall prediction error defined as out of sample MAE calculated for 30% randomly chosen observations.

The performance of random forest extrapolating results to the health states unmentioned in learning procedure differed among health states (Fig. 5). That means that for some health states it was easy for the model to predict the level of the utility loss and prediction error was even lower than

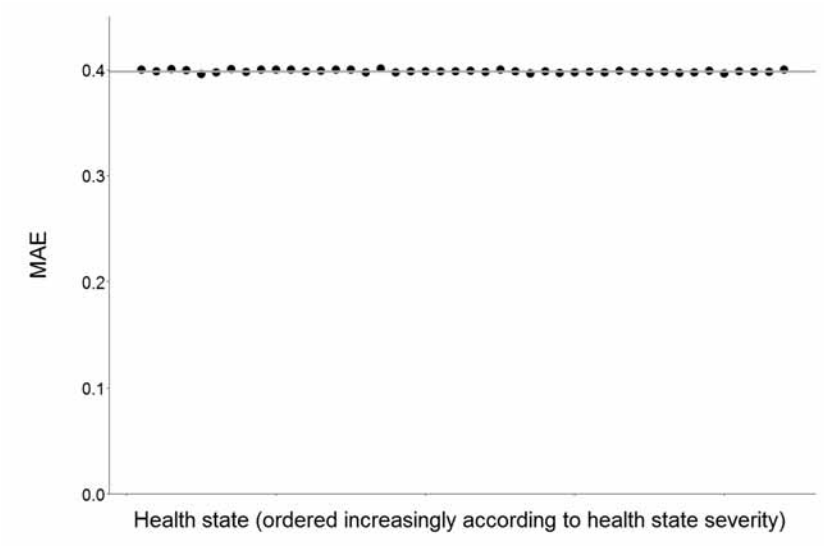


Figure 4: Prediction error due to exclusion of given health state from train set

overall prediction error. However, especially in terms of more severe health states, random forest found it difficult to extrapolate results of the model (for those health states the level of prediction error was the highest). That means, that there were health states that seemed to have unique characteristic and for them it was hard to extrapolate results of applied model. Moreover, the level of prediction error did not seem to correspond to the ability of model to predict mean value of utility loss for given health state. Using t-Student test concluded that values of the loss of health-related utility that were predicted by random forest were generally not statistically equal to empirical values for particular health state; such relationship was observed only in one third of cases (on Fig. 5 health states with equal means of empirical and predicted values are black, with unequal means – grey).

We showed that prediction error differed between the health states and did not seem to be affected by the relation between empirical and predicted mean of the loss of utility in health state. Also it seems that it was not caused by the absence of given health state in train set, but preference structure of health state, for which we predicted loss of utility. Therefore, we analysed

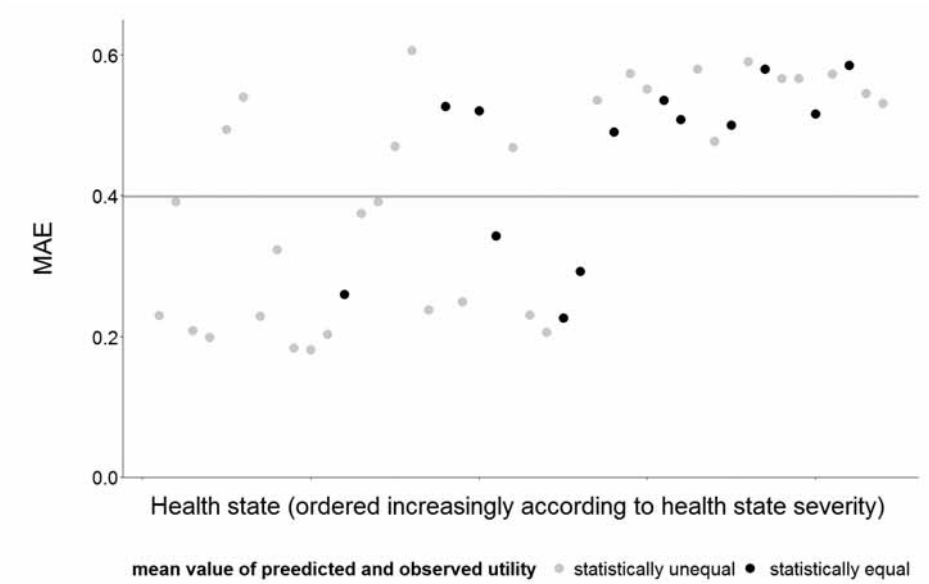


Figure 5: Performance of the random forest extrapolation

the relation between preference structure for given health state and accuracy of model predicting health-related utility, measured as out of sample MAE.

In general prediction error was highly correlated with mean value of the loss of utility in given health state, however for health states with the lowest utility in sample (and highest mean loss), such relationship was interrupted (Fig. 6 shows the relation between standard deviation, mean utility in health state and prediction error, where dashed line presents fitted linear regression and solid line shows fitted spline function). Additionally, prediction error measured as out of sample MAE was correlated with standard deviation of the loss of utility in health state, as a repercussion of relationship between mean utility in health state and MAE.

The relationship between model performance and distribution of health-related utility in given health state can be presented with the usage of boxplots (Fig. 7). Ordering health states on boxplot increasingly according to prediction error showed that the highest MAE had health states with the highest mean and standard deviation of the loss of utility.

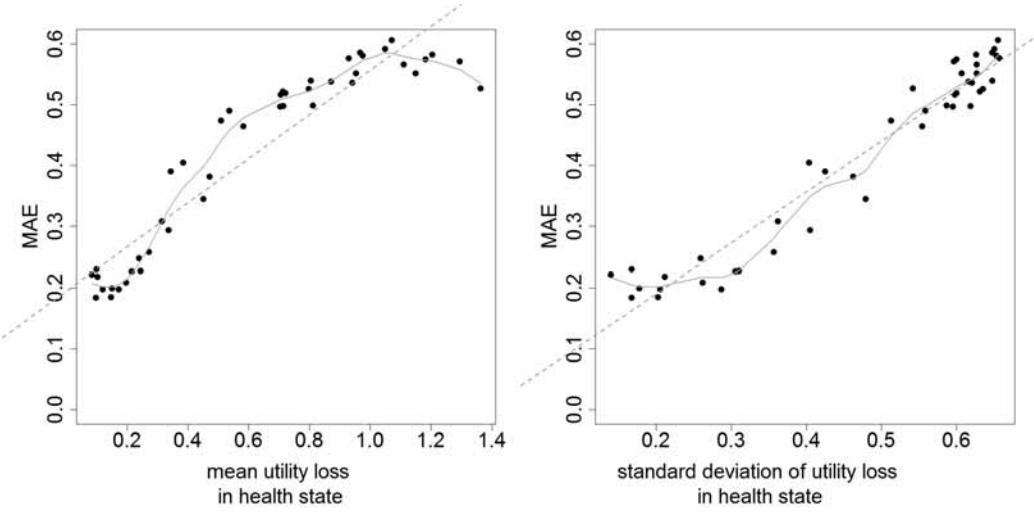


Figure 6: Relation between observed mean utility its standard deviation and prediction error

Presented results show that model's ability to extrapolate results on the out-of-sample health state depends highly on the health-state's characteristic. Health states with high mean values of utility loss, as well as high values of standard deviation were generally associated with worse results. For determining subset of health states that should be mentioned in the study, it is essential to describe the characteristic of health states, for which the loss of utility was the most difficult to predict. Therefore, in next step we defined several variables indicating features of health states and used them to build model explaining prediction error.

We showed that what influences the prediction error most is, i), the occurrence of more than once level three in health state (meaning severe problems in different dimensions), ii), mixing levels one (no problems) and three in a single state description, iii), mixing levels two (some problems) and three in a single state description (conclusions are based on decrease of model accuracy) (Fig. 8). This shows that health states including description of severe problems in some dimensions determined highly preference structure and differentiation of valuations between respondents.

Using random forest, did not show the direction of the relationship, i.e. if

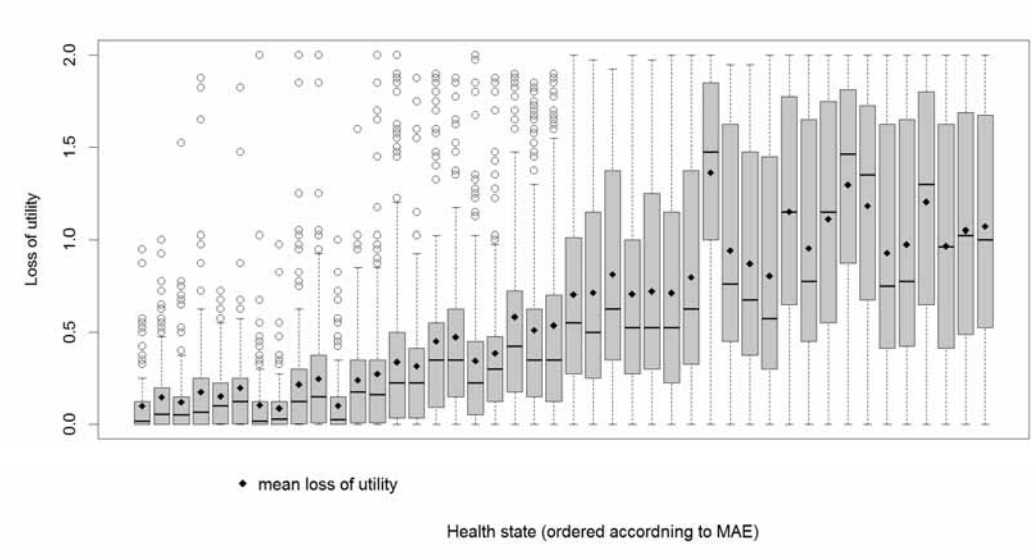


Figure 7: Distribution of the health state ordered by the error of prediction

given characteristic provided high or low values of prediction error. Information concerning direction of relationship could be provided by simple linear model, however using the same subset of predictors as in case of random forest was not allowed due to very high interdependencies between variables. Therefore, we built simple linear model, where we explained out of sample MAE by variables indicating number of levels occurrence in each health state. We also included interactions between variables (Tab. 1). Results show, that presence in health state level two or three increased chances of higher prediction error. The same relationship should be expected for simultaneous occurrence of levels 2 and 3. However, health status with levels 1 and 2 and with the lack of level 3 was characterized by lower prediction error.

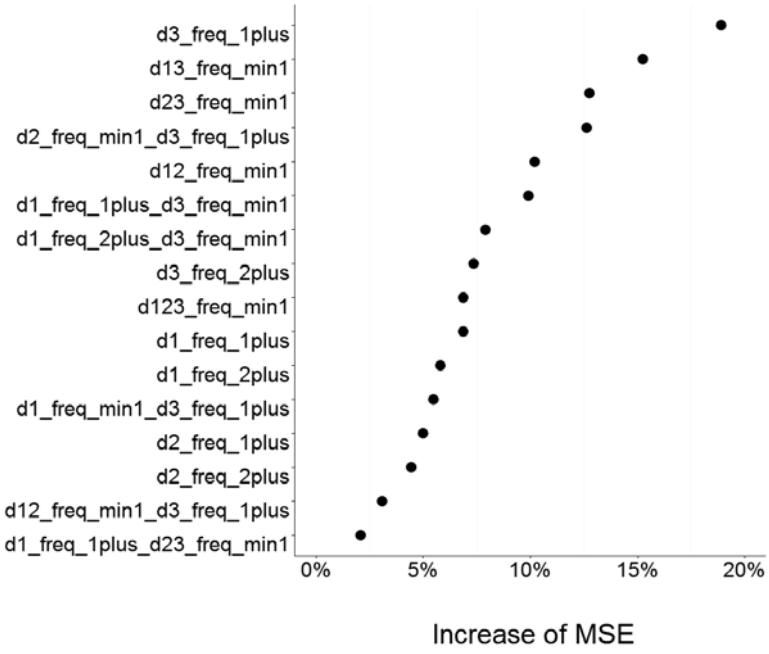


Figure 8: Influence of health state characteristic on prediction error

Variable	**Coefficient**	**Std. error**	**t value**	**p-value**	**statistical significance**
(Intercept)	0.257764	0.040074	6.432	1.64e-07	***
level2	0.011083	0.007948	1.394	0.171531	
level3	0.051587	0.009339	5.524	2.77e-06	***
level2:level1	-0.013847	0.006672	-2.075	0.044954	*
level2:level3	0.023664	0.005089	4.651	4.13e-05	***
level3:level1	0.025532	0.006896	3.703	0.000693	***
level2:level3:level1	0.006198	0.005931	1.045	0.302745	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 1: Level of MAE according to levels occurrence

4. Conclusion

Polish study concerning societal health-related utility included relatively big number of health states valued by a single respondent in comparison to other countries [7, 8]. This may be a reason why impact of omitting one health state from train data was not severe and similar for different health states. Moreover, health states present in Polish study differed with respect to how well the model was able to predict their average valuation based on valuations of other states. That leads to conclusion that by defining subset of health states present in valuation, we have to take into account model ability to extrapolate the results for particular health state. Also it is worth to mention, that different health states have different probability of occurrence in real life, so if we have health-state which will occur in rare cases, that difficulties with model extrapolation will not be very problematic.
Further study in this area should determine if as in random forest, valuations in responses cause problems with results extrapolation, or such relationship is present also in other methods that enable to extrapolate values of health-related utility (e.g. in case of modelling based on DCE [1]).

References

[1] De Bekker-Grob, E.W., Ryan, M., & Gerard, K. (2012). Discrete Choice Experiments in Health Economics: A Review of the Literature. *Health Economics*, 21(2), 145-172.

[2] Bleichrodt, H., Wakker, P., & Johannesson, M. (1997). Characterizing QALYs by Risk Neutrality. *Journal of Risk and Uncertainty*, 15(2), 107-114.

[3] Dolan, P. (1997). Modeling Valuations for EuroQol Health States. *Medical Care*, 35(11), 1095-1108.

[4] Dolan, P. (2000). The Measurement of Health-related Quality of Life for Use in Resource Allocation Decisions in Health Care. *Handbook of Health Economics*, 1, 1723-1760.

[5] EuroQol, (2015). *EQ-5D-3L User Guide. Basic Information on How to Use the EQ-5D-3L Instrument*. Retrieved on 19.06.2015, from http://www.euroqol.org/_leadmin/user_upload/Documenten/PDF/Folders_Flyers/UserGuide_EQ-5D-3L_UserGuide_2015.pdf.

[6] EuroQol, (2015). *EQ-5D-5L User Guide. Basic Information on How to Use the EQ-5D-5L Instrument.* Retrieved on 21.06.2015, from http://www.euroqol.org/_leadmin/user_upload/Documenten/PDF/Folders_Flyers/UserGuide_EQ-5D-5L_UserGuide_2015.pdf.

[7] Golicki, D., Jakubczyk, M., Niewada, M., Wrona, W., & Busschbach, J.J. (2010). Valuation of EQ-5D Health States in Poland: First TTO-Based Social Value Set in Central and Eastern Europe. *Value in Health*, 13(2), 289-297.

[8] Golicki, D., Jakubczyk, M., Niewada, M., Wrona, W., & Busschbach, J. (2013). Is Extending of a TTO Experiment to 23 States per Respondent Justifiable? An Empirical Answer from Polish EQ-5D Valuation Study. *Journal of Health Policy & Outcomes Research*, 1, 110-117.

[9] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* New York: Springer-Verlag.

[10] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning.* New York: Springer.

[11] Lamers, L.M., McDonnell, J., Stalmeier, P.F., Krabbe, P.F., & Busschbach, J.J. (2006). The Dutch Tariff: Results and Arguments for an Effective Design for National EQ-5D Valuation Studies. *Health Economics*, 15(10), 1121-1132.

[12] Liaw, A., & Wiener, M. (2015). *Breiman and Cutler's Random Forests for Classification and Regression – Package 'randomForest'.* Retrieved on 06.05.2015, from https://cran.r-project.org/web/packages/randomForest/randomForest.pdf.

[13] *Rozporządzenie Ministra Zdrowia z dnia 2 kwietnia 2012 r. w sprawie minimalnych wymagań, jakie muszą spełniać analizy uwzględnione we wnioskach o objęcie refundacją i ustalenie urzędowej ceny zbytu oraz o podwyższenie urzędowej ceny zbytu leku, środka spożywczego specjalnego przeznaczenia żywieniowego, wyrobu medycznego, które nie mają odpowiednika refundowanego w danym wskazaniu* (Dz.U. 2012 poz. 388).

[14] Scalone, L., Cortesi, P.A., Ciampichini, R., Belisari, A., D'Angiolella, L.S., Cesana, G., & Mantovani, L.G. (2013). Italian Population-based Values of EQ-5D Health States. *Value in Health*, 16(5), 814-822.

[15] Tsuchiya, A., Ikeda, S., Ikegami, N., Nishimura, S., Sakai, I., Fukuda, T., Hamashima, C., Hisashige, A., & Tamura, M. (2002). Estimating

an EQ-5D Population Value Set: The Case of Japan. *Health Economics*, 11(4), 341-353.

[16] Ustawa z dnia 12 maja 2011 r. o refundacji leków, środków spożywczych specjalnego, przeznaczenia żywieniowego oraz wyrobów medycznych (Dz.U. 2011 nr 122 poz. 696).

The Robust Clustering Methods

Ewa Szlachtowska, Dominik Mielczarek, Tomasz Burchard

AGH University of Science and Technology, Poland

Abstract

Classical statistical methods highly reliant on assumptions which are not always fulfilled in practice. In particular, it is frequently assumed that the data is normally distributed (at least approximately) or the sample is large enough to apply the central limit theorem to obtain a normal distribution of the error estimator. Unfortunately, very often, in practice, the assumptions are not fulfilled. The robust approach to statistical modelling and data analysis is to propose a statistical procedure giving credible estimates, which are useful not only in a situation where the data generated by the expected distribution, but also in a situation where the distribution data generating slightly deviates from the expected distribution. In the paper we present the robust clustering methods on the example of tclust algorithm.

Keywords: robust cluster analysis, trimming, heterogeneous clusters, tclust algorithm, k-means algorithm

1. Introduction

In the grouping task there is given an input data set described by specific attributes. Grouping consists in dividing the set into smaller groups, wherein elements included in the same group should be very similar to each

other, whereas the elements included into different groups - very different from each other. The result of grouping can be valuable, because groups are a generalization of the input information. For example, several million customers of mobile network, number, such that no man can cope with it, can be divided into several groups where you can find a characteristic customer. Such generalization can comprehend the available information and look at it from a different perspective.

Cluster analysis (CA) is the task of grouping a set of objects. CA is unsupervised learning method. This is a method of grouping elements in a relatively homogeneous class (groups, clusters). In the majority clustering algorithms are based on the similarity between the elements - expressed in terms of function, so-called measure of similarity and dissimilarity.

Below, we present examples of application of cluster analysis.

- Preliminary data analysis involving the separation of homogeneous groups (subpopulations), which are subject to separate further statistical and econometric analysis.

- Exploratory mining, where grouping is used among other things to customer allocation for certain subgroups.

- Information Retrieval with the task of streamlining and simplifying access to information (e.g. Google).

Classical statistical methods are highly reliant on assumptions that are not always fulfilled in practice. In particular, it is often assumed that the data are normally distributed (at least approximately), or the sample is sufficiently large to use the central limit theorem to obtain a normal distribution error estimator. Unfortunately, in practice, assumptions very often are not fulfilled. For example, the sample contains a few observations significantly deviating from the main part of data, which by assumptions are generated by the normal distribution. In addition, the normality of the distribution is always only an approximation - the real data are always quantized (discrete) and bounded. Naturally, this approximation is often so good that it still profitable to use it. Sometimes, however, the data are not normally distributed and standard statistical methods can produce unreliable results.

A robust statistical modelling approach proposes statistical procedures which give reliable estimates and which are useful not only in situations

where data generated by the assumed distribution, but when data generating distribution slightly deviates from the expected distribution. The procedure should have good properties, both when the sample does not have elements deviating from the main cloud of data (i.e. Outliers), but also when such elements are present. In addition, outliers can be detected by clusters analysis. Analysis is difficult, and in extreme cases impossible, when there are outliers in the dataset. In particular, methods and coefficients based on the assumption of normal distribution and linear relationships are not very robust to outliers, such as Pearson correlation, linear regression, etc. It is therefore necessary or removing outliers, or using robust statistical methods.

Robust cluster analysis (robust clustering) is part of the multivariate statistical analysis, including robust methods on slight departure from model assumptions (especially the presence of outliers) or on abandonment of certain assumptions.

2. Introduction to clustering algorithms

In cluster analysis commonly used algorithms are k-means algorithm and trimmed k-means algorithm. In the group of k-means methods clustering is pre-dividing the population into fixed predetermined number of classes. If the parameter $\alpha = 0$, we obtain k-means algorithm. While for $\alpha > 0$ we obtain trimmed k-means method, in which we reject the proportion α of trimmed observations.

k-means	**Trimmed k-means**
1. Draw k centers at random.	1. Draw k centers at random.
2. Keep observations to these k centers.	2. Keep the proportion $1 - \alpha$ of observations "closest" to these k centers.
3. Compute new k centers based on the observations.	3. Compute new k centers based on these "closest" observations.
4. Return the k centers leading to the "best" value of the target function.	4. Return the k centers leading to the "best" value of the target function.

Table 1: Comparison between k-means and trimmed k-means algorithms

2.1. Remarks on k-means algorithms

The main drawback of using this group of methods (k-means and trimmed k-means) is that the resulting clusters have a spherical shape, i.e. we are looking for spherically scattered groups. Moreover, both clustering methods ideally search for clusters with equal sizes. However, in many clustering problems, the clusters we are looking for are not necessarily spherical nor they have similar sizes. Moreover, k-means methods do not detect overlapping groups, i.e. large groups sometimes have a tendency to "absorb" small ones.

Consider a bivariate mixture of three components with very different scatters and with one small group. In this example, we examine the influence of different k by applying the k-means algorithm to the given data set.

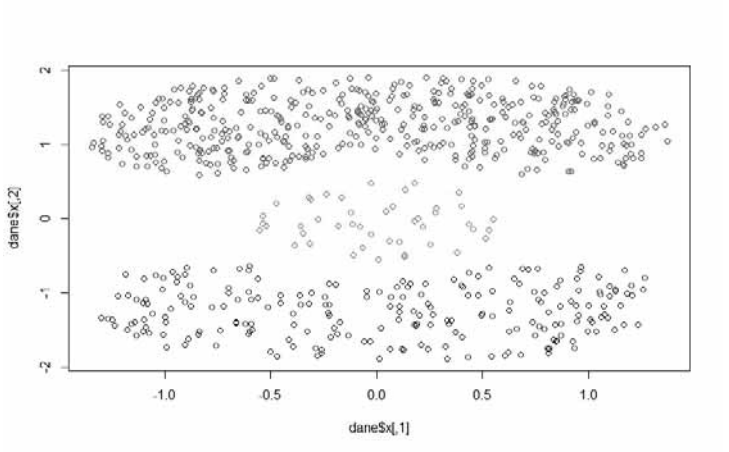
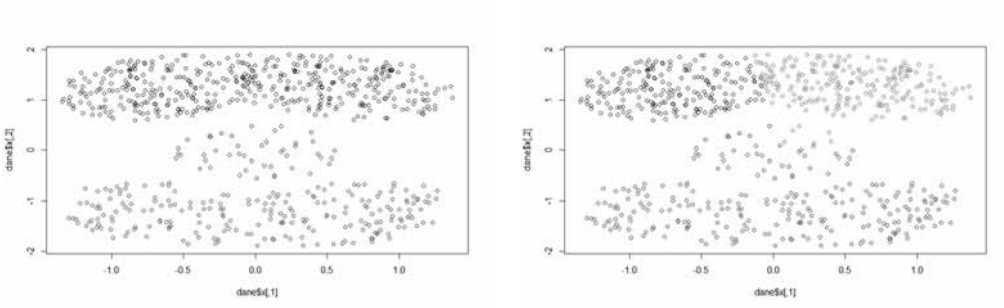


Figure 1: A scatter plot of the given data set. Different colors are used for the data points generated by each of the three bivariate components

In this example the drawback of k-means algorithm revealed. The algorithm ignores small (but obvious) data groups. In the Fig. 2, the points of small group were absorbed by other groups.

3. The tclust algorithm

It is easy to find connections between Forgys k-means algorithm (see [2]) and the fast-MCD algorithm (see [9]). These two widely applied algorithms plays a very important role in Cluster Analysis and in Robust Statistics, respectively. The algorithm **tclust** combines the capabilities of



(a) $k = 2$ (b) $k = 3$

Figure 2: Results of the clustering processes for the given data set for different parameters k, where k is the number of clusters. Different colors represent each observations individual cluster assignment

k-means method with the possibility robust estimation of the covariance matrix, which gives a fast algorithm for the minimum covariance determinant estimator (called Fast-MCD algorithm). **Tclust** algorithm was proposed by Heinrich Fritz, Luis A. García-Escudero, Agustín Mayo-Iscar (see [3, 4]).

New centers and new scatter matrices are updated by considering the cluster sample means and cluster sample covariance matrices. New cluster assignments are obtained by gathering the "closest" observations to the new centers. Unfortunately, such a combination of both algorithms does not provide reasonable clustering results. This is due to the fact that the large groups sometimes have a tendency to "eat" smallest, when the result is spurious or distorted groups with few almost collinear observations.

The **tclust** method based on constraining the relative sizes of the eigenvalues of the scatter matrices defining the elliptically contoured groups. The imposition of restrictions is one of the most important steps in the algorithm. This is the computational bottle-neck of this algorithm because a complex optimization problem must be solved in each concentration step.

3.1. Constrained robust clustering

Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^p$ be a sample of observations and $\phi(\cdot, \mu, \Sigma)$ the p.d.f. of a p-variate normal distribution with mean μ and covariance matrix Σ. We consider the general robust clustering problem, i.e. we search for a partition $\{\mathcal{R}_0, \mathcal{R}_1, \ldots, \mathcal{R}_k\}$ of the indexes $\{1, \ldots, n\}$ with $\#\mathcal{R}_0 = \lceil n\alpha \rceil$, centers $m_1, \ldots, m_k$, symmetric positive semi-definite scatter matrices $S_1, \ldots, S_k$

and weights $p_1, \ldots, p_k$ with $p_j \in [0, 1]$ and satisfying $\sum_{j-1}^{k} p_j = 1$ maximizing:

$$\sum_{j=1}^{k} \sum_{i \in \mathcal{R}_j} \log \left(p_j \phi(x_i, m_j, S_j) \right) \tag{1}$$

In this clustering problem the number of centers is predetermined. Moreover we allow clusters with elliptical shape and we reject the observations with indices from the set $\mathcal{R}_0$, i.e. outliers.

3.2. The "pros and cons" of tclust algorithm

The maximization of (1) when $\alpha = 0$ leads to well established clustering procedures depending on the constraints imposed on the weights p_j and on the scatter matrices S_j. For instance:

- If, for $\alpha = 0$ we assume that

  $$p_1 = \cdots = p_k \qquad \text{and} \qquad S_1 = \cdots = S_k = \sigma^2 I$$

  with I being the identity matrix and $\sigma > 0$, then we have k-means.

- Assuming that $\alpha > 0$ we obtain trimmed k-means.

- Note that a proportion $\lceil n\alpha \rceil$ of observations in $\mathcal{R}_0$, are not taken into account when calculating the objective function (1). Therefore, it can be avoided harmful effects of outlying observations.

- It is also important to note that the solution of previous problem when $k = 1$ (i.e., only a partition of data points onto trimmed $x_i \in \mathcal{R}_0$ and not trimmed ones $x_i \in \mathcal{R}_1$ is given) lead to the MCD problem.

The direct maximization of (1) without any constraint on the scatter matrices is not a well defined problem just by considering one of the S_j with $\det(S_j) \to 0$. Then, the objective function is unbounded.

In order to make the maximization of (1) a well defined problem, García-Escudero (see [3, 4]) propose the maximization of (1) but with scatter matrices $S_1, \ldots, S_k$ satisfying the following eigenvalue-ratio constraint:

$$\frac{\max\limits_{j,l} \lambda_{j,l}}{\min\limits_{j,l} \lambda_{j,l}} \leq c \tag{2}$$

with $\lambda_{j,l}$ are the eigenvalues, $(l = 1, \ldots, p)$, of the scatter matrices S_j $(j = 1, \ldots, k)$ and $c \geq 1$ is a constant that controls the strength of the constraint (2). If the constant c is selected larger, then the constraint on scatter matrices is "looser", which allows a greater variety of clusters. On the other hand, for small values of the constant c (close to one), we obtain a more uniformly scattered clusters. For $c = 1$, we obtain a trimmed k-means algorithm with weights.

3.3. Example

For an example data set, we examine the influence of different constraints on the size, shape and volume of the clusters.

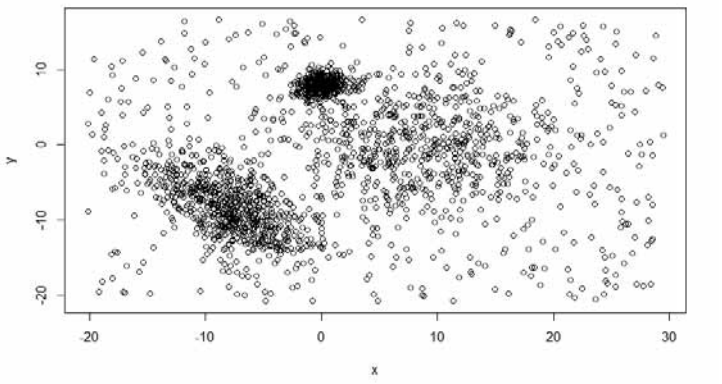
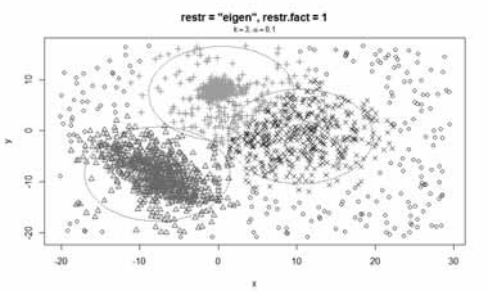


Figure 3: A scatter plot of the example dataset

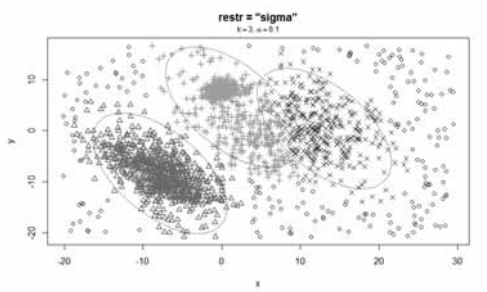
This data set, which accompanies the tclust package (see [4]), is a bivariate mixture of three simulated gaussian components with very different scatters. A 10% proportion of outliers is also added in the outer region of the bounding rectangle enclosing the three gaussian components. In Figure 3 there is a clear overlapping between two of these components.

Although different constraints are imposed, we are searching for $k = 3$ clusters and the trimming proportion is set to $\alpha = 0.1$ in all the cases.

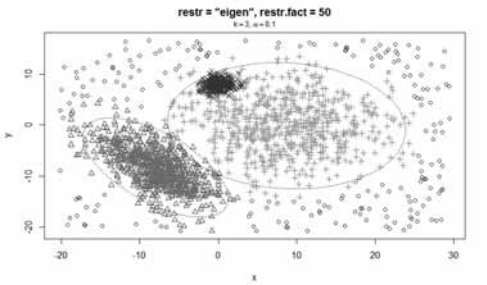
Assuming $c = 1$ and equal weights we obtain clusters using trimmed k-means. We received three clusters of similar structure and spherical shape. Note that for a sufficiently large fixed value c $(c = 50)$ the algorithm returns approximately three original clusters in spite of the very different cluster



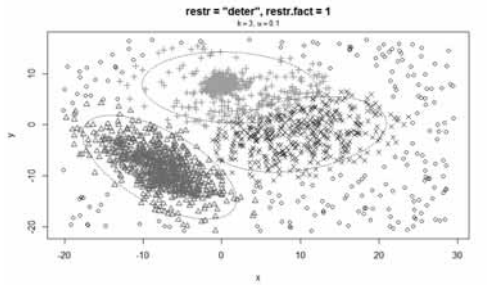
(a) $c = 1$, equal weights.



(b) The same cluster scatter matrix.



(c) $c = 50$, different weights.



(d) The same cluster scatter matrix determinant.

Figure 4: Results of the clustering processes for the given data set for different constraints on the cluster scatter matrices and the parameters $\alpha = 0.1$ and $k = 3$. Different colors represent each observations individual cluster assignment

scatters and different cluster sizes (see Fig. 4(c)). Moreover, this clustering procedure adequately handles the overlap of two clusters. The value $c = 50$ has been chosen in this case because the eigenvectors of the covariance matrices of the three gaussian components satisfy restriction (2) for this value.

4. Selecting the number of groups and the trimming size

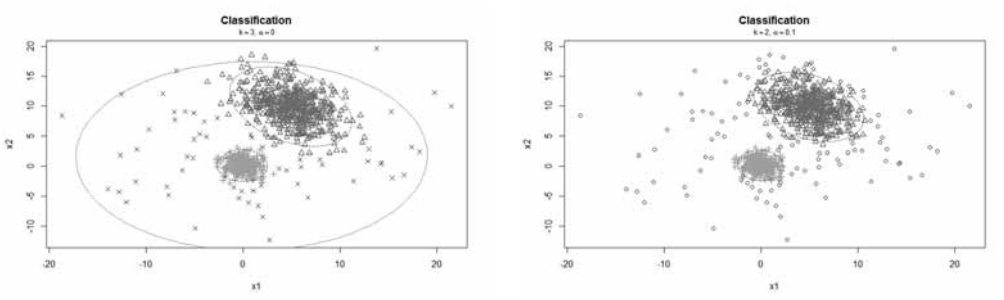
Probably one of the most complex problems when applying cluster analysis is the choice of the number of clusters, k. In some cases, we might have no problem specifying the number of clusters k, but usually k is completely unknown. In addition, in the presented approach, the level of trimming α also be selected without knowing the true contamination level. García-Escudero proposed some graphical tool that is useful in deciding on the number of groups k and the level of trimming α (see [5]).

As we will see in the following example, the choice of k and a choice of α are related problems, which should be solved simultaneously. It is important to note that the trimming level entails a specified number of clusters and vice versa. This dependence is due to the fact that the entire clusters can be completely trimmed when α excessively will increase. On the other hand, if the selected level of α is too low, then groups of outliers can create new false clusters. Thus, it seems that the number of clusters found in the data set is higher. In addition, simultaneous choice of k and α depends on the type of clusters, we're looking for as well as allowed differences between the sizes of the clusters.

4.1. The function ctlcurves in package tclust

Consider the data set, which could either be interpreted as a mixture of three components (see Fig. 5(a)) or a mixture of two components with a 10% outlier proportion (see Fig. 5(b)).

Let us assume first that the constant c have been fixed by the researcher, who applies the robust clustering methods. Then, even assuming that the $\alpha = 0$, choosing the appropriate number of clusters is not an easy task. The traditional method of selection of number of clusters, when $\alpha = 0$, is careful monitoring the maximum value of the objective function (1). However, increasing the number of clusters k always increases the maximum value of the function (1), which can lead to "overestimate" the number of clusters.



(a) $k = 3$, $\alpha = 0$. (b) $k = 2$, $\alpha = 0.1$.

Figure 5: Results of the clustering processes for the given data set for different parameters α and k

The function **ctlcurves** in package **tclust** sets the maximum value reached by the function (1) by successively applying the **tclust** function for a sequence of values of k and α. The default value c is set 50, but if desired, other values of c can be passed to **tclust** via **ctlcurves**.

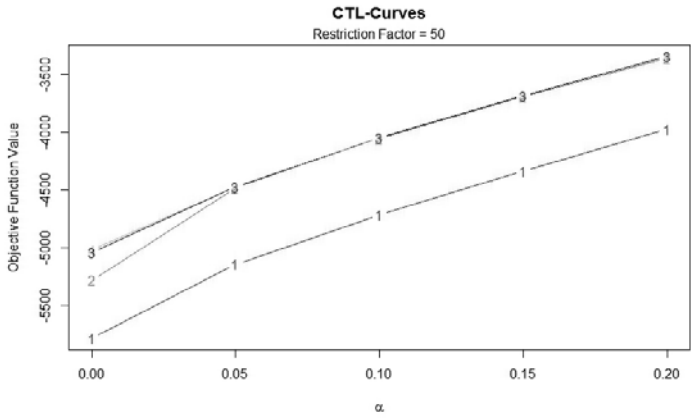


Figure 6: $k = 1, \ldots, 4$, $\alpha = 0, 0.05, \ldots, 0.2$, $c = 50$

Fig. 6 shows that increasing k from 2 to 3 is necessary assuming $\alpha = 0$, because the value of the objective function 1 varies greatly between $k = 2$ and $k = 3$. On the other hand, increasing k from 2 to 3 is not needed any more, when we assume the trimming levels equal to $\alpha = 0.05$. Then the third (more scattered) "cluster" vanishes when trimming 5% of the most outlying observations. So, there is no noticeable difference in the objective function

value at $\alpha \geq \alpha_0 = 0.05$ and $k \geq 2$. Increasing k from 3 to 4 is not needed in any case.

The described procedure requires an active role of the researcher in making sensible choices of parameters k and α. The type of restriction or the value of the eigenvalue-ratio constraint, which do not necessarily depend on the data set, must be predetermined. Consequently, the researcher's decision on the imposed restrictions, affects the proper determination of parameters k and α. For example, some specific clustering applications require almost spherical cluster that can be obtained by setting the constant c close to one.

Then, the obtained "sensible" value for k and α and the associated clusters should be carefully explored. For instance, **tclust** sends a warning when the returned clustering solution have been "artificially restricted" by the algorithm. This means, that the values $\max_{j,l} \lambda_{j,l}$ and$\min_{j,l} \lambda_{j,l}$ derived from the returned scatter matrices satisfy

$$\frac{\max\limits_{j,l} \lambda_{j,l}}{\min\limits_{j,l} \lambda_{j,l}} = c \tag{3}$$

because the algorithm has forced the chosen constraint. In this situation, if no specific constraints are required, the constant c can be gradually increased until the warning disappears.

4.2. The function DiscrFact in package tclust

Considering the **tclust** package, some additional graphical tools may be applied to assess the quality of clustering and trimming decisions. This is done by applying the function DiscrFact. An assessment is made by using the discriminant factors. These factors are used to assess the quality of the assignment decision of a non trimmed observation to the cluster as well as to assess the quality of the trimming decision. Hence, discriminant factors $\mathbf{DF}(i) \leq 0$ are obtained for every observation in the data set, whether trimmed or not. Observations with $DF(i)$ values close to zero indicate doubtful assignments or trimming decisions.

Silhouette plot(see [8]) can be used to summarize obtained discriminatory factors. Clusters in the silhouette plot with many large values of $\mathbf{DF}(i)$ indicate the existence of doubtful cluster assignment decisions. The most doubtful assignments with discriminant factor greater than a threshold value ($\mathbf{DF}(i) \geq \log(prog)$) are highlighted by the function **DiscrFact**.

We illustrate the result of applying the **DiscrFact** function to a clustering solution for the data set appearing in Fig. 5. Increasing k from 2 to 3 is only needed assuming $\alpha = 0$, because the value of the objective function (1) differs noticeably between $k = 2$ and $k = 3$ (see Fig. 6). Although Fig. 6 suggests to choose $k = 2$, when $\alpha = 0.1$, k has been increased to 3 in order to show how such a change leads to doubtful cluster assignment decisions, which will be visualized by the **DiscrFact** function.

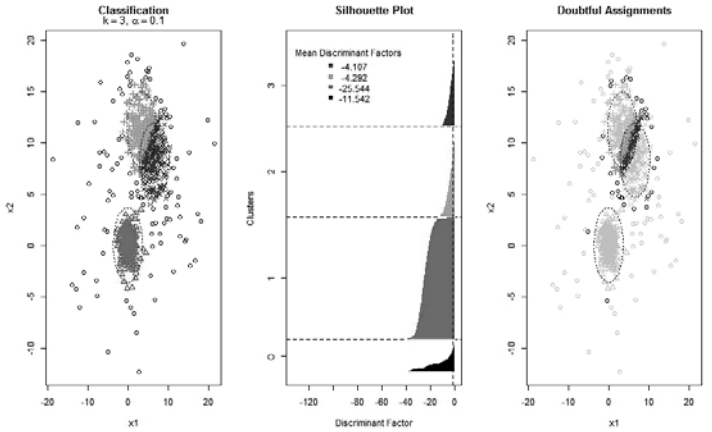


Figure 7: Graphical displays based on the $DF(i)$ values for a cluster solution with $k = 3$, $\alpha = 0.1$

Figure "Classification" simply illustrates the cluster assignments and trimming decisions. "Silhouette Plot" summarizes the obtained ordered discriminant factors, whereas the doubtful decisions are marked in "Doubtful Assignments".

All observations, such that $\mathbf{DF}(i) \geq \log(0.1)$, are plotted in black or in color. Most of the doubtful decisions are located in the overlapping area of the two artificially found clusters. Some doubtfully trimmed observations are located in the boundaries of these two clusters. When choosing $k = 2$ the doubtful cluster assignment decisions are much less, as illustrated in Fig. 8.

5. Conclusion

The **tclust** algorithm is an algorithm used to robust analysis of heterogeneous clusters. The algorithm imposes certain constraints on scatter

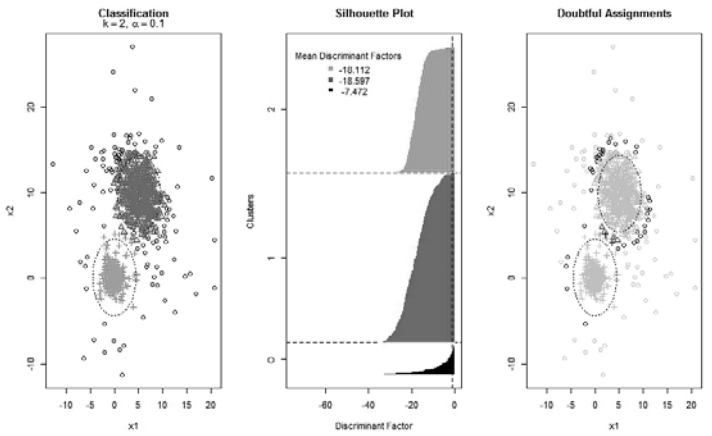


Figure 8: Graphical displays based on the $DF(i)$ values for a cluster solution with $k = 2$, $\alpha = 0.1$

matrices. The restrictions on the cluster scatters have to be changed in order to carry out different robust clustering algorithms. The **tclust** algorithm has successfully combined k-means algorithm and the fast-MCD algorithm.

References

[1] Ferraty, F., & Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. Springer Science & Business Media.

[2] Forgy, E.W. (1965). Cluster Analysis of Multivariate Data: Efficiency Versus Interpretability of Classifications. *Biometrics*, 21, 768-769

[3] Fritz, H., García-Escudero, L.A., & Mayo-Iscar, A. (2013). A Fast Algorithm for Robust Constrained Clustering. *Computational Statistics & Data Analysis*, 61, 124-136.

[4] Fritz, H., Garcıa-Escudero, L.A., & Mayo-Iscar, A. (2012). tclust: An R Package for a Trimming Approach to Cluster Analysis. *Journal of Statistical Software*, 47(12), 1-26.

[5] García-Escudero, L.A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2011). Exploring the Number of Groups in Robust Model-based Clustering. *Statistics and Computing*, 21(4), 585-599.

[6] Kosiorowski, D., Mielczarek, D., & Szlachtowska, E. (2015). Clustering of Functional Objects in Energy Load Prediction Issues. [in:] M. Papież & S. Śmiech (Eds.), *Proceedings from 9th Professor Aleksander Zeliaś*

International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Cracow: Foundation of the Cracow University of Economics.

[7] Górecki, T., & Krzyśko, M. (2012). Functional Principal Components Analysis. [in:] J. Pociecha & R. Decker (Eds.), *Data Analysis Methods and its Applications*. Warszawa: C.H. Beck.

[8] Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

[9] Rousseeuw, P.J., & Driessen, K.V. (1999). A Fast Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, 41(3), 212-223.

Simulation Comparison of Methods for Estimation the Forecasts Intervals for Time Series

Michał Miłek

University of Economics in Katowice, Poland

Abstract

In practice of economic research very often we meet the problem of formulating forecasts. Typically to predict the future values of the variables we base on their past realization. To determine the prognosis is usually utilized a variety of methods such as: nave method, models of Holt and Winters, ARIMA, GARCH as well as various simulation methods. Using these methods allows us to formulate point forecasts to determine the future value of the variable under consideration by extrapolation the model describing the studied phenomenon. Information obtained on the basis of such forecast is not always sufficient. Often, is required the knowledge about the probability of what the future value of a variable will be in a given prediction environment. This information can be very useful in many situations. The simulation study has been made to compare the efficiency of the interval forecasting methods.

Keywords: bootstrap, simulation, interval forecast

1. Introduction

By analyzing quantitative data we often meet a situation where it is necessary to predict the analyzed phenomena in the future. The problem of forecasting is widely discussed in the literature and there is numerous works based on a different approach to this issue. To forecast time series a variety of methods is used, for example: nave method, models of Holt and Winters, ARIMA, GARCH as well as various simulation methods. However, to apply those methods for a specific purpose relevant assumptions must be met. For example, when they are considered prediction intervals based on an assumed empirical distribution, it is required that the rest of the model have proper distribution and the forecast is unbiased:

$$E\left(e_n^2\left(t\right)\right) = var\left(e_n\left(t\right)\right) \tag{1}$$

2. Basic remarks

Stochastic process can be determined by symbols:

$$\{Y_t\}_{t=1,2,...} = (Y_1, Y_2, ...) \tag{2}$$

The purpose of analysis is to set the forecast for $k = 1, 2, \ldots, G$ periods ahead (forecast horizon will be denoted by G), the point forecast will be denoted as follows: $\hat{Y}_n, \hat{Y}_n, ..., \hat{Y}_n(G)$ whereas the values observed during the forecast period by $\hat{y}_n$, $\hat{y}_n$, ..., $\hat{y}_n(G)$. Forecast error can be written as follows: $e_n(t) = \hat{Y}_{n+t} - \hat{y}_n(t)$.

D.C. Montgomery et al. [12] consider two models of periodic time series: additive and multiplicative. In the following analysis an additive model will be considered. This model can be written as follows [12, 15, 1]:

$$Y_t = f(t) + w_t^+ + \varepsilon_t \tag{3}$$

where:
$t = 1, 2, \ldots n$
Y_t – observed value in period t
$f(t)$ – trend function
w_t^+ – seasonal components for an additive model $w_t^+ = w_{t+S}^+ = \ldots = w_{t+(k-1)S}^+$ which satisfy a given restriction $\sum_{t=1}^{G} w_t^+ = 0$
ε_t – random component with the properties: $E(\varepsilon_t) = 0$ i $D^2(\varepsilon_t) = \sigma_\varepsilon^2$
S – number of distinguished periods
In further analyzes the presence of a periodic component was omitted, while the trend was considered as a function of linear model. Then

$$Y_t = \alpha + \beta X_t + \varepsilon_t \tag{4}$$

$t = 1, 2, \ldots n$
α, β – unknown model parameters

To check the effectiveness of the interval forecasting methods various disturbances were introduced to the residual component.

3. Construction of interval forecasts

During the analysis of time series the point forecasts are often constructed to determine the future value of the variable under consideration by extrapolation model describing the studied phenomenon. This method does not involve the full information about the future value of a variable - the likelihood that the future realization will be in a given forecast environment is not known. However, with only one value we omit a lot of valuable information that can significantly affect the decisions made on the basis of such a forecast. Therefore, it may be warranted to construct interval forecast, which detail the accuracy of the prediction. This allows us to determine the uncertainty of the forecast, consider various strategies and compare forecasts obtained on the basis of several methods [3]. Interval forecast consists of a lower and upper limit, which future value will be projected with a certain probability. In literature, this interval is called differently: the scope of the predictions [2], the confidence interval [7] and the interval limits of the forecast are called forecast limits [14]. However, most often it appears the words: interval forecasts. There are many methods of determining intervals of forecasts, but there is no one that could be applied in all cases. In particular, the theoretical intervals cannot always be applicable because of the complexity of the models, especially multi-equation ones.

3.1. Construction of interval forecasts based on the theoretical distribution

Prediction interval is defined as follows:

$$P\left\{y_T^* - uv_T \leq y_T \leq y_T^* + uv_T\right\} = 1 - \alpha \tag{5}$$

where:
α – forecast reliability,
y_T^* – point forecasts value in the period T.

$$v_T = s \cdot \sqrt{\frac{(T-\bar{t})^2}{\sum_{t=1}^{n}(t-\bar{t})^2} + \frac{1}{n} + 1}$$

$$s = \sqrt{\frac{1}{n-(m+1)}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}$$

y_t – the true value of the Y variable at the moment or period t,

$\hat{y}_t$ – the theoretical value of the Y variable at the moment or period t,
n – number of observations,
m – the number of explanatory variables,
u – factor associated with the reliability of the forecast, forecasted variable distribution and the length of the time series ($u > 0$). If in the process of verifying the hypothesis of normal distribution of model residuals is not rejected, then the value of the coefficient can be read from normal distribution tables (for $n > 30$) or t-Student's distribution tables with n-2 degrees of freedom and probability 1- α. If this hypothesis has been rejected or has not been verified, then the value of the coefficient can be determined from the Chebyshev inequality:

$$P\left\{|Y - E(Y)| \leq u\sigma\right\} \geq 1 - \frac{1}{u^2} \tag{6}$$

$$u = \sqrt{\frac{1}{1-\alpha}} \tag{7}$$

where: $E(Y)$ – the expected value of forecasted variable Y,
σ – the standard deviation of forecasted variable Y.

3.2. The construction of interval forecasts based on bootstrap method

One of the methods of construction prediction intervals is the bootstrap method. This method is widely used in economical time series analysis (see: [8, 9]). In this approach, the forecast interval is determined based on samples from the empirical distribution of bootstrapping statistics [13] (see:[4, 6]). If the test statistic is unbiased as an estimator of the unknown parameter, the sampling distribution centers around the real parameter value. Cutting off the appropriate quantile of distribution statistics, we can obtain the appropriate intervals of the forecast. This method is applicable if the bias of bootstrap estimator is small.

Figure 1 presents the principle of the generating the bootstrap samples.

Considered is the form of model consistent with the formula. Analyzed time series Y_t free of the trend and seasonal fluctuations is reduced to the random component of the form:

$$X_t = \varepsilon_t \tag{8}$$

where: ($t = 1, 2, \ldots, n$)

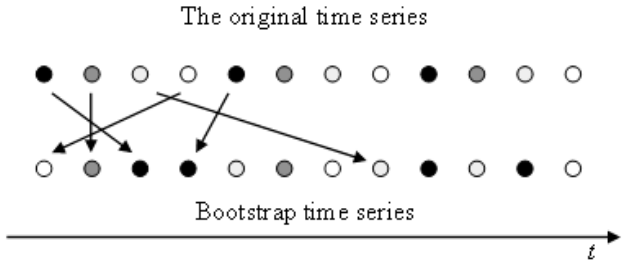


Figure 1: The idea of using the bootstrap method

In obtained X_t time series the subsequent realizations $x_1, x_2, \ldots, x_n$ can be denoted as follows:

$$X_t = x_1, x_2, \ldots, x_n \tag{9}$$

Let us introduce to bootstrap sample designation:

$$X_t^* = x_1^*, x_2^*, \ldots, x_n^* \tag{10}$$

Using such a sample we can determine the bootstrap series, where the trend is determined, in accordance with the received grade obtained on the basis of calculations of the output time series. The random deviations are added based on bootstrap sampling. So the resulting time series can be written as follow:

$$Y_t^* = f\,(t) + X_t^* \tag{11}$$

Retrieving in this way B-times the bootstrap samples we obtain B time series:

$$Y_t^{*(1)}, Y_t^{*(2)}, \ldots, Y_t^{*(B)} \tag{12}$$

For each of the series the forecast is determined. Based on B-forecasts for a fixed period $t+k$ quantile $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ is calculated. These are respectively the lower and upper bound of the prediction interval designated by the bootstrap method.

3.3. Block Bootstrap

B. Efron and R. Tibshirani [5] show the possibilities of using the bootstrap method for time series. This block bootstrap method (BB) is used to estimate the parameters of the autoregressive model. The idea of BB presented by B. Efron and R. Tibshirani [5] is presented schematically in Figure 2.

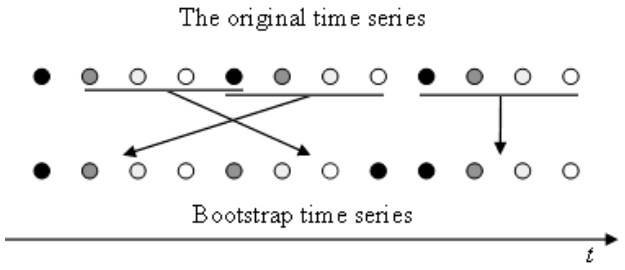


Figure 2: The idea of using the block bootstrap method

This method leads for the reverse sampling of full blocks of length G observation and insert them together into a time series. In the illustration shown in Figure 2 taken $G = 4$, $(G = kn)$.

In the obtained series next i-th element of G-block ($i = 1, 2, ..., k$) can be written by the following symbol:

$$x_i = x_{i+1}, x_{i+2}, \ldots, x_G \tag{13}$$

Let us introduce to block bootstrap sample designation:

$$(x_1^*, x_2^*, \ldots, x_k^*) \tag{14}$$

where $x_i^* = x_{i+1}, x_{i+2}, \ldots, x_G$, and $i \in \{1, 2, \ldots, k\}$
Like in the bootstrap method, a series of Y is constructed:

$$Y_t^* = f(t) + X_t^* \tag{15}$$

And then B time series is obtained as result of replication:

$$Y_t^{*(1)}, Y_t^{*(2)}, \ldots, Y_t^{*(B)} \tag{16}$$

The interval forecast is determined analogous to that in the case of bootstrap method. The advantage of block bootstrap method is that it is less dependent on the model of the time series than the classic bootstrap method. It can also be applied to the analysis of time series with autocorrelation. As one of the variants of the method there is a version that is assumed to levy the blocks on each other. In particular, this approach may be used for periodical series analysis (see: [10, 11]). In the following analyses it will be used to determine the interval forecasts.

4. Simulation

In order to compare discussed methods of the interval forecasting, we made a simulation which was carried out based on the following assumptions:

1. we perform 1,000 simulations,
2. considering is the time series with the length of 110 observation ($n+G=100$) generated according to the model,
3. a variety of disorders are introduced to the model affecting on the form of residuals,
4. forecast interval is determined at $G=10$ periods ahead based on $n=100$ first observations, by the classical method, bootstrap and block bootstrap,
5. we compared the results obtained with the considered methods.

The following indicators were used in order to compare the quality of predictions:

1. Q – Coverage range forecasts of the observed value of a time series over the forecast horizon,
2. W – The width of the forecast interval,
3. P – synthetic index (the ratio of the above indicators).

The following disorders were tested in a series of residues: no disturbance, no normal distribution, autocorrelation, the occurrence of an ARCH effect. The following chart shows a comparison of the results. Figure 3. Shows the comparison of the coverage by interval forecasts the observed value of a time series over the forecast horizon - Q. Figure 4 shows the comparison of the intervals wide - W, while Figure 5 shows the synthetic indicator P.

Analyzing the results, it found out that all variants of the bootstrap method behave very similarly. When we analyze the variant without the disorder, it turns out that the bootstrapping methods have a slight advantage over the classical method for interval width, which directly affects the value of the synthetic indicator. In the case of a first disturbances (residues are not normally distributed) and the emergence of the ARCH effect, methods give different results. Classical method preserves the width of forecasts while the bootstrap interval width increased significantly. This translated directly also in the coverage of the interval forecast of real value and the value of the synthetic indicator.

Q - indicator value

Type of disturbance	Forecast method			
	T	B	BB1	BB2
No disturbances	0.942	0.942	0.943	0.940
No normal distribution	0.205	0.954	0.946	0.949
Autocorrelation of residues	0.294	0.281	0.280	0.282
The occurrence of ARCH	0.138	0.936	0.943	0.944

W - indicator value

Type of disturbance	Forecast method			
	T	B	BB1	BB2
No disturbances	3.946	3.887	3.894	3.898
No normal distribution	3.946	28.619	28.573	28.584
Autocorrelation of residues	4.060	3.811	3.828	3.836
The occurrence of ARCH	3.946	50.849	51.108	51.006

P - indicator value

Type of disturbance	Forecast method			
	T	B	BB1	BB2
No disturbances	0.239	0.242	0.242	0.241
No normal distribution	0.052	0.033	0.033	0.033
Autocorrelation of residues	0.072	0.074	0.073	0.073
The occurrence of ARCH	0.035	0.018	0.018	0.019

Table 1: Results comparison

5. Conclusion

In this paper four methods of determining the interval forecast were compared - using computer simulation method were compared based on the formed classical probability distribution and three bootstrap method: percentile bootstrap and two variants of block bootstrap. By using three indicators the quality of forecast were tested. The use of bootstrap as an example of simulation method allow us to escape from the assumptions that are involved in the classic method. Failure to meet these assumptions will disqualify the possibility of using the classical method.

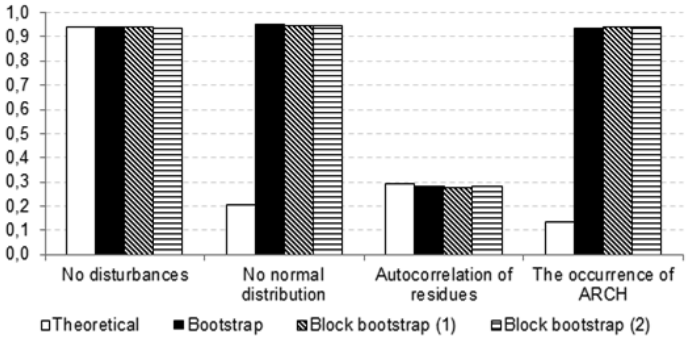


Figure 3: Q indicator value - interval forecasts covering the observed value of a time series over the forecast horizon

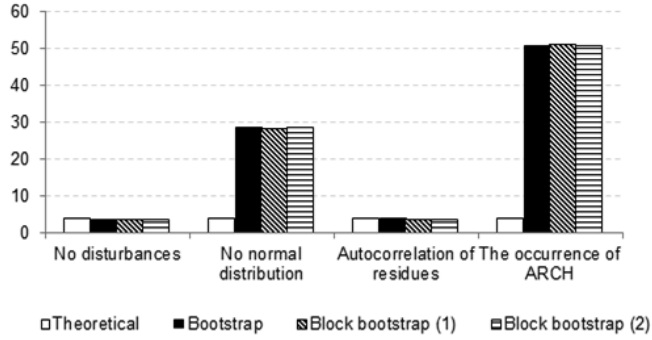


Figure 4: Q indicator value - interval forecasts covering the observed value of a time series over the forecast horizon

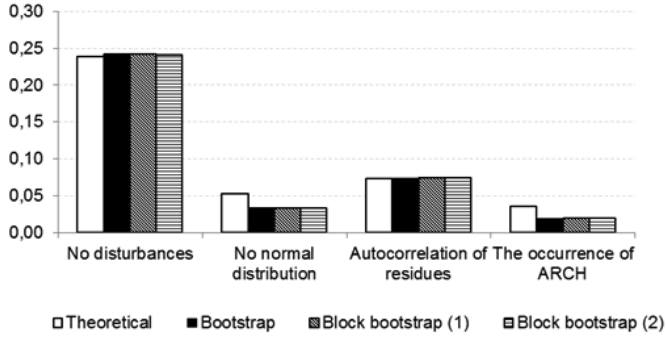


Figure 5: Q indicator value - interval forecasts covering the observed value of a time series over the forecast horizon

References

[1] Box, G.E., Jenkins, G.M., & Herer, W. (1983). *Analiza szeregów czasowych: prognozowanie i sterowanie.* Warszawa: PWN.
[2] Brockwell, P.J., & Davis, R.A. (2013). *Time Series: Theory and Methods.* Springer Science & Business Media.
[3] Chatfield, C. (1993). Calculating Interval Forecasts. *Journal of Business & Economic Statistics*, 11(2), 121-135.
[4] Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 1-26.
[5] Efron, B., & Tibshirani, R.J. (1994). *An Introduction to the Bootstrap.* CRC Press.
[6] Good, P.I. (2006). *Permutation, Parametric, and Bootstrap Tests of Hypotheses.* Springer Science & Business Media.
[7] Granger, C.W.J., & Newbold, P. (2014). *Forecasting Economic Time Series.* Academic Press.
[8] Irding, M., & Lyden, J. (2008). Is Historical Data a Good Estimate of the Future Risk of Funds? – A Study on the Swedish Hedge Fund Market. *Rapport nr.: Industriell och finansiell ekonomi 07/08:4.*
[9] Juan, S., & Lantz, F. (2001). Application of Bootstrap Techniques in Econometrics: The Example of Cost Estimation in the Automotive Industry. *Oil & Gas Science and Technology*, 56, 373-388.
[10] Kończak, G., & Miłek, M. (2014). Wykorzystanie metody moving block bootstrap w prognozowaniu szeregów czasowych z wahaniami okresowymi. *Studia Ekonomiczne*, 203, 91-100.
[11] Miłek, M. (2013). Wyznaczanie prognoz przedziałowych z wykorzystaniem metody Moving Block Bootstrap. *Rola informatyki w naukach ekonomicznych i społecznych. Innowacje i implikacje interdyscyplinarne*, 2, 193-205.
[12] Montgomery, D.C., Jennings, C.L., & Kulahci, M. (2015). *Introduction to Time Series Analysis and Forecasting*, John Wiley & Sons.
[13] Moore, D.S., & McCabc, G.P. (2006). Bootstrap Methods and Permutation Tests. [in:] D.S. Moore & G.P. McCabc (Eds.), *Introduction to the Practice of Statistics.* New York: W.H. Freeman.
[14] Wei, W. (1990). *Time Series Analysis.* Redwood City.
[15] Zeliaś, A., Pawelek, B., & Wanat, S. (2002). *Metody statystyczne: zadania i sprawdziany.* Warszawa: PWE.

Exploration of Internet User Activity with the {sm} Package

Agnieszka Gołdyn, Agnieszka Góra, Robert Olechnowicz

Cracow University of Economics, Poland

Abstract

Exploring the behavior of Internet users has become a significant part of modern marketing campaigns. When companies possess the knowledge of spatio-temporal characteristics of online user activity, such as the number of unique visitors to a webpage at a specific time or the number of views of the advertising banners, they can more effectively target potential customers and optimize the layout of their websites. In this paper, we demonstrate the capabilities of the sm package from the R statistics environment by performing non-parametric kernel estimation of the data on internet user activity. We obtain and visualize uni- and bivariate estimates of the density functions. The resulting distributions are shown to be fat-tailed and bimodal, which suggests variable and time-dependent patterns of user activity on the analyzed websites.

Keywords: kernel density estimation, online user activity, {sm} package

1. Introduction

In the recent years, data mining and data-driven marketing strategies have become an integral part of customer targeting and marketing strategy planning. Online advertisements have also become widespread, and maintaining them is now a major focus of numerous companies. If placed correctly, online advertisements are also highly impactful. Research has shown that Ads Click-Through Rate (CTR), which is a widely accepted measure of their performance, can be improved by as much as 670% if the correct user group is targeted [7].

However, in order to extract meaningful information and patterns from the large quantities of data from the Internet, one has to apply proper data processing and visualization techniques. One often has no prior knowledge of the patterns, and therefore the suggested approach is to obtain descriptive statistics and then proceed to further analysis [4]. When the data is highly irregular and parametric methods are likely to provide low quality approximations, the shift to non-parametric methods is justified. One of the statistical packages offering a wide variety of functionalities related to non-parametric statistics is the sm package, a part of the R environment [1].

In this article, we will demonstrate the capabilities of the {sm} package with regards to a specific non-parametric method, the kernel density estimation. Our data, consisting of daily observations of Internet user activity from two websites (the number of unique users and the number of views of online advertisements), will be subject to a three-step analysis: the interpretation of descriptive statistics, univariate density estimation and joint density estimation. All steps will be shortly discussed and illustrated with visualizations prepared with the default tools from the package.

The next section of this paper shortly discusses the widely accepted measures of user activity and the methods of collecting them. Section three is dedicated to the theoretical basis of kernel density estimation. It introduces the Rosenblatt-Parzen estimator and its properties for uni- and bivariate densities. Section four introduces the {sm} package and its functionalities, maintaining focus on its kernel estimation algorithms. Section five contains the analysis of sample data from the Internet. Section six concludes the results of our research and provides some recommendations. The last section contains the list of references.

2. User activity in the Internet

Due to technical considerations, there is no single method of measuring user activity on the Internet and the methods currently in use can only provide approximations of the real patterns of user behavior (for a short review of different methods, see [7]).

One of the measures is the number of unique users, although there is also no general agreement regarding the technical approach to obtaining such data. One method is based on recording the individual IP addresses of all

visitors to a website and assigning all actions on the website to the respective users. It prevents a single person from logging in multiple times and being recorder as different people. However, if the user has a dynamic IP address which is reset between visits to the website, redundant records might still happen. An alternative approach is to use the cookie files sent to the user's browser, for they also clearly identify the user and, unlike the IP address, are less likely to be deleted between visits to the website.

The second measure of activity is the number of page views, which reflects how many times the unique users have clicked online advertisement banners on a website. It is a reliable measure of activity, provided that the unique users have been properly identified in the previous step.

When sufficient data on user activity is gathered, it is possible to extract some valuable information or map the most frequently clicked parts of the website. They can then be visualized with the use of heatmaps, where brighter colors correspond to larger numbers of clicks on a specific spot.

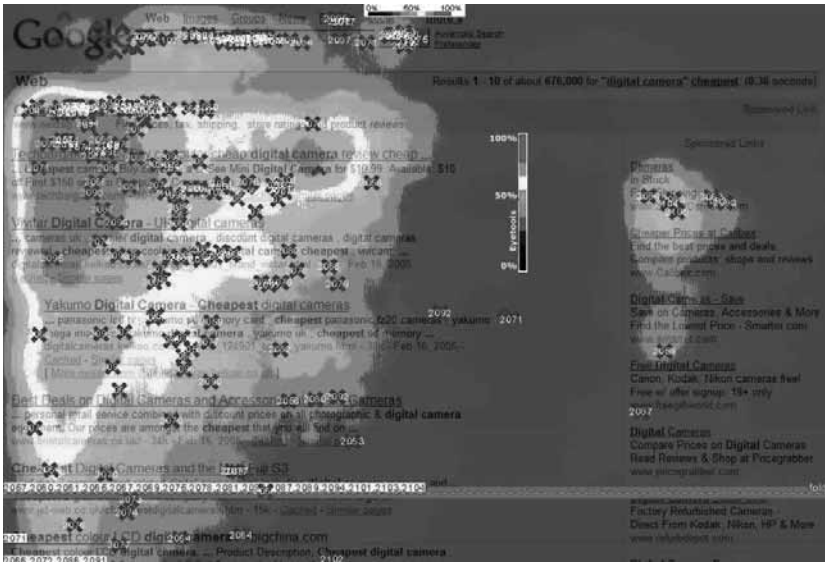


Figure 1: An exemplary heatmap [3]

Such information can be used for redesigning the webpage layout in a way which attracts the most users to some of its components, such as advertisement banners.

3. Kernel density estimation

The approach preceding the kernel estimation was the histogram approach, praised for its simplicity and practicality, but also widely criticized for its shortcomings [2, 3]. It is based on a series of bins, whose width, as well as the central points of the interval, have to be selected manually. In spite of some rules-of-thumb which provide accurate approximations and allow the researchers to create histograms automatically, there are still some features histograms lack as an approximation of the underlying density function. One of the missing properties is the smoothness of the function, the lack of which makes histograms non-differentiable due to the sharp edges of the boxes they consist of.

3.1. Univariate density estimation

The kernel estimation was introduced as a continuous, non-parametric alternative to histograms, capable of delivering the missing features. A smooth, symmetric weight function called a kernel was used in order to allow for differentiation, and it was centered over individual observations in order to replace averaging over the intervals. The estimator which was proposed as an alternative to the histogram is known as the Rosenblatt-Parzen estimator. It has received much attention as an unconditional estimator of probability density functions. It can be written in the form:

$$\widehat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} w(y - y_i; h) \tag{1}$$

where $w(.)$ is a probability density called a kernel function, h is the variance of this density function. There is a general agreement that the kernel function should be symmetric with mean 0, but its precise shape does not seem to be a very important issue. A normal (Gaussian) density function is often used as the kernel function, which leads to:

$$w(y - y_i; h) = \phi(y - y_i; h) \tag{2}$$

where $\phi(y - y_i; h)$ is the normal density function, with mean 0 and standard deviation h; the complete formula for the function is as follows:

$$w(u) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2}) = \phi(u) \tag{3}$$

h is often called a smoothing parameter or a bandwidth. It affects the spread of the probability of each observation. Choosing w to be smooth will also lead to a smooth density estimate. The kernel function $w(z)$ has to be non-negative and to satisfy the following properties:

$$\int w(z)dz = 1 \tag{4}$$

$$\int zw(z)dz = 0 \tag{5}$$

$$\int z^2 w(z)dz = \kappa_2 < \infty \tag{6}$$

In order to evaluate the properties of kernel functions, the pointwise mean square error (MSE) is often used. The bias and variance are dependent on the bandwidth (bias falls as h decreases, variance rises as h decreases). For this reason, bandwidth is the most important parameter and it can be subject to optimization, for example, by cross-validation; see [4] for details. The cross-validation method minimises the integrated squared error.

After the optimal bandwidth is obtained, one can attempt to obtain an optimal kernel function by choosing one from a given variety of functions, all of which reflect different properties of underlying distributions.

A kernel function optimal with regards to the integrated mean square error was proposed in the literature. This IMSE-optimal weighting function is known as the "Epanechnikov kernel" after the person who first proposed it as a kernel function for density estimation. However, there is a wider range of kernel functions which are also relatively efficient, including the Gaussian kernel. As it turns out, the kernel function does not have a big impact on the outcome and kernel functions are often chosen only on the basis of their respective computational complexity.

In contrary to the kernel function, the bandwidth has a large influence on the final shape of the resulting density estimate. The former is responsible for characteristics such as differentiability and smoothness, whilst the latter has more impact on how the sample affects the resulting estimate. For this reason, it was necessary to develop efficient ways of determining the optimal bandwidth. There are four methods widely proposed in literature [5]:

1. rule-of-thumb,
2. plug-in,
3. cross-validation,
4. bootstrap methods.

The seemingly fastest way is to utilize a reference distribution, such as the normal (Gaussian) distribution. The h is chosen according to the rule-of-thumb, which has the following formula for the Gaussian distribution:

$$h_{opt} = (4\pi)^{-\frac{1}{10}} \frac{3}{8}^{-\frac{1}{5}} \pi^{\frac{1}{10}} \sigma n^{-\frac{1}{5}} \tag{7}$$

thus,

$$h_{opt} = 1.06 \hat{\sigma} n^{\frac{1}{5}} \tag{8}$$

where $\widehat{\sigma}$ denotes the sample standard deviation. This choice of h_{opt} optimizes the integrated squared distance between the estimator and the true density. The calculated bias, variance, and mean square errors hold at any point of the domain, while the integrated mean square error (IMSE) is a global error measure of the MSE over the entire domain of the density yielding a global error measure.

By minimizing the IMSE function with respect to the bandwidth and the kernel function, it is possible to obtain "optimal bandwidths" and "optimal kernels". This expression also provides a basis for data-driven bandwidth selection, where prior knowledge of the patterns in the data is not necessary for successful estimation. It is worth noting that by using IMSE rather than MSE, one can choose the bandwidth in a way which results in a reliable estimate across the whole domain rather than one which is a perfect fit at a single point. The result of this procedure is a bandwidth which globally balances bias and variance by minimizing IMSE with respect to h.

However, it is also worth noting that data-driven bandwidth selection procedures do not guarantee satisfactory results at all times.

3.2. Multivariate density estimation

The kernel estimation was introduced as a continuous, non-parametric alternative to histograms, capable of delivering the features they missed. A smooth, symmetric weight function called a kernel was used in order to allow for differentiation, and it was centered over individual observations in order to replace the method of averaging over intervals. The estimator which was proposed as an alternative to the histogram is known as the Rosenblatt-Parzen estimator. It has received much attention as an unconditional estimator of probability density functions. It can be written in the form:

$$\widehat{f}(y_1, y_2) = \frac{1}{n} \sum_{i=1}^{n} w(y_1 - y_{1i}; h_1) w(y_2 - y_{2i}; h_2) \tag{9}$$

where (h_1, h_2) is a vector of smoothing parameters (bandwidths),
$w(.)$, the kernel function, is a product of two univariate kernels.

If necessary, the method can also be generalized to higher dimensions. However, it is recommended to assume that the individual kernels of each of the components of the product kernel are identical in order to avoid potential issues resulting from high computational complexity. A useful visualization aid in displaying the results of bivariate density estimation comes in the form of a contour plot; using contour lines, one can display data on a two-dimensional plane, which allows to indicate the height of the density, and therefore the frequency of individual pairs of observations.

4. The {sm} package

The {sm} package was developed by Bowman and Azzalini. It contains a variety of tools for performing non-parametric estimation and regression, either uni- or multivariate. Some of the R-specific functions available to the users are: sm.autoregression, sm.density, sm.pca, sm.regression or sm.ancova (an overview of the functions was presented in [6]).

The function of most interest to our research is sm.density, which plots the density functions of univariate and multivariate variables. A contour

plot of the density estimate is also included in the three-dimensional case. The features of the estimate and its display can be controlled through an interactive panel if the {rpanel} package is used.

In order to use the sm.density function, it is necessary to provide the required data x as an argument. It can be a list if the data is one-dimensional, or a two-column matrix in the bivariate case. The argument h is responsible for the smoothing parameter (or two smoothing parameters if there are two dimensions), and, if none is provided, a normal optimal smoothing parameter or a product of two independently calculated default parameters are used.

From the computational point of view, it is possible to make calculations on large datasets less demanding for the processor by making use of binning procedures. If the binning procedure is turned on, binned data shall be used for kernel density estimation, with the number of bins specified by the user or calculated automatically.

5. Construction of density estimates

In our research, we have used daily data on the number of unique users and page views, collected from two distinct, popular Polish websites over the period of February 1, 2012, until December 31, 2013. The dataset consists of 17,544 observations for each of the websites and each of the examined variables.

Due to the technical specifics of the data gathering process, one should remember that there is a connection between both measured variables. Since each of the users is automatically classified as a visitor to the website, there is a lower bound to the number of page views, as it cannot be smaller than the number of unique users. It can, however, grow larger without bounds if each or some of the users click multiple advertisement banners.

The basic characteristics of the data can be initially examined by pairwise comparison of the boxplots in terms of user count and page views:

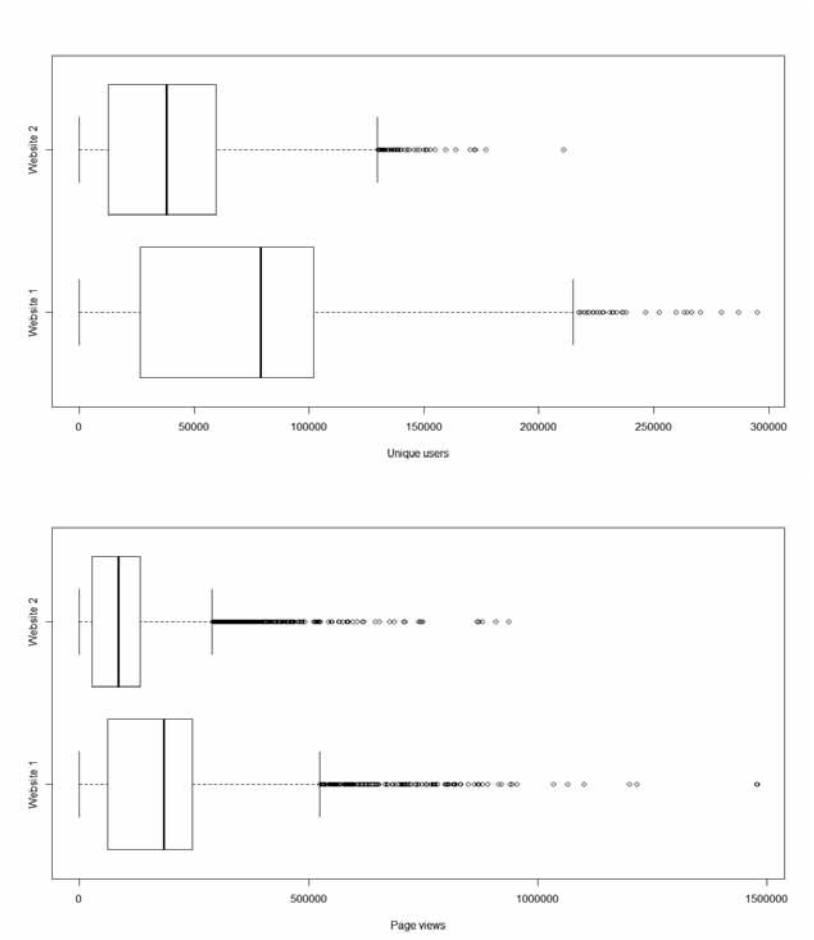


Figure 2: Box plots for users and views data

One can see that the data for the first website is characterized by larger median values as well as a more pronounced spread around the average values. The second website, on the other hand, appears to experience less intense and less variable user activity, both in terms of the number of unique users and the number of views.

The differences between the two metrics can also be observed; the views count is visibly larger than the number of unique users, which can be attributed to the aforementioned specifics of the data; the number of outliers, whose values are significantly larger than the median ones, is also noticeably larger for the views data.

After the initial look at the data, we proceed to non-parametric estimation of the data with the {sm} package. Initially, we obtain estimates for the univariate density functions by applying the sm.density function and choosing the Gaussian kernel coupled with the default rule-of-thumb approach for bandwidth selection.

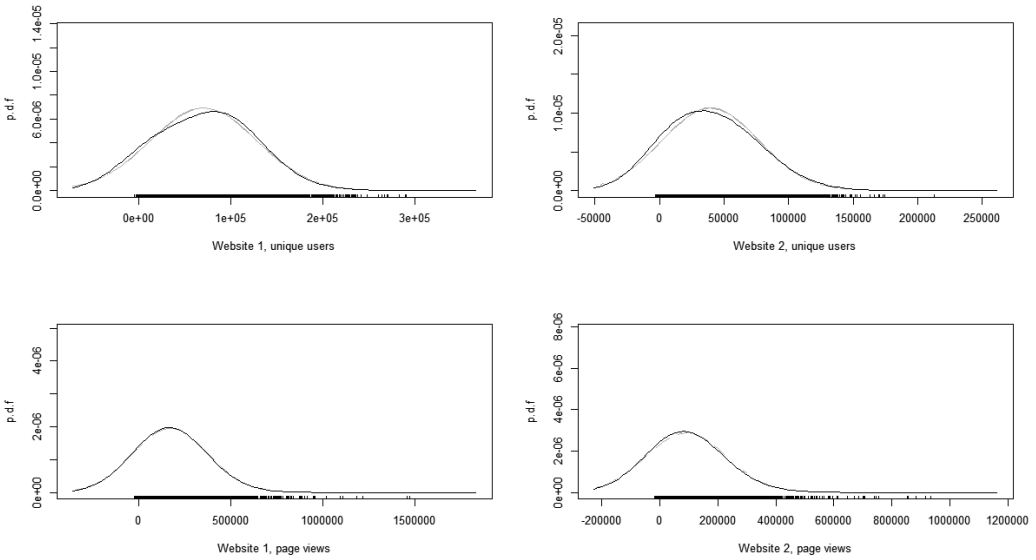


Figure 3: Estimated univariate density functions

The presence of outliers is shown to be skewing all four distributions by affecting their right tails. Their shape, however, is reminiscent of a normal distribution, whose curve is marked on the plots with a lighter color. The overlapping of the empirical and normal distributions is more visible for page views data.

The next step in the analysis is to obtain multivariate estimates for the data as a whole, taking advantage of the relations between the number of users and view count, which cannot be captured by univariate distributions alone. By using the sm.density command on vectors of data, we have automatically obtained 3D plots of bivariate distributions and their respective contour plots. The default product kernel based on individual Gaussian kernels was used along with the rule-of-thumb approach to bandwidth selection.

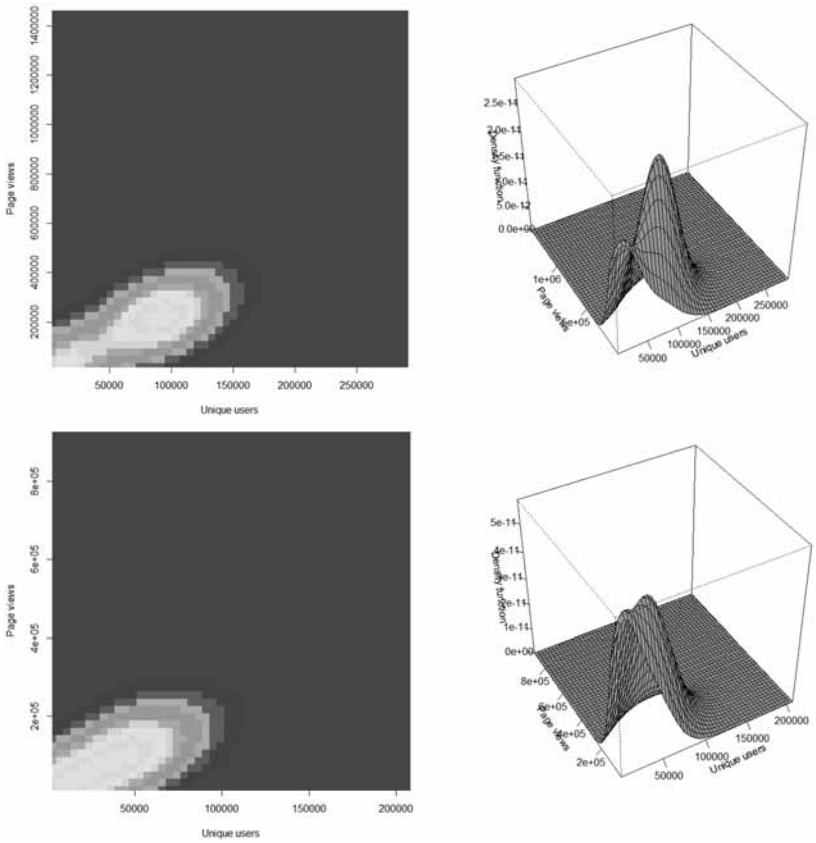


Figure 4: Estimated bivariate density functions

One can see, from both the perspective and the contour plots, that the distribution for the first website appears to be clearly bimodal. This characteristic is also present in the other distribution, albeit to a lesser degree. The perspective plot appears to be showing two peaks in the data, but the contour plot shows that the corresponding values of the density function for these points do not belong to different quartiles (as both peaks lie within the same-colored area of the plot). All plots are also affected by outliers which expand their range while remaining insignificant in terms of relative frequency among the observations.

The bivariate distributions also have steep slopes, which might confirm the relation between the number of users and page views. Uncorrelated observations, where e.g. a relatively small number of unique users corresponds to a relatively large number of page views, are encountered less frequently (have

smaller corresponding values on the z axis) than those where the frequency of one variable is similar to that of the other one.

6. Conclusion

Exploring and classifying patterns of user behavior is an important part of modern marketing research. The data regarding user activity is often unordered and very large in quantity, and thus needs to be properly processed and visualized in order to allow for in-depth investigation. The {sm} package has proved itself to be an accurate non-parametric toolkit for performing such visualizations. Even though it is described as sensitive to the initial setting of parameters, such as the appropriate bandwidth, it has produced reasonable estimates of the density functions without being given any additional input or specific knowledge of the data. By performing kernel estimation of density functions, it creates accurate plots which can be further examined to find additional patterns within the data.

The densities we have obtained possess certain distinct features. The univariate density functions for the number of unique users and the number of page views are both shown to be approximately normally distributed, albeit fat-tailed due to the presence of outliers. Bivariate density functions are shown to be following an approximately bimodal distribution, which might point to some patterns of user activity where the number of users is more subtly related to the number of page views.

Our research has shown that the sm package is well-suited for discovering basic features in the data, but such an approach is far from exhaustive. Further research may include exploring the hidden patterns by applying more sophisticated methods of estimation or using robust methods to process the data without the strong influence of outliers. A thorough understanding of user activity patterns would be beneficial to all companies that operate in the Internet, as it would allow them to launch effective strategies and target potential customers with more efficiency.

References

[1] Bowman, A.W., & Azzalini, A. (2014). *R Package 'sm': Nonparametric Smoothing Methods (Version 2.2-5.4)*. University of Glasgow, UK and University di Padova, Italia.

[2] Bowman, A.W., & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations* (Vol. 18). OUP Oxford.

[3] Google Heatmap (2015). Retrived on 20/06/2015, from http://www.cms.rk.edu.pl/w/p/sledzenie-kliniec-i-mapy-cieplne/.

[4] Härdle, W., & Simar, L. (2007). *Applied Multivariate Statistical Analysis* (3rd ed). Berlin: Springer.

[5] Racine, J.S. (2008). Nonparametric Econometrics: A Primer. *Foundations and Trends in Econometrics*, 3(1), 1-88.

[6] Scrucca, L. (2001). Nonparametric Kernel Smoothing Methods. The sm Library in Xlisp-Stat. *Journal of Statistical Software*, 6(7), 1-49.

[7] Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009). How Much Can Behavioral Targeting Help Online Advertising? [in:] *Proceedings of the 18th International Conference on World Wide Web* (pp. 261-270). ACM.

On Using Permutation Tests in the Data Homogeneity Analysis

Dominika Polko, Grzegorz Kończak

University of Economics in Katowice, Poland

Abstract

The chi-square test for homogeneity of proportions is used when some independent samples are categorized on a single dimension. If the data are homogenous the proportions of the observations in the j-th category will be equal in all of the samples. In the case 2×2 contingency tables it is possible to employ a one-tailed alternative hypothesis. The proposal of the method for testing directional hypothesis is presented in the paper. Idea of application the proposed method is illustrated with an empirical example.

Keywords: contingency tables, directional hypothesis, permutation test, Monte Carlo

1. Introduction and basic notations

There are two chi–square tests that can be applied with contingency tables. The first one is the independence test and the second is the homogeneity test. The chi–square independence test can be employed if a single sample is categorized on two variables. It is assumed that the sample is randomly selected from the population.

The chi–square test for homogeneity can be used if r independent samples are categorized on a single dimension which consists of c categories ($r \geq 2$, $c \geq 2$). This test assumes that the sums of r rows are determined by the researcher and represents the number of observations in each sample.

Table 1 presents the model for the data from r populations. The entries n_{ij} ($i = 1, 2, \ldots, r$, $j = 1, 2, \ldots, c$) are the counts for every two–way combination of row and column (every cell).

Population	**Column variable**				**Row**
	$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{\ldots}$	$\boldsymbol{y}_c$	**sums**
1	n_{11}	n_{12}	$\ldots$	n_{1c}	$n_{1\bullet}$
2	n_{21}	n_{22}	$\ldots$	n_{2c}	$n_{2\bullet}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
r	n_{r1}	n_{r2}	$\ldots$	n_{rc}	$n_{r\bullet}$
Column sums	$n_{\bullet 1}$	$n_{\bullet 2}$	$\ldots$	$n_{\bullet c}$	n

Table 1: Contingency table (r populations)

The total number of observations in Table 1 is equal to

$$n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = \sum_{i=1}^{r} n_{i\bullet} = \sum_{j=1}^{c} n_{\bullet j}. \tag{1}$$

The distribution of the column variable for each population is presented in Table 2.

Population	**Column variable**				**Row**
	$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{\ldots}$	$\boldsymbol{y}_c$	**sums**
1	p_{11}	p_{12}	$\ldots$	p_{1c}	1
2	p_{21}	p_{22}	$\ldots$	p_{2c}	1
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
r	p_{r1}	p_{r2}	$\ldots$	p_{rc}	1
Column sums	$p_{\bullet 1}$	$p_{\bullet 2}$	$\ldots$	$p_{\bullet c}$	1

Table 2: The probabilities in contingency table model

The value p_{ij} represents the probability that the observation from i^{th} population is equal to y_j. This probability can be written as follows $p_{ij} = P(Y = y_j|X = i)$. The probabilities p_{ij} are usually unknown and can be estimated using the following formula $\hat{p}_{ij} = \frac{n_{ij}}{n_{i\bullet}}$. The null hypothesis in homogeneity testing can be stated as follows: "in the underlying populations the samples represent all of the proportions in the same column of the $r \times c$ table are equal". The alternative hypothesis is usually formulated: "in the underlying populations the samples represent all of the proportions in the same column are not equal for at least one of the columns". Formally these hypotheses can be written as follows:

$$\begin{aligned} H_0 : p_{11} &= p_{21} = \cdots = p_{r1} \\ p_{12} &= p_{22} = \cdots = p_{r2} \\ &\cdots \\ p_{1c} &= p_{2c} = \cdots = p_{rc} \end{aligned} \tag{2}$$

and the alternative

$$H_1 : p_{ik} \neq p_{jk} \tag{3}$$

for some i,j=1,2,...,r and k=1,2,...,c.
It could be used chi–square statistics to test the homogeneity hypothesis

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{4}$$

where $\hat{n}_{ij}$ for $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$ are the expected frequencies. Where there are more than two rows and at least two columns instead of the chi-square test the Freedman-Halton test can be used [3]. The statistic (3) has asymptotic chi–square distribution with (r-1)(c-1) degrees of freedom. The chi–square test for contingency tables is based on the following assumption [1, 8]:

a) Categorical or nominal data for $r \times c$ mutually exclusive categories are employed in the analysis.

b) The data that are evaluated represent a random sample comprised of n independent observations. This assumptions reflects the fact that each subject or object can only be represented once in the data.

c) The expected frequency of each cell in the contingency Table is 5 or greater.

The chi–square distribution only provides an approximation of the exact sampling distribution for a contingency table. The accuracy of the chi–square approximation increases as the size of sample increases and, except for instances involving small sample sizes, the chi–square distribution provides an excellent approximation of the exact sampling distribution.

2. Testing directional hypothesis

The special case of the contingency table is table with 2 rows and 2 columns (Tab. 3).

	Column 1	**Column 2**	**Row sums**
Row 1	n_{11}	n_{12}	$n_{11} + n_{12}$
Row 2	n_{21}	n_{22}	$n_{21} + n_{22}$
Column sums	$n_{11} + n_{21}$	$n_{12} + n_{22}$	n

Table 3: Model for 2×2 contingency table

The null and the alternative hypotheses usually have the following form

$$\begin{aligned} H_0 &: p_{11} = p_{21} \\ H_1 &: p_{11} \neq p_{21} \end{aligned} \tag{5}$$

In the case of 2×2 contingency tables it is possible to employ a directional (one tailed) alternative hypothesis [4]. The null hypothesis and directional alternative hypothesis can be stated as follows [5]:

$$\begin{aligned} &H_0 : p_{11} = p_{21} \\ &H_1 : p_{11} > p_{21} \text{ or } H_1 : p_{21} > p_{11} \end{aligned} \tag{6}$$

For small samples with expected values in cells less than 5 C.R. Rao in [7] suggests to use Yates continuity correction for chi–square statistic. A. Agresti argues that there is no longer any reason to use this approximation (see: [1]). Modern software makes it possible to use Fisher exact test. When evaluating contingency tables with r rows ($r > 2$) and c columns ($c > 2$) then directional multi–tailed alternative hypothesis can be employed.

The chi–square statistic (4) can be used for testing the hypothesis (2) against (3) only if expected numbers for all cells are greater than 5. In the chi–square test rows and columns are treated equivalently. The analysis of two way tables can be extended to the multi–way tables [6]. To test the hypothesis (2) against the one–directional hypothesis the permutation test can be used. In permutation testing the principal issue is deciding on an appropriate test statistic. Let us consider the hypothesis (2) with the alternative

$$H_1 : p_{ik} > p_{jk} \tag{7}$$

for some $i, j = 1, 2, \ldots, r$ and $k = 1, 2, \ldots, c$.

The test statistics can be written as follows

$$T = \hat{p}_{ik} - \hat{p}_{jk} \tag{8}$$

where $\hat{p}_{ik} = \frac{n_{ik}}{n_{i\bullet}}$.

In permutation testing we derive the achieved significance level (ASL) from the permutation distribution.Having observed T_0, the ASL is defined to be probability of observing at least that large value when the null hypothesis is true [2]

$$ASL = P_{H_0}\{T \geqslant T_0\} \tag{9}$$

The smaller the value of ASL, the stronger evidence against H_0. Formally we choose a significance level α and reject H_0 if ASL is less than α. ASL can be approximated by

$$\hat{ASL} = \frac{card\{i : T_i \geqslant T_0\}}{N} \tag{10}$$

where T_i denotes the value of the test statistic in i^{th} permutation.

3. Empirical example

The idea of application the proposed method presents following example. Table 4 presents collected data from Social Diagnosis in contingency table. Data contains an assessment of level of satisfaction with own family's financial situation in 2013 (http://www.diagnoza.com 15.05.2015). Included in the table data related to selected four Polish voivodeships. The degree of satisfaction with the financial situation of the family is defined on six levels (1 – 6): very satisfied (1), satisfied (2), rather satisfied (3), rather dissatisfied (4), dissatisfied (5) and very dissatisfied (6).

The null hypothesis (1) in homogeneity testing stated as follows: "in the voivodeships the samples represent all of the proportions of persons with the level of satisfaction with their own family's financial situation are the same" was being verified, towards three alternative hypotheses:

a) $H_{1a} : p_{21} > p_{11}$, saying that "proportions of persons that are very satisfied with own family's financial situation is greather in małopolskie voivodeship than in dolnośląskie voivodeship",

b) $H_{1b} : p_{42} > p_{32}$, saying that "proportions of persons that are satisfied with own family's financial situation is greather in śląskie voivodeship than in mazowieckie voivodeship",

c) $H_{1c} : p_{36} > p_{26}$, saying that "proportions of persons that are very dissatisfied with own family's financial situation is greater in mazowieckie voivodeship than in małopolskie voivodeship".

In Table 4 were underlined cells (values for two selected populations – compared voivodeships) within each tested category for satisfaction with financial situation of the family in 2013.

Voivodeship	**Level of satisfaction with their own family's financial situation**						**Row sums**
	1	**2**	**3**	**4**	**5**	**6**	
dolnośląskie	62	399	556	249	271	184	1721
małopolskie	96	395	712	291	250	135	1879
mazowieckie	102	734	1109	524	449	249	3167
śląskie	167	778	934	410	317	172	2778
Column sums	427	2306	3311	1474	1287	740	9545

Table 4: Assessment of the degree of satisfaction with the financial situation of the family in 2013

The calculation procedure for verification the hypothesis using a permutation test can be written as follows:

1) The level of significance α has been assumed,

2) The value of test statistics T_0 on the basis of (8) has been calculated for data in contingency table,

3) N permutations of variables were performed and values of statistics T_i (i =1, 2, ..., N) have been determined,

4) The decision is made on the basis of ASL value. If $ASL < \alpha$, then H_0 is rejected, otherwise H_0 hypothesis cannot be rejected.

All calculations were performed in R program (http://cran.r-project.org). Empirical distributions of test statistics T and values of statistics T_0 are presented on Figure 1. Significance level $\alpha = 0.05$ in all performed tests was assumed. N=10000 permutations of variables were used.

Calculation results are presented in Table 5. Permutation tests for three selected directional (one tailed) alternative hypothesis were performed. The

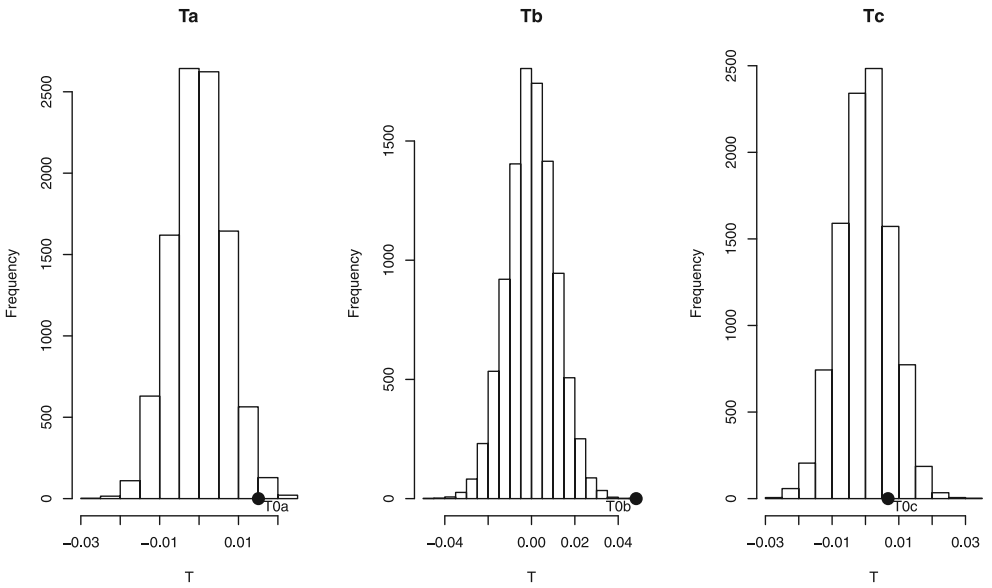


Figure 1: Empirical distributions of test statistics

formula (8) was used as test statistic. In subsequent cases of considered alternative hypotheses test statistics can be written as:

$$T_a = \hat{p}_{21} - \hat{p}_{11} \tag{11}$$

$$T_b = \hat{p}_{42} - \hat{p}_{32} \tag{12}$$

$$T_c = \hat{p}_{36} - \hat{p}_{26} \tag{13}$$

ASL values for hypothesis on data homogeneity with the alternative hypothesis formulated according to the first comparison of voivodeships: małopolskie – dolnośląskie equals 0.0148 and according to the second comparison of voivodeships: śląskie – mazowieckie equals 0. *ASL* values calculated with empirical distributions of statistics are lower than assumed significance level α. In these cases the hypothesis on homogeneity H_0 is rejected in favor of hypothesis H_1. In the third case (comparison of voivodeships: mazowieckie – małopolskie) *ASL* value equals 0.1918 therefore hypothesis H_0 cannot be rejected. Calculations results indicate that:

a) in małopolskie voivodeship "very satisfied" category occurs more often than in dolnośląskie voivodeship,

Comparison number	**Compared two voivodeships (populations)**	**Selected category (column variable)**	**Test statistics (T_0)**	***ASL* value**
1	małopolskie–dolnośląskie	very satisfied	0.0151	0.0148
2	śląskie–mazowieckie	satisfied	0.0483	0.0000
3	mazowieckie–małopolskie	very dissatisfied	0.0068	0.1918

Table 5: Calculation results

b) in śląskie voivodeship "satisfied" category occurs more often than in mazowieckie voivodeship,

c) there is no basis for rejection null hypothesis saying that "in the voivodeships the samples represent all of the proportions of persons with the level of satisfaction with their own family's financial situation are the same".

4. Conclusion

The homogeneity test for contingency table is analyzed in the paper. The most interesting for the researcher are directional alternative hypotheses. The proposal of testing directional hypotheses for $r \times c$ contingency tables was presented in the paper. The method is based on data permutations. The principal issue in permutation analysis is deciding on an appropriate test statistic that discriminates between the null hypothesis and the alternative. Permutation tests are subject to two sources of variability: the sample is chosen at random from the population, and the resamples are chosen random from the original sample. The idea of the proposed method was illustrated in example. The data was taken from the Social Diagnosis project in which the subjective quality of live in Poland is determined in years 2000-2013.

References

[1] Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: John Wiley & Sons.

[2] Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
[3] Good, P. (2005). *Permutation, Parametric and Bootstrap Tests of Hypotheses*. New York: Springer Science Business Media, Inc.
[4] Good, P. (2006). *Resampling Methods. A Practical Guide to Data Analysis*. Boston: Birkhauser.
[5] Kończak, G. (2012). On Testing Multi-directional Hypotheses in Categorical Data Analysis. [in:] A. Colubi, E.J. Kontoghiorghes, K. Pokianos & G. Gonzalez-Rodriguez (Eds.), *Proceedings of COMPSTAT 2012*.
[6] Polko, D., & Kończak, G. (2014). On the Method of Comparing Populations Structures Based on the Data in the Contingency Tables. *Folia Oeconomica, Acta Universitatis Lodziensis*, 3(302), 81-89.
[7] Rao, C.R. (1982). *Modele liniowe statystyki matematycznej*. Warszawa: PWN.
[8] Sheskin, D.J. (2004). *Handbook of Parametric and Nonparametric Statistical Procedure*. Boca Raton: Chapman & Hall/CRC.

A Note on the Nonstationary Functional Time Series

Daniel Kosiorowski[a], Jerzy Rydlewski[b], Małgorzata Snarska[a]

[a] *Cracow University of Economics, Poland,*
[b] *AGH University of Science and Technology, Poland*

Abstract

Functional data analysis [8, 7] is a part of modern multivariate statistics that analyses data providing information about curves, surfaces or anything else varying over a continuum. We often have to work with a functional data, where we cannot easily decide, whether they are to be considered as stationary or nonstationary data. However the definition of nonstationary functional data is a bit vogue. A quite fundamental issue is that before we try to statistically model such data, we need to check whether these curves (suitably transformed, if needed) form a stationary functional time series. We present precise mathematical definitions of such a data. Inference for functional data almost always assumes that the data are stationary, possibly after some transformation and strict stationarity is strongly preferred. At present there are no suitable tests of stationarity for such functional data. Finally we set our own proposition.

Keywords: functional data analysis, nonstationaty time series, methods

1. Functional nonstationary process

There are considered in literature [1, 3] only processes with values in Banach or Hilbert space (which is also a Banach space). The definition of strictly stationary process is quite understandable.
Assume that T is an additive subgroup of $\mathbb{R}$ e.g., $T = \mathbb{R}$ or $T = \mathbb{Z}$. Let consider a real process $X = (X_t, t \in T)$ and let

$$\tau_X^h = (X_{t+h}, t \in T), h \in T \tag{1}$$

Let P represent a cumulative distribution function of the joint distribution of X. In [1] it is carefully proved that such cds's exist for functional processes

with values in Banach or Hilbert space.
The process X is strictly stationary if

$$P_{\tau_X^h} = P_X, h \in T \tag{2}$$

The author shows that process X is strictly stationary if and only if

$$P_{X_{t_1+h},...,X_{t_k+h}} = P_{X_{t_1},...,X_{t_k}}, k \geq 1, t_1, ..., t_k, h \in T \tag{3}$$

The definition and property directly extend to processes with values in a separable Banach space or in a separable Hilbert space.
Let now B be a separable real Banach space with $||\cdot||$ norm and let B^* be a space topologically dual to a B space – which is a space of all linear and continuous functionals on B. $||x^*|| = \sup\{|x^*(x)|, x^* \in B^*, x \in B, ||x|| \leq 1\}$ is a most common norm in B^*.
A Banach space valued random variable X is weakly integrable if $x^*(X)$ is integrable for all $x^* \in B^*$ and if there exists an element of B, denoted by EX, such that

$$E(x^*(X)) = x^*(EX),\ x^* \in B^* \tag{4}$$

We call EX the weak expectation or weak integral of random variable X. Random variable X is said to be integrable (or strongly integrable) if $||EX|| \leq \infty$ If X is strongly integrable, then EX is called the expectation of X. Let X be a squared integrable Banach valued random variable and let $EX = 0$. The covariance operator of X, denoted by C_X, is the bounded linear operator from B^* to B, defined by

$$C_X(x^*) = E[x^*(X)X],\ x^* \in B^* \tag{5}$$

If $EX \neq 0$, we set $C_X = C_{X-EX}$. The operator C_X is completely determined by the covariance function of X, which is defined as

$$c_X(x^*, y^*) = y^*[C_X(x^*)] = E[x^*(X)y^*(X)] = Cov(x^*(X), y^*(X)) \tag{6}$$

for $x^*, y^* \in B^*$.
Let's now define weakly a stationary process. Let $X = (X_n, n \in \mathbb{Z})$ be a Banach space valued random process. X is said to be a (weakly) stationary if $E||X_n||^2 < \infty$, $n \in \mathbb{Z}$, $EX_n = \mu$ does not depend on n and the covariance functions of X satisfy the following

$$E(X^*(X_{n+h} - \mu)y^*(X_{m+h} - \mu)) = E(x^*(X_n - \mu)y^*(X_m - \mu)) \tag{7}$$

$n, m, h \in Z$ and $x^*, y^* \in B^*$.
In their book [3] Horváth and Kokoszka do assume stationarity. They look at the random curve $X = \{X(t), t \in [0,1]\}$ as a random element of the space $L^2 = L^2([0,1])$ equipped with the Borel $\sigma-$algebra. The L^2 is a separable Hilbert space with the inner product $< x, y >= \int x(t)y(t)dt$. Thus all definitions and theorems formulated in Banach space are valid as well. They can be rewritten in the following form. X is integrable if $E||X|| = E[\int X^2(t)dt]^{\frac{1}{2}} < \infty$. If random curve X is integrable, there exists a unique random function $\mu \in L^2$ such that for any $y \in L^2$ we have

$$E < y, X >=< y, \mu > \tag{8}$$

So we have $\mu(t) = E[X(t)]$ for almost all $t \in [0,1]$. Note, that if X is integrible the operator does commute with any bounded operator Φ: $E\Phi(X) = \Phi(EX)$. If X is square integrable and $EX = 0$ we can define covariance operator of X by

$$C(y) = E[< X, y > X], y \in L^2 \tag{9}$$

It is $C(Y)(t) = \int c(t,s)y(s)ds$, where $c(t,s) = E[X(t)X(s)]$.
To sum up, when we have observations $X_1, ..., X_N$ which are iid and which have the same distribution as square integrable X we can define the following:

$$\begin{gathered} \mu(t) = E[X(t)] \\ c(t,s) = E[X(t) - \mu(t))(X(s) - \mu(s))] \\ C = E[< (X - \mu), \cdot > (X - \mu)] \end{gathered} \tag{10}$$

called mean function, covariance function and covariance operator, respectively. They need to be estimated in practice. Their estimators are:

$$\hat{\mu}(t) = \frac{1}{N}\sum_{i=1}^{N}[X_i(t)] \tag{11}$$

which is sample mean function,

$$\hat{c}(t,s) = \frac{1}{N}\sum_{i=1}^{N}(X_i(t) - \hat{\mu}(t))(X_i(s) - \hat{\mu}(s))] \tag{12}$$

which is sample covariance function

$$\hat{C}(x) = \frac{1}{N}\sum_{i=1}^{N} < X_i - \hat{\mu}, x >)(X_i - \hat{\mu}), x \in L^2 \tag{13}$$

which is sample covariance operator.
It should be noticed, that sample covariance operator $\hat{C}$ maps the space L^2 into a finite dimensional subspace spanned by $X_1, ..., X_N$. Because a naturally finite sample can show an infinite dimensional object with a limited precision, this can be regarded as the illustration of limitations of statistical inference in a case of functional observations.
Horváth and Kokoszka show that the presented estimators are unbiased and MSE consistent estimators. Horváth and Kokoszka prove more lemmas and theorems on the preceding estimators. However they always use the property of stationarity.

2. Testing stationarity

In their paper Horváth et al. [4] authors formalize the assumption of stationarity in the context of functional time series and proposes several procedures to test the null hypothesis of stationarity. The properties of the tests under some rigorously defined alternatives are studied and new properties present only in the functional setting are uncovered. The theory is illustrated by a simulation study.
As the authors have noted the spectral analysis of nonstationary functional time series has not been developed to a point where usable extensions could be readily derived, so in their paper Horváth et al. developed a general methodology for testing the assumption that a functional time series to be modeled is indeed stationary. A test should be applied before fitting one of the known stationary models. Often it should be applied to transformed functions in order to remove seasonality or trend.
The authors considered stationary functional time series represented as

$$X_n(t) = \mu(t) + \sum_{j=1}^{+\infty} \sqrt{\lambda_j} \xi_{jn} v_j(t) \tag{14}$$

where n is the time index that counts the functions, t is the argument of each function. μ is the mean function and v_j are the functional principal components. They are unknown deterministic functions which depend on the stochastic structure of the series X_n. They are obviously estimated by random functions $\hat{\mu}$ and $\hat{v_j}$.
Authors note that if X_n is not stationary the estimators do not converge to μ and v_j because these population quantities will not exist then. For

this reason authors developed an analysis of the behavior of the tests under alternatives.
Linear functional time series have the form

$$X_n = \sum_j \Psi_j(\epsilon_{n-j}) \tag{15}$$

where ϵ_i are iid error functions, and the Ψ_j are bounded linear operators acting on the space of square integrable functions. Authors assumed that $X_n = \overline{f}(\epsilon_n, \epsilon_{n-1}, ...)$, for some, possibly nonlinear, function $\overline{f}$.

Functional autoregressive process and for the more general nonlinear moving averages these functional are in a class which is customarily referred to as weakly dependent or short memory time series. The authors state the conditions for the error process, denoted by $\eta = \{\eta_j\}_{-\infty}^{+\infty}$ which they used to formulate the null and alternative hypotheses.
Let η_j be a sequence of Bernoulli shifts i.e. $\eta_j = f(\epsilon_j, \epsilon_{j-1}, ...)$ for some measurable function f and for iid functions ϵ_j with values in a measurable space. Moreover, $\epsilon_j(t) = \epsilon_j(t, \omega)$ is jointly measurable in (t, ω).

$E\eta_0(t) = 0$ for all t, and $E||\eta_0||^{2+\delta} < \infty$ for some $\delta \in (0, 1)$.
The special assumption taken by the authors states that the sequence $\{\eta_n\}_{n=-\infty}^{+\infty}$ can be approximated by l-dependent sequences $\{\eta_{n,l}\}_{n=-\infty}^{+\infty}$ i.e. $\sum_{l=1}^{\infty}(E||\eta_n - \eta_{n,l}||^{2+\sigma})^{\frac{1}{\kappa}} < \infty$ for some $\kappa > 2 + \sigma$ where $\eta_{n,l} = g(\epsilon_n, \epsilon_{n-1}, ..., \epsilon_{n-l+1}, \epsilon^*_{n,l})$ where $\epsilon^*_{n,l} = (\epsilon^*_{n,l,n-l}, \epsilon^*_{n,l,n-l-1}, ...)$, where $\epsilon^*_{n,l,k}$ are independent copies of ϵ_0 which are in turn independent of ϵ_i. The assumption means that the function g decays relatively fast and that shocks back in the past are small enough so that they can be replaced by their independent copies. It slightly changes the distribution of the process.
Assumptions similar to those stated by Horváth et al. have been used commonly in recent theoretical works. It is due to the fact that all stationary time series models in practical use can be represented as Bernoulli shifts. Bernoulli shifts in turn are stationary by construction.
Finally, after taking above assumptions, the authors want to test a hypothesis:
$H_0 : X_i(t) = \mu(t) + \eta_i(t)$, $1 \leq i \leq N$, and $\mu \in L^2$.
The mean function μ is unknown. The null hypothesis says that the functional time series is stationary and weakly dependent, with the structure of dependence quantified by conditions formulated above. The most general alternative is that H_0 does not hold, however the authors considered some specific alternatives. They focused on the following three alternatives.

1. Change point alternative:
$H_{A,1}: X_i(t) = \mu(t) + \delta(t) I_{\{i>k^*\}} + \eta_i(t), 1 \leq i \leq N$, with some integer $1 \leq k^* < N$.
Functions $\mu(t)$, the size of the change $\delta(t)$, and the time of the change, k^*, are all unknown parameters. Authors assumed that the change occurs away from the end points, i.e. $k^* = [N\tau]$ with some $0 < \tau < 1$.
2. Integrated alternative:
$H_{A,2}: X_i(t) = \mu(t) + \sum_{l=1}^{i} \eta_l(t), 1 \leq i \leq N$.
3. Deterministic trend alternative:
$H_{A,3}: X_i(t) = \mu(t) + g(i/N)\delta(t) + \eta_i(t), 1 \leq i \leq N$ where $g(t)$ is a piecewise Lipschitz continuous function on $[0, 1]$. The trend alternative includes many change point alternatives, including $H_{A,1}$, and alternativese in which change can be gradual.
The tests developed by the authors are consistent against any other sufficiently large departures from stationarity and weak dependence. In order to test the hypotheses the following statistics are defined.

$$T_N = \int \int Z_N^2(x,t) dt dx \tag{16}$$

where $Z_N(x,t) = S_N(x,t) - xS_N(1,t), 0 \leq x, t \leq 1$,
with $S_N(x,t) = N^{-1/2} \sum_{i=1}^{[Nx]} X_i(t), 0 \leq x, t \leq 1$ and the second test statistic is defined as $M_N = T_N - \int \left(\int Z_N(x,t) dx \right)^2 dt = \iint \left(Z_n(x,t) - \int Z_N(y,t) dy \right)^2 dx dt$.
A general idea for tests is that under hypothesis H_0, a statistics $Z_N(x,t)$ computed from the observations X_i is the same as if it were computed from the unobservable errors η_i. Under alternative H_A, deterministic or random trends show up in $Z_N(x,t)$. The same idea as in the scalar case. The difficulty is, how to define the test statistics and their limits for functions.
The null limit distributions of T_N and M_N depend on the eigenvalues of the long-run covariance function of the errors:

$$C(t,s) = E\eta_0(t)\eta_0(s) + \sum_{l=1}^{\infty} E\eta_0(t)\eta_l(s) + \sum_{l=1}^{\infty} E\eta_0(s)\eta_l(t) \tag{17}$$

It is proven in Horváth et al. [4] that the series is convergent in L^2. The covariance function $C(t,s)$ is positive definite, and therefore there exist $\lambda_1 \geq \lambda_2 \geq ... \geq 0$ and orthonormal functions $\phi_i(t), 0 \leq t \leq 1$, satisfying

$$\lambda_i \phi_i(t) = \int C(t,s)\phi_i(s) ds, 1 \leq i < \infty \tag{18}$$

Kernel estimators $\hat{C}$ can be constructed which give empirical counterparts of the eigenvalues λ_i and the eigenfunctions ϕ_i:

$$\hat{\lambda}_i\hat{\phi}_i(t) = \int \hat{C}_N(t,s)\hat{\phi}_i(s)ds, 1 \leq i \leq N \tag{19}$$

The theorem, proven by the authors, specifies limit distributions of T_N and M_N under the stationarity null hypothesis. Throughout the paper let $B_1, B_2, ...$ be independent Brownian bridges.
Theorem states that if assumptions and H_0 hold, then

$$T_N \to \sum_{i=1}^{\infty} \lambda_i \int B_i^2(x)dx \tag{20}$$

where the convergence is in distribution and

$$M_N \to \sum_{i=1}^{\infty} \lambda_i \int \left(B_i(x) - \int B_i(y)dy\right)^2 dx \tag{21}$$

where the convergence is in distribution. As the sum $\sum_{i=1}^{\infty} \lambda_i < \infty$ the variables T_0 and M_0 are finite with probability one.
The theorem shows that for functional time series a simple normalization with a long-run variance is not possible, and approaches involving the estimation of all large eigenvalues must be used. The limits in the theorem (approximating the null distribution) can be approximated with $\sum_{i=1}^{d} \hat{\lambda}_i \int B_i^2(x)dx$ and $\sum_{i=1}^{d} \hat{\lambda}_i \int \left(B_i(x) - \int B_i(y)dy\right)^2 dx$, where d is suitably large and $\hat{\lambda}_i$ and the $\hat{\phi}_i$ are consistent estimators only under hypothesis H_0.
Horváth et al. [4] considered asymptotic distributions depending on the eigenvalues λ_i, which are regarded as an analog of the long-run variance. By projecting on the eigenfunctions $\hat{\phi}_i$ it is possible to construct statistics whose limit null distributions are parameter free.
In order to have uniquely defined, up to the sign, eigenfunctions the authors assumed $\lambda_1 > \lambda_2 > ... > \lambda_d > \lambda_d + 1 > 0$ and defined

$$T_N^0(d) = \sum_{i=1}^{d} \frac{1}{\hat{\lambda}_i} \int < Z_N(x,\cdot), \hat{\phi}_i >^2 dx \tag{22}$$

and

$$T_N^*(d) = \sum_{i=1}^{d} \int < Z_N(x,\cdot), \hat{\phi}_i >^2 dx \tag{23}$$

The statistic $M_N^0(d)$ and $M_N^*(d)$ are also defined this way.
If certain assumptions hold, then

$$T_N^0(d) \to \sum_{i=1}^{d} \int B_i^2(x)dx \tag{24}$$

and

$$T_N^*(d) \to \sum_{i=1}^{d} \lambda_i \int B_i^2(x)dx \tag{25}$$

where the convergence is in distribution. Similar theorems can be formulated for the other statistic M_N^0 and M_N^*.
T_N^* and M_N^* are d-dimensional projections of T_N and M_N. The distribution of the limit can be found in books. The distributions of the limits can also be expressed in terms of sums of squared normals. It is easy to derive normal approximations, what the authors do with the central limit theorem.
Note that the test based on $T_N^0(d)$ is the usual asymptotic test with parameter free limit distribution and it depends on d. The authors considered four other tests, similar in spirit. They compared all six tests theoretically and using simulations.
The authors [4] considered an asymptotic analysis of the tests introduced. They warn that in the functional setting, there is a fundamentally new aspect i.e. convergence of a scalar estimator of the long run variance must be replaced by the convergence of the eigenvalues and the eigenfunctions of the long run covariance function. Authors derived precise rates of convergence and limits for this function, and use them to study the asymptotic power of the tests. Authors implemented and conducted the finite sample performance of the tests. However, several issues must be considered: the choice of the kernel, the smoothing bandwidth. They found that to implement Monte Carlo tests based on statistics whose limits depend on the estimated eigenvalues, a fast method of calculating replications of these limits must be used. Then the eigenvalues $\hat{\lambda}_i$ and normalized statistics $T_N^0(d)$ and $M_N^0(d)$ can be computed. In next step they use the tests based on the asymptotic distribution of their limits. Then the critical values can be computed. Alternatively, the critical values can be obtained by calculating a large number of replications of $T_N^0(d)$ and $M_N^0(d)$ for any specific functional time series. The authors used iid Brownian motions. However, they note that this method is extremely computationally intensive, if its performance is to be assessed

by simulations - the authors needed almost two months of run time on the University of Utah Supercomputer to obtain the empirical rejection rates for their tests for samples of size 100 and 250 and values of d between 1 and 10.

3. Detecting trend

In their article Fraiman et al. [2] considered functional time series, where a trend is expected. They defined trend and then show test that enable detecting them. The definitions were needed, because an obvious increasing/decreasing trend is doubtful to obtain in functional data.

Let $\{x_t(s) : s \in [a,b]\}_{t=1}^{T}$ be a sequence of T real functions defined on a compact interval $[0,1]$. The proportion of time that each curve matches with the previous maxima or attains new maxima is examined i.e. the authors track sequentially in t the proportion of time s, during the observed time interval $[0,1]$, that each function $x_t(s)$ has attained the maxima accumulated up to year t, amongst $\{x_1(s),...,x_t(s)\}$. For each t the $r_t(\cdot)$ denotes the maximum function up to t-th curve, where $r_t : C[0,1]^t \to C[0,1]$, $r_t(s) = \max\{x_1(s),...,x_t(s)\}$ for $t \in [0,1]$. This maximum function up to t-th curve is called the record function up to year t. The proportion of the time that the curve $x_t(\cdot)$ attains or exceeds the maximum function up to the t-th curve $r_t(\cdot)$ is called $w_t = \int_0^1 I_{\{x_t(s)=r_t(s)\}}ds$, where I_A is the indicator function of set A. Analogously the minimum function up to t-th curve is defined as $r_{-t}(s) = \min\{x_1(s),...,x_t(s)\}$ for $s \in [0,1]$, and the proportion of time that the function $x_t(\cdot)$ overlaps with $r_{-t}(\cdot)$ as $w_{-t} = \int_0^1 I_{\{x_t(s)=r_{-t}(s)\}}ds$.

Definition of strong increasing trend is the following. The sequence $\{x_t(s) : s \in [0,1]\}_{t=1}^{T}$ is said to have a strong increasing trend if it satisfies the following inequalities:

$$0 < w_2 - \frac{1}{2} < ... < w_T - \frac{1}{T} \tag{26}$$

and

$$0 > w_{-2} - \frac{1}{2} > ... > w_{-T} - \frac{1}{T} \tag{27}$$

Definition of weak increasing trend is the following. The sequence $\{x_t(s) :$

$s \in [0,1]\}_{t=1}^{T}$ is said to have a weak increasing trend if it satisfies the following inequalities:

$$\min_{t \in \mathbb{T}} \left(w_t - \frac{1}{t} \right) > 0 \tag{28}$$

where $\mathbb{T} = \{2, 3, ...T\}$.
If there is no trend in the sequence, we expect to obtain $w_t = \frac{1}{t}$ for all t. The equation's consequence is that every curve spends more time than expected at the maximum function up to year t. In many cases the definition may be too restrictive, so the authors formulated third definition, which may capture some pattern of increasing trend.

Definition of k_0-weak increasing trend is the following. The sequence $\{x_t(s) : s \in [0,1]\}_{t=1}^{T}$ is said to have a k_0-weak increasing trend if there exists $k_0 \in \mathbb{T}$ such that for all $k \geq k_0$

$$\overline{w}_{[k]} > 0 \tag{29}$$

where $\overline{w}_t = w_t - \frac{1}{t}$ for $t \in \mathbb{T}$ whose corresponding order statistics are denoted by $\overline{w}_{[2]} \leq \overline{w}_{[3]} \leq \overline{w}_{[T]}$.

The authors developed the nonparametric tests for the proposed increasing trends for a sequence of functional data. They also established their results for a multiple time series of functional data. Thus this kind of nonstationary functional time series is possible to be dealt with. Finally, they perform the tests for Antarctic temperature data.

4. Unit sphere projection approach

D. Liebl [6] introduced a robust version of functional principal component analysis for sparse non stationary data. He used an iterating optimization algorithm to fit an orthogonal factor model to the data and claim that functional principal component analysis (for details see [8, 5]) may not be able to deal with sparse and non stationary functional time series data and extended the procedure to the context of non stationary time series data.
He assumes the curves come from a stochastic process (X_t) for $t = 0, 1, 2, ...$ with realizations in the space of square integrable functions $H = L^2(U)$ on a compact set $U \subset R$. Similarly to multivariate statistics, stationary functional stochastic process are often described by their time invariant mean function and covariance operator. But what happens, if the series of curves,

considered by the authors, has a stochastic and non stationary trend? The author considered a functional random walk model, i.e.

$$X_t = \delta + X_{t-1} + e_t, \text{ with } t = 0, 1, 2, ... \tag{30}$$

with a linear trend, where $\delta \in H$, an initial value from a random function, Z_0, which is normally distributed with mean, μ_Z, and covariance operator, C_Z. $(e_t) \in H$ is a white noise process with mean zero and covariance operator, C_e. We have $EX_t = \mu_Z$ for all t, but the covariance operator depends on t, such that the process defined above is non stationary.

Note that the mean function of the functional random walk is independent of t and the author might investigate, without loss of generality, the properties of the demeaned process $(X_t^*) = (X_t - \mu_Z)$. This operation provides virtually the same functional random walk process as above but with an initial functional random variable that has it's mean equal to zero. We obtain $X_t^* = \delta + X_{t-1}^* + e_t$.
Stationarity is needed in order to use the well established functional principal component analysis. Such traditional techniques as transformation procedures i.e. diferentiation of the time series, (X_t^*), cannot be used, because the process are not observed at equidistant values. The new transformation procedure is proposed. It decomposes the original series into its stationary functional component, $(\hat{X}_t^*)$, and its non stationary univariate random walk component, (Θ_t). The decomposition is inspired by the unit sphere projection of functional data.

In other words the author conducts principal component analysis for non stationary data. Further he transforms the process by the unit sphere projection, that is often used in multivariate robust statistics of iid samples. He proposes to decompose the series (X_t^*), into a functional component $(\hat{X}_t^*) \in H$ and an univariate component$(\Theta_t) \in R$, such that

$$(X_t^*) = (\hat{X}_t^* \Theta_t) \tag{31}$$

The first component is called spherical component while the second is called the scaling component.

A decomposed de-meand random walk process $(\hat{X}_t^*)$ has the following properties (proved as Theorem 2.2):

a) Its spherical component, which can be written in a form $(\hat{X}_t^*) = (\pi X_t^*)$ is a stationary process, where $\pi = (\cdot)/||\cdot||_2$ is the unit sphere projection operator, with $||\cdot||_2 = \sqrt{\int_U (\cdot)^2}$ a norm in H.

b) The covariance operators for the process $(\hat{X}_t^*)$ $t = 1, 2, ...$ are elements of the same space as their non spherical counterparts (the covariance operators for the process (X_t^*)).

As a consequence, the eigenfunctions of their covariance operators are the same as of the covariance operators for (X_t^*).

So the author has obtained that asymptotically, the covariance operators of the non stationary original process, (X_t^*), are the same as the covariance operators of the spherical process $(\hat{X}_t^*)$ except for scale differences. Hence we can estimate the original covariance operators from the stationary spherical series $(\hat{X}_t^*)$. Then we rescale the estimated covariance operator by the scaling component (Θ_t) that has absorbed the scale differences. The K eigenfunctions related to the first K eigenvalues of the spherical covariance operator meet the optimality criterion saying that the mean square error of the projection in functional PCA is minimized.

5. Conclusion

We can clearly combine the above described methods. We can easily be able to detect trends such as in [4, 2]. To sum up, when we deal with a functional problem with additive error, we can apply some transformation procedures. Then can detect trends or we can turn procedures round and apply transformations after removing trend. We are able to test whether after performed transformations we obtain a stationary time series. A stationary functional time series can be treated with the functional principal component analysis.

Year by year the new methods which can help to analyze nonstationary time series do appear. We take a look on some promising ones. However, this field of statistics is still in progress.

References

[1] Bosq, D. (2000). *Linear Processes in Function Spaces*. Springer.
[2] Fraiman, R., Justel, A., Liu, R., & Llop, P. (2014). Detecting Trends in Time Series of Functional Data: A Study of Antarctic Climate Change. *Canadian Journal of Statistics*, 42(4), 597-609.
[3] Horváth, L., & Kokoszka, P. (2012). *Inference for Functional Data with Applications*. Springer Science & Business Media.
[4] Horváth, L., Kokoszka, P., & Rice, G. (2014). Testing Stationarity of Sunctional Time Series. *Journal of Econometrics*, 179(1), 66-82.
[5] Krzyśko, M. (2012). Kernel and Functional Principal Component Analysis. *Multivariate Statistical Methods Lecture Series,* Polish Statistical Society.
[6] Liebl, D. (2010). Modeling Hourly Electricity Spot Market Prices as Non Stationary Functional Times, University of Cologne. Retrieved on 19/06/2016, from http://mpra.ub.uni-muenchen.de/25017/series.
[7] Ramsay, J.O., & Silverman, B.W. (2005). *Functional Data Analysis*. Springer.
[8] Ramsay, J.O., Hooker, G., & Graves, S. (2009). *Functional Data Analysis with R and Matlab*. Springer.

Why Forecasting of Crude Oil Price is Difficult Task? Results from Comparison of Large Set of VAR Models

Sławomir Śmiech

Cracow University of Economics, Poland

Abstract

There are two objectives of the paper. The first one is to identify specification of four dimensional VAR models which generate the most accurate forecast of crude oil. The second one is to asses usefulness for crude oil price forecasting a set of predictors related to different area of economy and commodity market. Brent crude oil are predicted for one step-ahead, with recursive shame, and the forecast period covering 01:2004 -10:014. As a result we can point the best specification of VAR models, as well as the best set of predictors. Unfortunately, the optimal predictors set is time dependent, so it must be chosen very carefully. The analysis is carried out not only to say whether it is possibly to accurate forecast crude oil price, but also try to find answer why.

Keywords: forecast, VAR models, crude oil

1. Introduction

Literature offers two views on the possibility of forecasting the price of crude oil at short horizons. According to the first one, represented by, for example, Hamilton [12] and Davies [10], it is impossible to generate more accurate forecast of crude oil prices than the random walk. Hamilton [12] further argues that low accuracy of crude oil prices forecast is a result of low flexibility of its supply and demand at short horizons. That is why even slight fluctuations in supply or demand, which might be caused by difficult to predict factors, can lead to substantial changes in crude oil prices. The other, more common view on the issue in question is represented by

researchers looking for efficient forecasting models. Baumeister et al. [7] identify three main approaches in literature dealing with forecasting crude oil prices. The first one refers to the predictive accuracy of oil futures prices and is analysed by, for example, Knetsch [19], Alquist and Kilian [3], Reeve and Vigfusson [20] and Alquist et al. [2]. The second one investigates the forecasting efficiency of professional and survey forecasts (see: [21, 2, 8]. The third approach focuses on analysing the forecasting abilities of variables describing supply and demand in the global oil market, especially crude oil inventories and crude oil production as well as macroeconomic fundamentals and exchange rates (which can be seen in studies by [9, 5, 6, 18, 16]). The accuracy of forecasts generated by models using variables describing the crude oil market and macroeconomic fundamentals is comparable to the accuracy of forecasts generated by the benchmark models, i.e. a random walk model and a simple autoregressive model. The results reported in these studies lead to the conclusion that "economic fundamentals help forecast the real price of oil, at least during times of large and persistent movements in economic fundamentals, but only at short horizons" [7]. The objective of this study is assessment of forecasting models for crude oil price which are estimated on a large set of predictors. The use of a large set of variables would show whether the available market information allows one for effective forecasting. The resulting predictions are compared with benchmark models (i.e. the forecast obtained from random walk model). Collected set of variables includes this related to real and financial sphere of world economy as well as euro area and U.S; total crude oil demand; prices of another energy commodity. The study analyzed the different specifications of the VAR models as well as all possible subset of predictors. Forecasting models are developed on the basis of monthly data from the period between January 1995 and October 2014, and they generate one-month ahead forecasts. The accuracy of forecasts constructed by recursive scheme was evaluated for the three sub-periods: i.e. from January 2005 to December 2006, from January 2007 to December 2009, from January 2010 to October 2014. This allows one to assess whether the superior specification of VAR models is resistant to sample selection, and whether the same subsets of predictors in all sub-periods allow to receive the most accurate forecasts. In addition, the effectiveness of the models in the case of changes the direction of the trend in oil prices are analyzed. The use of many variables as predictors, and many specifications help to answer

the question whether it is possible to anticipate turning points using information available on the market.

2. Data

2.1. Variable of interest

The Brent spot price of crude oil is chosen for the verification of the possibility of forecasting prices of fossil fuels. This price, together with WTI, is considered the world benchmark (see, e.g. [6]). The International Monetary Fund serves as a source of data. Crude oil spot price is expressed as real, in constant prices in 2010. The consumer price index in the USA CPIUS is used as the GDP deflator. The analysis is based on monthly data from the period between January 1995 and October 2014. This means that the whole sample period contains 238 monthly observations.

2.2. Predictors used in forecasting real crude oil prices

According to reach literature in the study we use 23 predictors for crude oil price. The whole variable set is divided into three following subsets: macroeconomic variables, financial variables, energy prices. Variables describing real economy include: the global industrial production index (IP_W) and in the euro area (IP_EA) (e.g. [1]) and variables referring to the economic activity (see [16]): the ISM manufacturing index in the U.S. (ISM_US), the PMI manufacturing index in the euro area (PMI_EA) (Purchasing Managers Index - Markit Eurozone Manufacturing PMI), and the German Ifo index (IFO) for the business climate among entrepreneurs in trade and industry published by the Ifo Institute; the Baltic Dry Index (BDI) (see [5]) and the global real economic activity index (IK) proposed by Kilian [15]. The remaining variables refer to the global oil market: the global crude oil production (PR_OIL) (see e.g. [15, 17]) and the world crude oil inventories (INV) [17]. Variables in the second set include: the real short-term interest rates in the U.S. (IR_US) and in the euro area (IR_EA); the real money supply M1 in the U.S. (M1_US) and in the euro area (M1_EA) (see, e.g. [4, 11]); the real effective exchange rates deflated by the consumer price index (narrow index) (2010=100) (RN_US) published by the Bank of International Settlements; the US dollar-euro exchange rate (REX) (see Chen et al., 2010); the Standard and Poor's 500 (SP500) stock price index, the German stock index - DAX (DAX) [22]; the Chicago Board Options Exchange Market Volatility Index (VIX) [14]. The third group of variables

contains the following energy prices: WTI crude oil spot price (WTI), steam coal price in Australian ports (NEWC), steam coal price in Richards Bay port (the Republic of South Africa) (RB), Russian natural gas border price in Germany (NG_RUS) and natural gas spot price in the U.S. (NG_US). The data are taken from the International Monetary Fund (IMF). All prices are expressed as real, in constant prices in 2010. The consumer price index in the USA CPIUS is used as a deflator.

3. Empirical results

3.1. The evaluation of forecast generated by VAR models with all possible combinations of predictors

This part presents the evaluation of forecast accuracy generated by (four dimensional) VAR models estimated for variables$[z_t, x_{1,t}, x_{2,t}, x_{3,t}]$', where z_t is either real Brent crude oil price, or the first difference of crude oil price, $[x_{1,t}, x_{2,t}, x_{3,t}]$' is a set of variables selected from all variables described in section 2. The VAR models taken into consideration include models with and without trend. The number of lags is 1,2,3, and 12 (as Hamilton and Herrera, 2004 show the importance of allowing for long lags in the crude oil price models). Additionally, the VAR models with one lag for the first difference of crude oil prices are evaluated. Taking into consideration the number of variables in the set, each VAR specification (after establishing the lag order and deterministic components) require estimating 1540 models, covering all possible combinations of 22 variables, which yields 13860 forecasting models. Due to the number of models considered in the paper, the assessment of their properties in the sample is omitted. In all these models the vector of forecast is computed following the recursive scheme, which is later assessed with the use of forecast accuracy measures (RMSE).

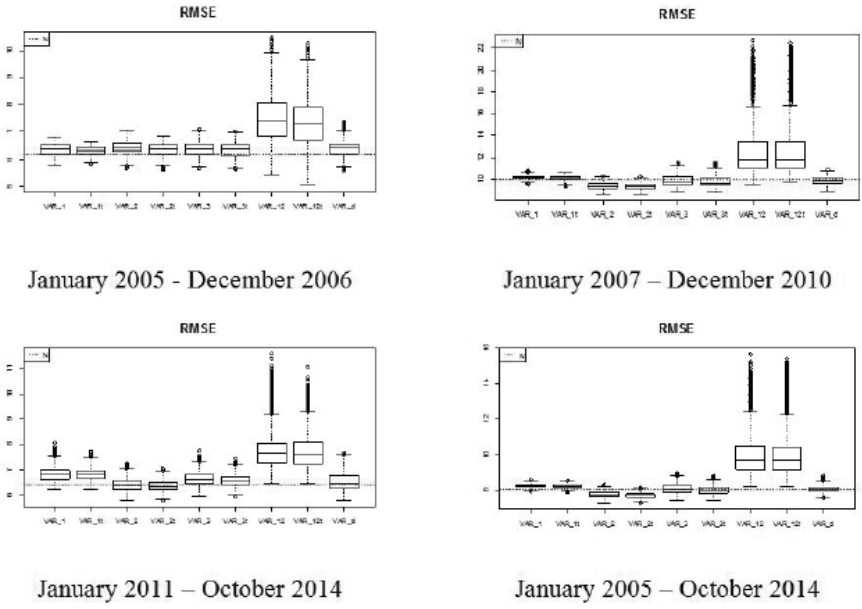


Figure 1: Distribution of forecast error (RMSE) for all considered sub-periods and different specification of VAR models

3.2. Distribution of forecast errors for VAR models with all possible combinations of predictors

Figure 1 presents the distribution of RMSE for all specifications of the VAR models considered in the study and all analysed sub-periods. Symbol 't' added to the name of the model indicates that a deterministic trend is used in this model (e.g. VAR(p)_t). The models for first differences are marked as VAR(1)_d. In the first sub-period (from January 2005 to December 2006) distribution of RMSE obtained for VAR models with lag 1,2,3 with trend and without trend seem to be quite similar. In all this cases, about a quarter of models (combinations of predictor's subset) generate the forecast models for which the RMSE was smaller than that of the naive forecasts. Distributions of errors (measured by RMSE) for VAR(12) and VAR(12)t are characterized by high volatility. However, in the case of these models it is possible to obtain the most accurate forecasts. The lowest value of the RMSE (5.017), is obtained for a subset of variables including: PMI_US, IFO and REX, which was used to estimate the VAR (12)t model.

In the second sub-period (January 2007 – December 2010), the accuracy of forecasts for all models turned out to be low. Even the best model specification, i.e. VAR(2) and VAR (2)t models, generate a prediction error RMSE which was higher than 8.6 (US dollars per barell). On the other hand, in the case of these types of models, almost all subsets of predictors led to construct models characterized by a lower forecast errors than naive forecast. Again, the huge dispersion of forecasting error is obtained for VAR with 12 lags. The lowest forecast error RMSE, which is 8.675, is obtained for the VAR(2) model, which includes following variable: IP_W, VIX and RN_USA. In the third sub-period (January 2011 – October 2014), the most accurate forecast are obtained for the three specifications it is: VAR (2), VAR (2)t and VAR_d. In this case, the median of RMSE distribution is as high as root square error of the naive forecasts. Again the least accurate forecast are obtained for VAR (12) and VAR (12)t models. The smallest forecast error (RMSE 5.767) is obtained for the VAR(2) model and a set of variables consisting of NG_RUS, IFO and RN_US. In the whole forecasting period (January 2005 - October 2014) the best, due to an RMSE error, are VAR(2) and VAR(2) t models. For the VAR(2) models, more than three-quarters of subsets of variables allows to construct more accurate forecasts that this generated by the random walk. Almost all of predictor's subsets for the VAR(2)t give forecast that beat naive forecast. In the case of VAR with 3 lags, and VAR for differences, median of the distribution of RMSE corresponds approximately to the level of a Root Mean Squared Error of naive forecast. The least accurate forecast are generated by VAR with 12 lags. The lowest RMSE (7.289), is obtained for the VAR(2) model, which takes into account variables IP_W, VIX and RN_US. The results obtained demonstrate that in considered periods both an error rate of forecasts and subsets of variables in the best due to forecast accuracy changed.

3.3. Forecast and forecast errors

Due to the lowest level of forecast errors obtained from the VAR with two lags, further analyzes consider only those specifications. Figure 2 presents all paths of forecast obtained from VAR(2) (red color) and VAR(2)t (blue) models for all possible combinations of predictors. Yellow color represent forecast obtained using naive model, while real crude oil price is draw in the black. The picture indicates first, that all models considered produce similar forecast, second all models have problems with anticipations of changes in the directions of the trend of the real price of oil. For example

a huge decrease of oil price in about 20th observations (since September 2006), which was precede by long period of price increase is unpredicted. Similarly, the drop of crude oil price caused by the global financial crisis (it is from 42 observation on the figure).

Prediction errors for different models are presented in the Figure 3. Red and blue color are, like in the previous case, used to indicate VAR(2) and VAR(2)t models respectively. The results obtained confirm that the largest errors occurs when crude oil price decrease sharply. At that time, models for all possible subset of predictors generated biased forecast. It suggest that there are no information in the data sets considered in the study that permit to predict most of turning points. Thus, neither real and financial indicators of world economy nor crude oil supply characteristics provide reason to expect shifts in crude oil price trend. On the other hand there are periods (for example from 50 to 80 observations) when some combination of predictors enable models to generate forecast with limited errors (smaller than in the case of naive forecast).

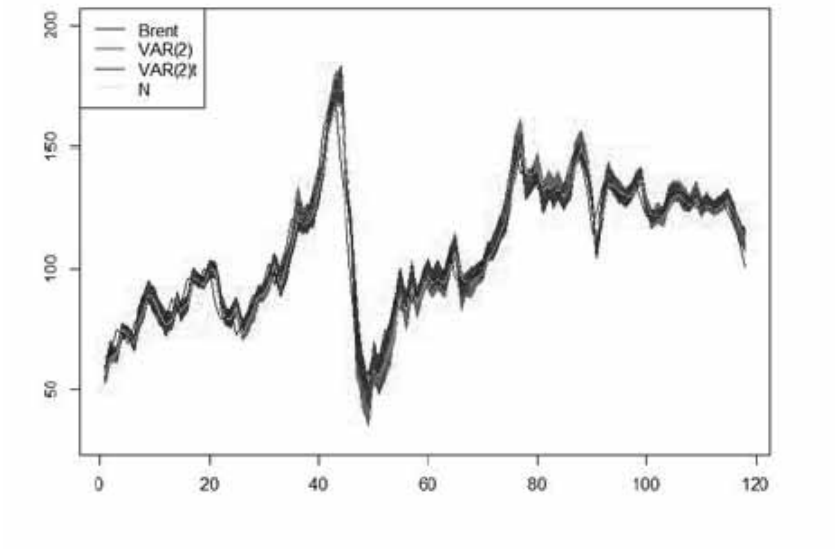


Figure 2: Forecast of crude oil price given by all possible combination of predictors from models VAR(2), VAR(2)t

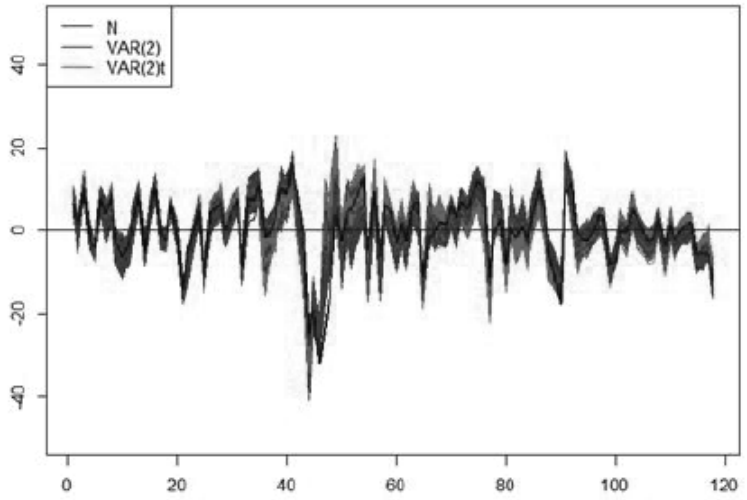


Figure 3: Prediction errors obtained from all possible combination of predictors from models VAR(2), VAR(2)t

3.4. The best predictors for Brent crude oil prices in different periods of forecasting

The results presented in the previous section reveled that it is impossible to identify a universal set of variables, which would lead to the systematically accurate crude oil forecast. In fact, different model specifications or sub-period, usually lead to different subsets of predictors which minimize forecast error. On the other hand, the difference between the error level obtained for a fraction of the best models i.e. VAR with two lags (due to a given criterion) are limited. This suggests that there are some series, which probably represent the same economic category and thus have similar paths. As consequence this variables contain the same source/set of information which provide similar forecasts. It is interesting to find out which variables occur most frequently in the best prediction models. To check this, the distribution of frequency of variables in the best 10% of forecasting models for crude oil are presented in Figure 4. The analysis are carried out in the three sub-periods: i.e. from January 2005 to December 2006, from January 2007 to December 2009, from January 2010 to October 2014. In this

way one is able to assess whether the same set of variables contribute to the improvement of forecasting oil prices in different periods.

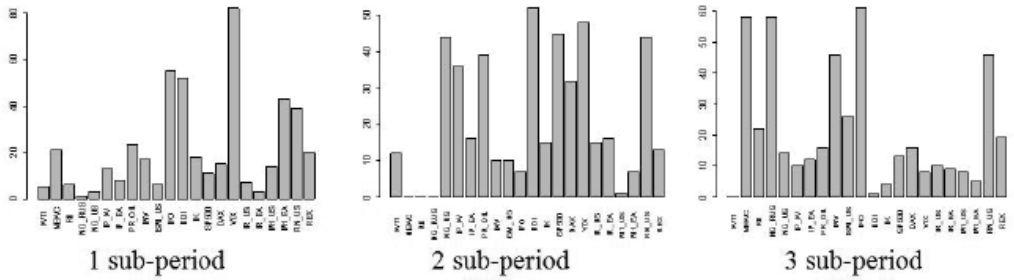


Figure 4: Frequency of predictors in 10 percent of best prediction models (VAR(2) with trend)

In the first sub-period (see Fig. 4, left), most frequently the following variables appeared: real M1 money supply in the euro area (i.e. M1_EA); the VIX index; variables describing market conditions, i.e. IFO and BDI. On the other hand, seldom in the best models appear variable describing: interest rates in the US and the euro area, the ISM manufacturing index in the U.S. and the prices of energy commodity prices (with the exception of the Australian coal prices NEWC). In the second sub-period (see Fig. 4, middle) accurate forecast are obtained using models containing: stock market indices (VIX, DAX, SP500), BDI, gas prices in the US, the real effective exchange rate of the dollar, global industrial production and global oil production. The third sub-period another set of predictors are used to obtain the most accurate forecast of crude oil. Then the most frequent variables include: prices of raw materials, in particular Russian gas and coal in Australia. In addition, in order to improve prediction accuracy characteristic of enterprise conditions in the German (IFO) and USA (ISM_US), the real exchange rate of the euro against the dollar and the real effective exchange rate of the dollar (RN_US and REX), as well as the global level of crude oil inventories (INV) should be include in the models. Further analysis focuses on analysing the relationship between the distribution of the frequency of the different predictors in the set of 10 percent best models and the three sub-periods of forecast sub-period. Figure 5 presents

the biplot, which shows the results of correspondence analysis, where the first category is the number of particular variables in the best models, and the second category are the three sub-periods of forecasts. The first sub-period is strongly associated with the real money supply Ml in the U.S. (M1_US) and in the euro area (M1_EA). This sub-period can be also connected with financial variables VIX and the business climate index in German (IFO). The second sub-period is associated with price of American gas and financial variables, i.e. the interest rate of the EU and the US stock market (SP500). In the third sub-period, contrary to previous sub-period, one can observe increase of importance of energy commodity prices, particularly coal and Russian gas. Two dollar exchange rate (i.e. Effective and dolar euro exchange rate) are the only variables that appeared equally often in the best prediction models in all the analyzed sub-periods.

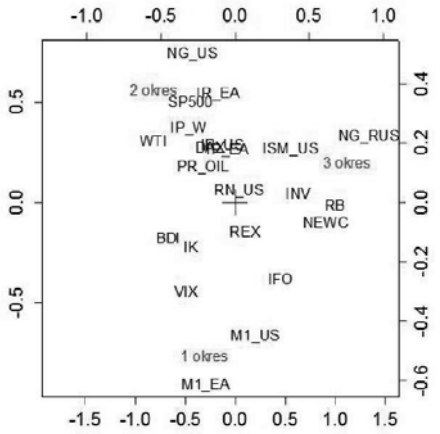


Figure 5: The results of correspondence analysis for frequency of predictors in 10 percent of the best prediction models and the forecast sub-periods

4. Conclusion

To answer the question formulated in the title of the article large (in comparison with other tests) data set was collected and different VAR specifications were investigated. As a result of the analysis two conclusions can be made. Firstly, an optimal specification for crude oil price forecasting seem to be VAR with two 2 lags. Forecast obtained by this model were more accurate in every sub-period of analysis. Second, the optimal set

of predictors used to forecast oil prices, was different in each sub-period. In the first sub-period the best forecasting models include variables describing the supply of money (especially in the euro area), economy activity indexes (IFO_BDI) and the VIX. In the second sub-period, the best prediction models include both variables characterizing economic activity (this can be connected with fast-growing economies of China and India at the time), the stock market (the reason for that is particularly strong financialization), and the American market of raw materials, especially natural gas prices (the beginning of the shale revolution). In the third sub-period the most accurate predictions could be obtained be using (as in the first sub-period) variable describing the economic activity in the euro area (IFO), and which is specific to this period, the prices of other energy resources and the world oil capacity. The common factor for most successive prediction models and for all sub-periods was the dollar exchange rate (RN_US and REX). The sensitivity of oil prices on the different factors for different time sub-periods make forecasting difficult. What is more, even if one succeeds in accurately identify a set of predictors, it will not offer any guarantee, that he could predict turning points in the trend in oil prices. If there are any information that can help predict trend changes in the price of oil there should be seek outside the economic factors considered in the paper. On the other hand, as results show, there are sub-periods (sometimes quite long), in which the VAR models are very competitive in comparison to naive forecasts. In this cases, the information contained in the prepared set of predictors are sufficient to effectively predict oil prices.
Use section and subsections to organize your document. Simply use the section and subsection buttons in the toolbar to create them, and we'll handle all the formatting and numbering automatically.

Acknowledgements

Supported by the grant No. 2012/07/B/HS4/00700 of the Polish National Science Centre.

References

[1] Akram, Q.F. (2009). Commodity Prices, Interest Rates and The Dollar. *Energy Economics*, 31(6), 838-851.

[2] Alquist, R., Kilian, L., & Vigfusson, R.J. (2013). Forecasting the Price of Oil. [in:] G. Elliott & A. Timmermann (Eds.), *Handbook of Economic Forecasting*, 2, Amsterdam: North-Holland.

[3] Alquist, R., & Kilian, L. (2010). What Do We Learn from the Price of Crude Oil Futures? *Journal of Applied Econometrics*, 25(4), 539-573.

[4] Anzuini, A., Lombardi, M.J., & Pagano, P. (2013). The Impact of Monetary Policy Shocks on Commodity Prices. *International Journal of Central Banking*, 9(3), 125-150.

[5] Baumeister, C., & Kilian, L. (2012). Real-time Forecasts of the Real Price of Oil. *Journal of Business and Economic Statistics*, 30(2), 326-336.

[6] Baumeister, C., & Kilian, L. (2014). What Central Bankers Need to Know About Forecasting Oil Prices. *International Economic Review*, 55(3), 869-889.

[7] Baumeister, C., Kilian, L., & Zhou, X. (2013). *Are Product Spreads Useful for Forecasting? An Empirical Evaluation of the Verleger Hypothesis (No. 2013-25)*. Bank of Canada Working Paper.

[8] Bernard, J.-T., Khalaf, L., Kichian, M., & Yelou, C. (2013). *On the Long-term Dynamics of Oil Prices: Learning from Combination Forecasts*. Mimeo: Carleton University.

[9] Chen, Y.C., Rogoff, K.S., & Rossi, B. (2010). Can Exchange Rates Forecast Commodity Prices? *Quarterly Journal of Economics*, 125(3), 1145-1194.

[10] Davies, P. (2007). *What's the Value of an Energy Economist?* Speech Presented at the International Association for Energy Economics, Wellington, New Zealand.

[11] Gargano, A., & Timmermann, A. (2014). Forecasting Commodity Price Indexes Using Macroeconomic and Financial Predictors. *International Journal of Forecasting*, 30(3), 825-843.

[12] Hamilton, J.D. (2008). *Understanding Crude Oil Prices (No. w14492)*. National Bureau of Economic Research.

[13] Hamilton, J.D., & Herrera, A.M. (2004). Comment: Oil Shocks and Aggregate Macroeconomic Behavior: The Role of Monetary Policy. *Journal of Money, Credit and Banking*, 36(2), 265-286.

[14] Issler, J.V., Rodrigues, C., & Burjack, R. (2014). Using Common Features to Understand the Behavior of Metal-commodity Prices and

Forecast Them at Different Horizons. *Journal of International Money and Finance*, 42, 310-335.

[15] Kilian, L. (2009). Not All Oil Price Shocks Are Alike: Disentangling Demand and Supply Shocks in the Crude Oil Market. *The American Economic Review*, 99(3), 1053-1069.

[16] Kilian, L., & Hicks, B. (2013). Did Unexpectedly Strong Economic Growth Cause the Oil Price Shock of 2003-2008? *Journal of Forecasting*, 32(5), 385-394.

[17] Kilian, L., & Murphy, D.P. (2014). The Role of Inventories and Speculative Trading in the Global Market for Crude Oil. *Journal of Applied Econometrics*, 29(3), 454-478.

[18] Kilian, L., & Vega, C. (2011). Do Energy Prices Respond to US Macroeconomic News? A Test of the Hypothesis of Predetermined Energy Prices. *Review of Economics and Statistics*, 93(2), 660-671.

[19] Knetsch, T.A. (2007). Forecasting the Price of Crude Oil via Convenience Yield Predictions. *Journal of Forecasting*, 26(7), 527-549.

[20] Reeve, T.A., & Vigfusson, R. (2011). *Evaluating the Forecasting Performance of Commodity Futures Prices. FRB International Finance Discussion Paper, (1025).* Board of Governors of the Federal Reserve System.

[21] Sanders, D.R., Manfredo, M.R., & Boris, K. (2009). Evaluating Information in Multiple Horizon Forecasts: The DOE's Energy Price Forecasts. *Energy Economics*, 31(2), 189-196.

[22] Schalck, C., & Chenavaz, R. (2015). Oil Commodity Returns and Macroeconomic Factors: A Time-Varying Approach. *Research in International Business and Finance*, 33, 290-303.

The Application of the Matrix Flow Diagram for the Errors in a Quality Control in Production Area

Angelina Rajda-Tasior

University of Economics in Katowice, Poland

Abstract

The paper presents a simply method for analyze errors occurred and detected in variety production processes. The new approach is to use of describing tool for verification of proper production and control process a specially quality control because the main role of quality control is to confirm products characteristics with requirements.

The proposed solution can give an information about an economic loss of quality, it indicates the relationships and dependencies between the processes where errors can occur and where can be detected. It indicates how much cost the untimely error. It also provides a clear and consistent image of errors placement. The matrix flow errors specifies processes where the non-compliance provided that the data for the analysis are reliable.

Keywords: quality control, errors pacement, production

1. Introduction

The issue of manufacturing high quality products concern on a proper measuring process and are linked to the expressed or unexpressed expectations of the buyers. It is very important to use correct parameters for the designed product. There is therefore a need for active impact on the quality. It requires a solution to the problem in the area of quality control. Quality control is a process of ensuring that a manufactured product fulfils quality criteria and meets requirements of producer and consumer. Products that don't comply with the specifications are rejected or returned to improve. Quality control includes monitoring processes (actions) and

eliminating the causes of errors at all stages of the product life cycle. Next to traditional quality management techniques like Ishikawa, Pareto-Lorenz diagrams, there are matrix diagrams.

Matrix diagrams, enable to specify relationships and dependencies between different quality characteristics which concern the value offered to the customer. This can be for example the relationship between the causes and consequences of inadequate quality.

2. Defectiveness a incompatibility. Characteristic of errors

Defectiveness refers to product life cycle. In the process of creating the value of product some errors may occur (Tab. 1). The arrangement of the errors in the process determine their size. It is therefore necessary to identify the location of faults and measure their value to change their location on more favorable to be able to capture in the process control. The term **Defective product** not only covers the production stage, but goes far beyond the company. Defectiveness of the product also includes its impact on the environment and the generation and degradation in accordance with the principle of respect for the earth's resources. Defective product in an innovative approach may therefore be incompatible with the requirements of in various stages of development and include marketing, development, design and construction defects as well as information and management defects.

From a statistical point of view, assessment of the product or its part, is often expressed by indicating the nonconformities or by classifying the product as a defective. Non-compliance is a condition of feature, which means failure to comply with the requirements.

Types of errors	Characteristics of errors
Marketing defects	Relate to the incompleteness of specified customer requirements, shortage of appropriate methods for testing its satisfaction [4]. Related to improper recognition of the features of a competitive product.
Development defects	These errors refers to the risk of progress [6]. Relate to the some characteristics of the product which are comply with the requirements at time t but become defective over time t + 1 as a result of the development of science and technology (materials with asbestos, mobile phones).

Design defects	Construction errors. Refers to the sample product. In accordance with the principles adopted in the design stage, product should fulfill the functions for which it is intended and not pose a danger to the environment. Otherwise, errors may occur in area of reliability, product handling, the possibility of use the product in extreme conditions (car disaster).
Production defects	The progress of science and technology allowed for mechanization and automation of manufacturing stages of the product, thus reducing the probability of failure. Also, the management of a defective product was launched just in manufacturing processes for this reason that each copy of the product can be compared with the standard – sample product. Production errors concern on technological processes as well as storage, transport and quality control.
Information defects	Associated with the label information and warnings against possible dangers of using the product. These include uncommunicative information, unclear, non-transparentand incomprehensible instructions included in the product or its documentation [7].
Management defects	Related to accuracy and delay of the decision-making process.

Table 1: Types and characteristics of errors

3. Model flow for the errors

Producing process requires a proper measuring all processes. Each process consists of actions (Fig. 1), which were separated in the following processes.

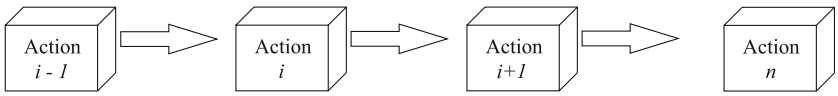


Figure 1: Activities at the sequence process

For example for one action in the process error can be permissible (defective parts) because components were not detected and in the other action errors can be created (faulty parts). Both of the actions are not always detected and can flow to the process. Among these actions can be found actions of the nonconformities. This is a close to approach Taiichi Ohno [1], who developed the model of seven types of actions which do not create value for the customer. These are the actions related to overproduction,

the expectation (for example delayed delivery), the unnecessary transport, redundant storage, improper preparation of production **and non-compliances regarding to the quality requirements (errors).**

3.1. Matrix flow diagram for the errors

Let rows of the matrix X determine the number of errors occurring in the action *i*, and detected both in operation and as well as in further (*i +1, i + 2, ..., n*). Columns of the matrix determine the number of errors occurring in previous actions (*i - 1, i - 2, ..., 1*) and in action *i*, which were detected in action *i*. Matrix X with dimensions (*n - 1*)*n*:

$$X = [x_{ij}] \tag{1}$$

i \ *j*	1	2	3	4	5	...	*n*
1	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	...	x_{1n}
2	0	x_{22}	x_{23}	x_{24}	x_{25}	...	x_{2n}
3	0	0	x_{33}	x_{34}	x_{35}	...	x_{3n}
4	0	0	0	x_{44}	x_{45}	...	x_{4n}
5	0	0	0	0	x_{55}	...	x_{5n}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n-1	0	0	0	0	0		*x(n - 1)n*

Table 2: Matrix flow errors for production processes

where:
x – number of errors,
i – actions, where errors occurred,
j – actions, where errors detected.

3.2. The measures of the flow errors

Based on the proposed matrix flow errors is possible to calculate a set of parameters that define the relationships between different measures of their flow. According to the following picture determined the parameters by formulas (2-12).

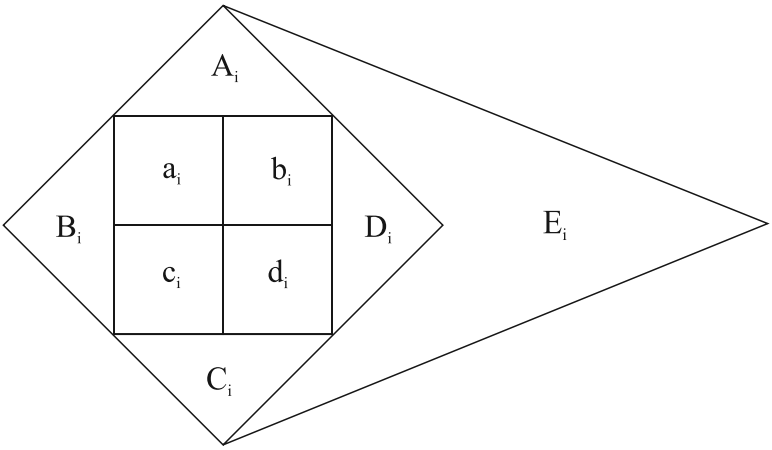


Figure 2: Synthetic figure of the matrix diagram types referred to the measures of defect flow

$$W_1 = \frac{a_i}{A_i} \tag{2}$$

$$W_2 = \frac{b_i}{A_i} \tag{3}$$

$$W_3 = \frac{c_i}{C_i} \tag{4}$$

$$W_4 = \frac{d_i}{C_i} \tag{5}$$

$$W_5 = \frac{a_i}{B_i} \tag{6}$$

$$W_6 = \frac{c_i}{B_i} \tag{7}$$

$$W_7 = \frac{b_i}{D_i} \tag{8}$$

$$W_8 = \frac{d_i}{D_i} \tag{9}$$

$$W_9 = \frac{A_i}{E_i} \tag{10}$$

$$W_{10} = \frac{c_i}{E_i} \tag{11}$$

$$W_{11} = \frac{B_i}{E_i} \tag{12}$$

where:
a_i – errors occurred and detected in action i,
b_i – errors occurred in action i, and detected in following actions ($i+1$, $i+2$, ..., n),
c_i – errors occurred in previous actions ($i-1$, $i-2$, ..., 1), and detected in action i,
d_i – errors occurred in previous actions (i-1, i-2, actions (i+1, i+2,..., n),
A_i – errors occurred in action i,
B_i – errors detected in action i,
C_i – errors moved to the action i,
D_i – errors not detected in action i,
E_i – errors located in action i during time t.

4. Verification of the proposed solution

To verify the solution by the matrix flow model for errors for furniture production were conducted. The relevant process concerned on production as sofas, corners, armchairs, beds and mattresses. In production process specified several actions which refer to the process of innovation, operational and support after sales. The information come from the quality control and complaints department. For the design of the array of matrix the diagram type L was used. It takes into account the relationship of two sets of quality characteristics: 9 actions where errors can occur i and 13 actions where errors can be detected j.

Actions i, where errors can occur:
1. product development section (RW),
2. calibration section (W),
3. carpenter section (S),
4. assembly section (M),

5. swing section (SZ),
6. upholstery section (T),
7. packing section (P),
8. warehouse section (MA),
9. transport section (TR).

Actions j, where errors can be detected: 1-9 actions and also:

10. quality control (upholstery section) (KJ),
11. the use in the customer area (KL),
12. monthly sampling stock control (KJ M),
13. sales section (H).

Control charts and complaints were used to isolate errors that were occurred in concrete actions and places where errors were detected and quantity. Analysis of errors related to one month and is presented in Table 3.

	RW	W	S	M	SZ	T	P	MA	TR	KJ	KJM	H	KL	Defects occurred
RW	-	2	-	-	-	-	4	-	2	1	-	-	1	10
W	-	-	5	4	6	8	2	-	-	1	-	-	2	28
S	-	-	-	5	-	2	-	-	-	3	2	-	5	17
M											3	-	2	5
SZ	-	-	-	-	-	70	-	-	-	25	35	-	15	145
T						12	4	-	-	70	40	-	50	176
P								13	27	-	15	50	41	146
MA									5	-	2	30	-	37
TR									3	-	-	99	66	168
Defects detected	-	2	5	9	6	92	10	13	37	100	97	179	182	732

Table 3: Matrix flow errors for production processes

Diagram analysis allows to determine activities which involve the identification of the causes of non-compliances, necessary corrective actions and to plan prevent actions. Based on the analysis of the matrix flow errors have been marked actions in which the error occur the most (Tab. 3). 70% of defects (errors) occur in **upholstering and transport**. Errors which occur on the upholstery section were detected in operations as: upholstery, packaging, quality control one hundred percent, quality control sampling and the use of the client. To illustrate the scale of all defects diagram Pareto-Lorenz was constructed. Another diagram presents that 60% of all

the defects which occur in the transport relates to the holes in the fabric on the furniture and damaged wood.

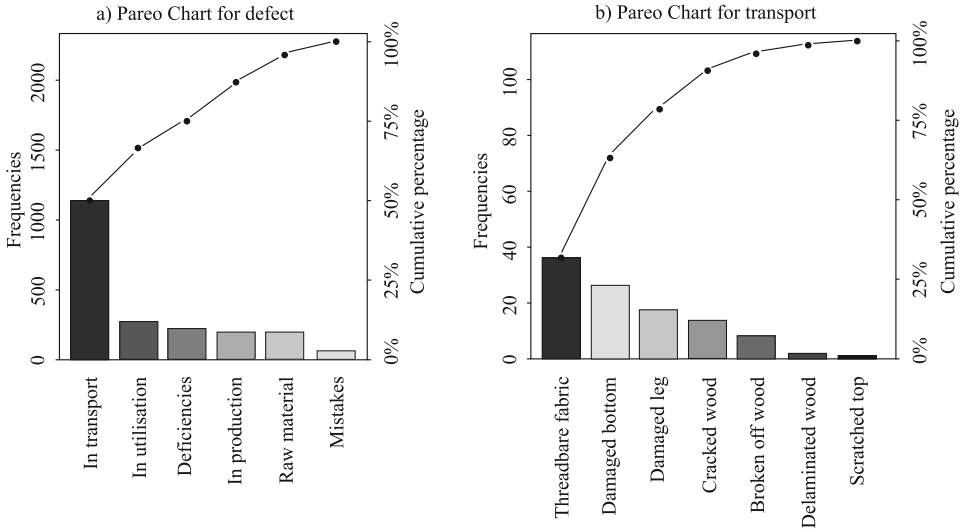


Figure 3: Pareto-Lorenz diagram (a – types of defects, b – faults in transport)

In Table 4 specified actions where were detected the most errors. During storage at the warehouse and customers using is reflected most defects. In quality control process many of the errors flow into the next actions (KJ). This is a disturbing situation and should be checked in order to improve.

	RW	W	S	M	SZ	T	P	MA	TR	KJ	KJM	H	KL	Defects occurred
RW	-	2	-	-	-	-	4	-	2	1	-	-	1	10
W	-	-	5	4	6	8	2	-	-	1	-	-	2	28
S	-	-	-	5	-	2	-	-	-	3	2	-	5	17
M											3	-	2	5
SZ	-	-	-	-	-	70	-	-	-	25	35	-	15	145
T						**12**	**4**	**-**	**-**	**70**	**40**	**-**	**50**	**176**
P								13	27	-	15	50	41	146
MA									5	-	2	30	-	37
TR									**3**	**-**	**-**	**99**	**66**	**168**
Defects detected	-	2	5	9	6	92	10	13	37	100	97	179	182	732

Table 4: Matrix flow errors for production processes

In order to monitoring the monthly level of defect Shewhart u-control chart has been applied (Fig. 4). The chart was based on the statistics where the average number of defects is the ratio of the number of products sold and the number of complaints in a specified month.

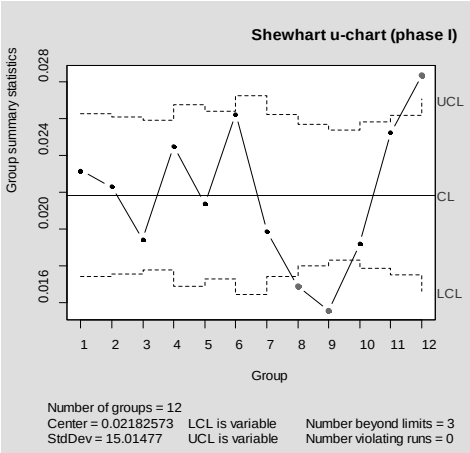


Figure 4: Shewant u-quality control chart

Control lines for each month were determined separately. For August and September marked deflection parameters below LCL (lower control limit). It actually means an improvement of production parameters and in this case the lines are not regulation lines regulation, in contrast to December. However, should be looked for the causes that led the improvement results to obtain the stable process.

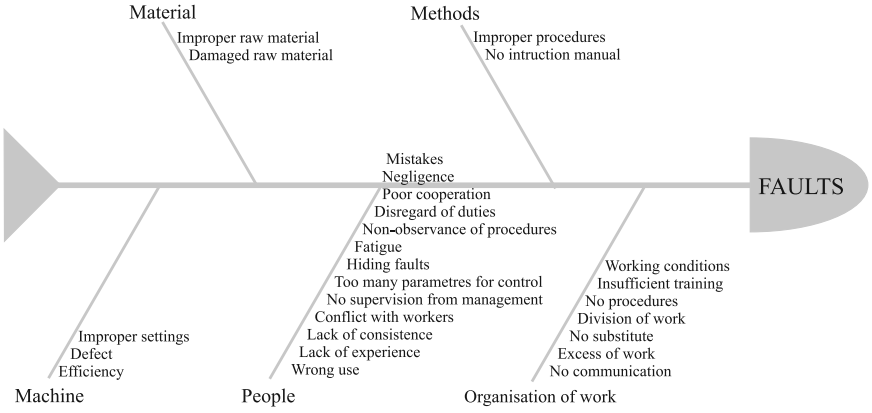


Figure 5: Ishikawa diagram

5. Conclusion

The proposed solution in the form of a matrix approach to losing quality in the process has many advantages. First organizes information about an economic loss of quality, and thus about the costs (for example: human, machine, material). Secondly it indicates the relationships and dependencies between the actions where errors occur and the actions where they are detected. Next it indicates how much cost when untimely error is detected. It also provides a clear and consistent image of errors placement in the actions relating to the production process and allows graphic shot of innovative, operational and customer service actions if they will be designated. The analysis shows the solution which can transform the complaint process to the improvement tool. Because the client is an absolute judge who evaluated the quality of the product, it was considered complaints as a good tool to analyze errors (defects, unconformities). In this study assumed that the company is a set of actions where errors occur and where they can be detected. These actions are always associated with deviations from quality requirements (errors). Therefore it proposed solutions refers to the errors associated with the product development cycle and product life time. That kind approach to the problem of occurring and detecting errors requires to use of an appropriate tool as a matrix model flow for errors.

References

[1] Drummond, H. (1998). *W pogoni za jakością,* Warszawa: Dom Wydawniczy ABC.
[2] Hamrol, A., Zymonik, Z., & Grudowski, P. (2013). Zarządzanie jakością i bezpieczeństwem, Warszawa: PWE.
[3] Iwasiewicz, A. (1985). *Statystyczna kontrola jakości w toku produkcji.* Warszawa: PWN.
[4] Kaplan, R.S., & Norton, D.P. (2001). *Strategiczna karta wyników. Jak przełożyć strategię na działania*, Warszawa: PWN.
[5] Kończak, G. (2007). *Metody statystyczne w sterowaniu jakością produkcji.* Wydawnictwo AE w Katowicach.
[6] Von Marschall, W. (1988). Z zagadnień odpowiedzialności za produkt. *Państwo i Prawo*, (3).
[7] Wasilewski, L. (1994). *Modele strategii jakości firm przemysłowych.* Warszawa: Instytut Organizacji i Zarządzania w Przemyśle „ORGMASZ".

[8] Zymonik, Z. (2013). Koszty jakości w zarządzaniu przedsiębiorstwem. Wrocław: Oficyna Wydawnicza Politechniki Wrocławskiej.

Human Capital Management in Services Providing Enterprise

Weronika Toszewska-Czerniej

Koszalin University of Technology, Poland

Abstract

The primary objective of human capital management is to help create a management process that will maximize the impact of employees on the company ability to achieve goals. The advantage of human capital management (HCM) is the implication of the process can not only improve the use of the competence of employees, but also productivity and profitability. The main purpose of this article is to present the results of research on human capital of the service delivery enterprise. The participants of the study were employees and managers of enterprise units located in Koszalin region. The study includes evaluation of the existing human capital management process highlighted as crucial in the services delivery process. Results obtained from the study shows significant inconsistency of views and perceptions of the capital employed in the company in the studied region. As a conclusion based on collected data an attempt to create the outline of the process was made, which can let for increasing the effectiveness of human capital management linking the process of the service delivery with the process of exploiting the potential of employees.

Keywords: human capital management (HCM), service delivering process, research on human capital

1. Introduction

The human capital theory helps to better understand the role of employees in the enterprise. Human potential is gaining great significance in adding the value to delivering services. Enterprise approach to human capital

management determines how creating value through the efficient creation and use of employees contributes to business performance [8] (p. 201).

The main objective of human capital management (HCM) is to understand the employees impact on the business and their ability to achieve goals. The advantage of human capital management is the implication that the process can not only improve the use of the competence of employees, but also productivity and profitability.

The aim of this paper is to present the opinion of employees and managers on chosen issues relating to HCM process. Results of the study shown in this paper includes the evaluation of the existing human capital management process highlighted as crucial in the services delivery process. The main purpose of this paper is to present that in every organization process in which is involved a man must be a revaluation consisting of understanding and using the value that is put on the effects of the activities of each human individual. Human capital determines the performance of the company becoming a key resource organization.

2. Human capital management process

This provision of human capital management process need to be paired with the strategy and vision of the organization. The key objective is to determine the impact of employees on the functioning and the ability to achieve the objectives of the organization. It is therefore necessary strong relationship between the processes of the organization. The essence of human capital management is to acquire, analyze and report data to inform directorates of value-adding decisions in the field of people management [1] (p. 301).

Process (HCM) must be preceded by the identification of needs and potential of the business. They are determined by internal and external factors. The ability of an organization to adapt to certain conditions, determine its level of flexibility. Flexibility HCM process involves the formation procedures and actions aimed at employees that will enable the organization to adapt to changes in human capital, environment and internal conditions [3] (p. 55). HCM therefore requires an understanding that business objectives can be achieved only through effective use of resources. Crucial importance in this case have applied processes aimed at exploiting the human potential.

One of the levels of human capital management is the level of recognition opportunities and threats. It helps to recognize the main points of knowledge of workers [3] (p. 78). The basic premise of modern HCM is the recognition of employees as the most valuable parts of the company [4] (p. 24). The company should implement a system to verify employees opinion. It allows to obtain useful information about the social mood, level of acceptance operations units and matching them to the needs of employees. However, first of all it gives the possibility to verify HCM process implemented. An employee is an essential unit of affecting the ability of businesses to adapt to existing conditions. Verify its opinion on selected issues concerning the HCM process is needed. To get the full picture the article also included opinion of superiors of the workers.

Considered aspects have been selected in a way as to allow revision of the three dimensions of human capital management, namely enterprise, service process and workplace. Taking into account different scales is possible to obtain an assessment of individual levels of activities that due to the scope adopt a different character.

The effective use of human capital is dependent on properly formulated and carried out management process. Appropriately implemented HCM process allows apply human capital to achieve a sustainable competitive advantage by expanding the value of employees [5] (p. 349). The purpose of human capital management in this case is to determine the effect of employees in the service activities of the company.

Properly implemented and improved business processes allows effectively operate in highly competitive market. First of all, because they provide a greater ability to create value for customers [9] (p. 92). Included in study questions are constructed to indicate the efficiency of human capital management process. Effectiveness means doing the right things. It is related to how well the company understands reacts and affects the environment and their potential [10] (p. 15). Workers evaluates effectiveness of the action through opinion on the following factors:

- knowledge of corporate goals,
- provided opportunities for development,
- working conditions,
- requirements for employees,
- loyalty.

3. Respondents characteristics

All participants were workers of the service delivering enterprise from Koszalin region, which included area of former Koszalin province and tree municipalities from province Słupsk. A scope of the relation with the customer constituted selection criteria. Workplaces which the highest level of the customer contact were chosen.

Data were collected using questionnaires, separate for each group of respondents. Participants assessed the HCM process in the five-point scale. Results obtained from the study show inconsistency of views and perceptions of the capital employed in the company in the studied region. Evaluation of the managers is identified with the HCM strategy adopted by the enterprise. Managers from the region formed a distinct group of 20 randomly selected respondents who evaluated the identical issues of HCM. The implementation of the business strategy directly depends on the activities of employees. The evaluation made by the workers shows the point of view of owners of capital.

Employees were divided according to the category of their place of employment, age, sex, education and the work experience. The total sample size was 628 units. They are individuals which are investing their capital in the process of the service delivery, as well as can be assessed by the client. The division of participants in examining according to exchanged criteria was put in Table 1.

Criterion \ Category of unit		Urban area	Country area
		Quantitative data	Quantitative data
sex	women	231	29
	men	208	34
education level	primary	10	2
	vocational	74	17
	average	304	37
	higher	51	7
work experience	up to year	16	0
	1-5	88	19
	6-10	62	8
	11-15	80	5
	16-20	68	9
	above 20	125	22

Table 1: Characteristics of respondents

Conclusions from the study are based on the demonstration of an entities divergence. Disagreements between the entities that implement activities in the HCM field and the owners of human capital affects the realization of the process. The result is a diverse distribution of the evaluation of self-assessment component. Actual values should be adapted to the needs and requirements of the company.

4. Characteristic of the research

Simplifying it can be stated that the purpose of human capital management is proving the value of people and creating added value by using they potential. The level of the value created by workers is conditioned by the balance between the needs of the recipient and the resources provided by the company. The ability of employees to create value-added process verifies HCM, whose effectiveness is reflected in the level of human capital adjustment to the conditions.

Presenting the results will be preceded by a description of issues that concerned the questionnaire. The scale of the questions, identical for all, adopt the greatest value for the vast consent, equals five, ending on the vast discrepancies equals one. Table 2 includes questions about the assessment HCM process by company employees.

Level of HCM process	Verified area	Question number	Justification
Employee	The level of development opportunities offered to employee.	**8**	The possibility of the development of knowledge and skills is an indicator of the level of investment in human capital. HCM include investing in human capital.
	The level of satisfaction with career paths that are created for each employee.	**9**	The ability to adjust directions of development of the workers is a measure of adequacy of adopted HCM process. Investing in human capital is more effective when is well adapted to needs.

Service delivering process	Understanding the level of requirements posed to the employee in the services delivery process	**10 a**	Can carried out HCM process makes the employee knows what to do achieve effective use of their capital. Is HCM able to meet the requirements of the position.
	The expected level of skills necessary to implement the services provision process.	**10b**	Is ensured action in the field of HCM allows to provide services at the highest level. Is the process providing employees with adequate support and sufficient skills to provide services.
	Assessment of the possibility of obtaining remuneration for the capital employed in process	**10c**	Whether the employee is aware of dependencies and the impact of capital to revenue growth from the use of human capital.
Enterprise	Perceived level of employment stability of human capital owner	**11 a**	Evaluation of the effects of the process, does it affect the proper attitude to work. HCM process should create commitment and devotion to the individual in the context of relationships.
	Declared level of loyalty to the company.	**11b**	Evaluation of the effects of the HCM process, does it create a high level of loyalty. Full commitment to employees depends on the actions of forming an atmosphere of trust, honesty and mutual respect.
	The level of acceptance of the actions undertaken in the area of creating corporate culture	**11c**	Evaluation of the effects of the HCM process, does it affects the culture of the organization. Culture is a collection of shared norms and values serves to create an individual identity and determines the process of communication, teamwork and activities in field of knowledge sharing.
	Acquaintance and understanding of the strategic objectives of the enterprise.	**1**	HCM process should be consistent with the strategy of the company. The employees to pursued it must be familiar with the main objectives of the company.

Table 2: Description of HCM levels

In the order of proper used human capital should be matched to the strategy and structure of the company and to satisfy market requirements. Therefore, the study divided organizational units due to the area on which they operate. Given the assumptions HCM striving for optimal adaptation process to the potential of employees is crucial. For this reason, the division was made into rural and urban areas assuming the terms of service varied depending on the location.The main task for management is to create a unique present and future architecture of components that will be created the core values of the future [14] (p. 164). The realization of this goal depends determine which companies are the most important ingredients for success now and in the future.

5. HCM process analysis

To determine the directions of improvement of human capital is essential to show that the verified employment issues include the factors shaping and determining the development of the employee and his actions. The degree of level of services is conditioned by the ability of the employee to provide it. The potential unit determines the management process and the actual level considered the building blocks of its capital. The studies deal with three levels of effectiveness, which together determine the effectiveness of the entire company, taking into account the processes and the capital of the enterprise on a holistic approach to efficiency [12] (p. 66).

Table 3 includes elements of effectiveness dealt with at the level of organization, process, service, and jobs. The effectiveness of human capital management process depends on optimal adjustment factors influencing the results achieved at different levels of human capital of the company.

A primary consideration is to set targets for testing the effectiveness of the company and developing states of desired results for the implemented processes. First of all objectives should be defined through standards arising with customer expectations. Defining objectives without considering the specificity of the process, the needs and abilities of employees will not be effective. This confirms the necessity of considering the views of several groups of respondents thus extending the reliability of the data obtained. Regardless of the precise definition of goals, properly designed process can't expect optimal results business without effective management. By defini-

tion it is a set of actions aimed at corporate resources carried out with the intention of achieving the objectives [12] (p. 69).

Level of process	Factors affecting level	Objective of the measure	Verification of the results achieved	process needs
Enterprise	Usage of human capital, HCM strategy, company structure.	The aims of the strategy ZKL determine the optimal use of human capital	Are the objectives are clearly defined, understood by employees? Is trends were communicated to employees?	Proper management, efficient procedure for use of human capital and relationship with the other dimensions of the organization.
The process of provision of services	The process of the service	The objectives of the process must be directly related to internal and external customers. Should follow with organizations goals and customer expectations.	he process allows maximum efficiency of capital involved in its realization.	Verification of the allocation of capital adequacy to the objective of the process, control of the adequacy of the specific results of each stage of the process.
workplace	Responsibilities applied labor standards, training system.	Achieving the optimal level depends on the objectives of the position and the ability of the employee.	The term impact of the effects of the process, verifying whether the work results meet the requirements of the recipient..	Creating a system of development potential of employees, taking into account the level of results from adequate development, motivation, talent and working conditions.

Table 3: HCM performance levels

Taking into account the contemporary approach to the study of entrepreneurial activity the most important goal of human capital management is to ensure long-term development of employees [13] (p. 11). Bringing the human capital to increase trough proper processes is a main factor to improve value level of services. Human capital management should apply

to financial and non-financial performance measures, particularly focusing on the following areas:

- applied leadership development programs,
- the involvement of employees,
- level of access to knowledge,
- optimization workplace
- the scope of teaching, the boundaries of perception [7] (p. 300).

Included in the study service company in order to increase its competitiveness should create a HCM process that will take into account the abovementioned key attributes of the organization.

6. Analysis of urban units employees opinion

Results of the analysis will begin with a presentation of average values for each of the questions received by the employees of the company deposited in urban areas. Presented in Figure 1. data applies to the total number of units. They indicate the level of acceptance representatives of the position statements contained in the questionnaire.

The highest level of fairness has obtained a statement regarding the level of knowledge of the rights and obligations arising from work positions. Knowledge of the requirements affects their ability to fulfill, as evidenced by the significantly lower level of response to the question about delivery of services.

It should be emphasized that the management in described enterprise is centralized, which in limited scope depends on the managers of individual units. The low level of satisfaction with the way management is evident premise mismatching of human capital process to which it relates. Inflexible activities in the field of creating organizational culture, creating a stable working environment adversely affect the image processes carried out across the enterprise. Matching this area of activities by the managers with employees individuals needs and opportunities would provide the opportunity to better reflect the actual conditions of the organization. The Board must understand that creating the right atmosphere will guarantee not only achieving goals, but also conducive to achieving better results and motivate employees to work efficiently [11] (p. 49).

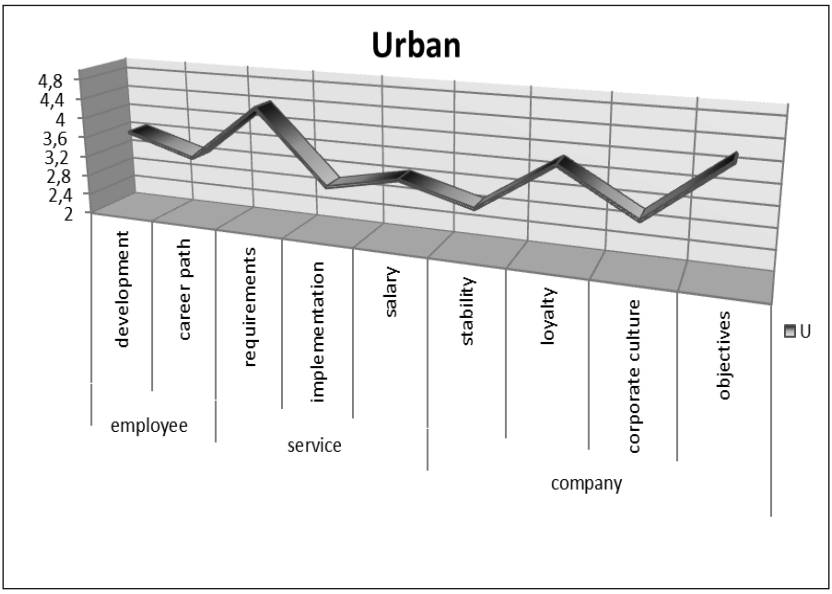


Figure 1: Results of the urban units

Noticeable for workers in this area is also a big disparity between the level of confidence in a stable position of the company providing favorable conditions of employment. It can be assumed that partly responsible for this state of the market situation, however skillful building relationships with employees should be allowed to maintain more consistent evaluations.

Presented division shows that none of the levels of the evaluation process HCM does not have integrity. The weakest area according to workers are activities to ensure the stability of employment.

7. Analysis of rural units employees opinion

Another discussed category are organizational units located in rural areas. A characteristic feature of opinion in this group of employees is a significant disproportion between the different answers.

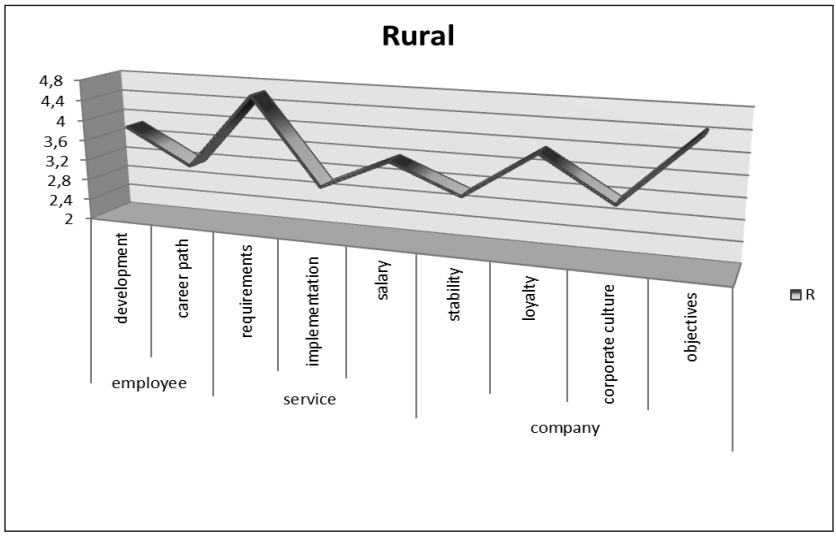


Figure 2: Results of the rural units

The scope of knowledge of the requirements by the employees received a high evaluation of reference, however, the capacity to implement the process of service received the lowest score among all verified in the study. In part this can be read as the result of the activities carried out at the workplace and therefore to provide opportunities for growth and the creation of career paths that have received low scores.

Again, employees of rural areas low evaluated the activity of companies in the area of creating organizational culture and building a sense of stability of employment. The results indicate insufficient compatibility of HCM process to the internal conditions of the organization. Workers are characterized not satisfied with the conditions to develop their own potential and direction of actions taken by the company. Mismatch HCM process to capital which it relates prevents its effectiveness, thus making the ability to obtain flexibility unreal.

8. Comparison of analyses

The estimates presented employee should be complemented by the opinions of individuals who are responsible for implementation of the human capital management process. Managers selected units assessed the same elements as workers thereby subjecting the assessment process designed by the highest executives in implementation of which they participate. Figure 3

takes into account the list of values for the two groups of respondents, taking the opinions of employees with the assessment made by the managers of organizational units.

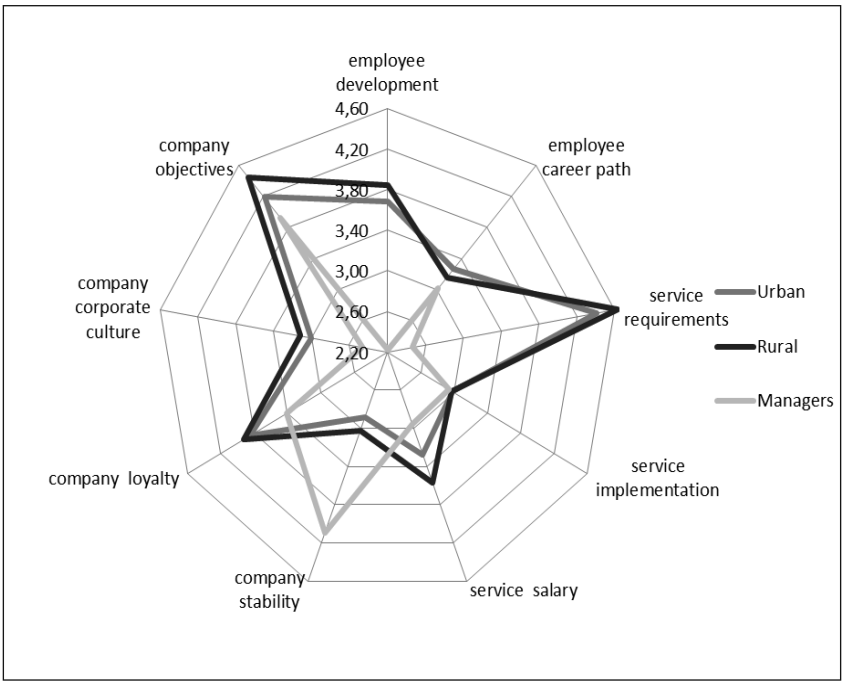


Figure 3: Combined opinions

The conclusion posted on the basis of the presented results is the need for transformation process of human capital management company. Despite knowledge of the requirements, tasks and goals employee does not achieve maximum results. Opportunities for development and the way a company manages employees is not consistent with the needs of employees. Without a doubt, it affects negatively on the performance of the employee's duties. The company should build its competitive advantages based on intangible assets, making them a unique bundle of value for customers [2] (p. 36). Companies are not able to compete effectively using only enough in this process is essential tangible commitment in the process of intangible assets of the organization [6] (p. 371).

What is very unexpected that the managers have so different opinion about ability to fulfill service requirements. Second aspect with low cohesion is opinion about employees development possibility. Also quite interesting

is fact that the opinions about stability of organization is also separate for each group of respondents. Such separate results with no doubt attest to the large inconsistencies in the implementation and needs of HCM process.

9. Conclusion

At high formalization process and with not advanced technologically process of providing services human factor is a very important element. A well designed process aimed at maximizing the human potential will be conditioned its results. Figure 4 recognizes scheme, taking into account the activities of joining the service area of personnel processes in the objectives of optimizing the results of activities of the company. The main objectives of the efficiency of the proposed scheme is reliable and systematic measurement and of well-functioning process of information flow.

Through shown opinion of employees and managers company can design a process for more effective services portfolio management and exploit existing employees potential. In the opinion of employees HCM process is not particularly fit to the embodied their potential, which does not allow its appropriate use. It shows that opinions aren't compatible enough to create a HCM process which is able to create high level of involvement. This is a precondition for this conception, which without this action taken will not be effective.

Obtaining substantial disparities in the results workers and managers does not show on properly designed and implemented HCM process. as a result, the company has far less chance of ensuring a high level of service to customers. You cannot expect optimal results in a situation in which evaluation of management process assumes that different results. The first step therefore should be striving to achieve a convergence of views, and further action should produce the best possible results of action.

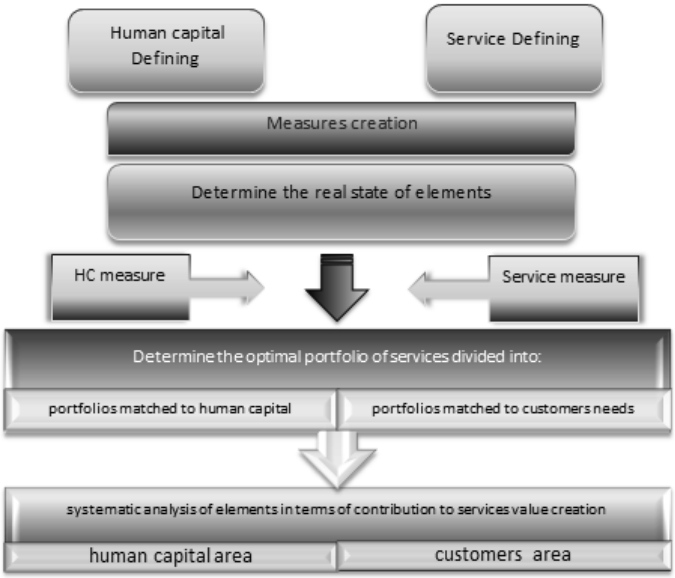


Figure 4: Implementation of services based on human capital

The communication aims of the research was to present the opinion of the service providing enterprise employees on HCM. The results indicate a modest acceptance of the actions taken by the company, and inadequate management process to fit the accumulated potential. This is a stimulus for further transformations aimed at optimizing capital utilization of employees.

References

[1] Jamka, B. (2011). *Czynnik ludzki we współczesnym przedsiębiorstwie: zasób czy kapitał? Od zarządzania kompetencjami do zarządzania różnorodnością*. Warszawa: Oficyna Wolters Kluwer business.

[2] Jarecki, W., Kunasz, M., Mazur-Wierzbicka, E., & Zwiech, P. (2010). *Gospodarowanie kapitałem ludzkim*. Szczecin: Wydawnictwo Economicus.

[3] Juchnowicz, M. (Ed.). (2007). *Elastyczne zarządzanie kapitałem ludzkim w organizacji wiedzy*. Warszawa: Difin.

[4] Lewicka, D. (2011). *Zarządzanie kapitałem ludzkim w polskich przedsiębiorstwach*. Warszawa: PWN.

[5] López-Cabrales, Á., Real, J.C., & Valle, R. (2011). Relationships between Human Resource Management Practices and Organizational Learning Capability: The Mediating Role of Human Capital. *Personnel Review*, 40(3), 344-363.
[6] Pearse, N.J. (2009). The Role of Experiences in Creating and Developing Intellectual Capital. *Management Research News*, 32(4), 371-382.
[7] Pickett, L. (2005). Optimising Human Capital: Measuring What Really Matters. *Industrial and Commercial Training*, 37(6), 299-303.
[8] Samul, J. (2012). Wynagradzanie efektywności kapitału ludzkiego w praktyce przedsiębiorstw branży budowlanej.
[in:] P. Wachowiak (Ed.), *Człowiek w organizacji, Teoria i praktyka*. Warszawa: Oficyna Wydawnicza SGH w Warszawie.
[9] Skrzypek, E., & Hofman, M. (2010). *Zarządzanie procesami w przedsiębiorstwie. Identyfikowanie, pomiar, usprawnianie*. Warszawa: Oficyna Wolters Kluwer business.
[10] Skrzypek, E. (2000). *Jakość i efektywność*. Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.
[11] Ścibiorek, Z. (2010). *Zarzadzanie zasobami ludzkimi*. Warszawa: Difin.
[12] Tylec, A., & Wielgórka, D. (2011). Istota i wielowymiarowość efektywności organizacyjnej. *Zeszyty Naukowe Politechniki Częstochowskiej, Zarządzanie*, 2, 59-71.
[13] Tyrańska, M. (2009). Jakość kapitału ludzkiego a efektywność przedsiębiorstwa. *Problemy Jakości*, (5), 9-16.
[14] Urbanek, G. (2011). *Kompetencje a wartość przedsiębiorstwa. Zasoby niematerialne w nowej gospodarce*. Warszawa: Wydawnictwo Wolters Kluwer business.

On Non-classical Methods of Design of Experiments Using the R Program

Małgorzata Szerszunowicz

University of Economics in Katowice, Poland

Abstract

Design of experiments, apart from control charts and acceptance sampling, is one of the tools of statistical quality control. The effective use of experimental design methods leads to improvements in the technological and economical results of a production process. The essential phase of design of experiments is the estimation of response surface function, which best characterizes the process under study. In practice, experimental designs which are most frequently used by production companies include full and fractional factorial designs. The methodology of design of experiments also deals with the theory of production process optimization. The aim of this paper is to propose the introduction of design of experiments procedures which could be used in cases in which the use of classical statistical methods is not possible. Moreover, the proposed methods will be presented using the R program language.

Keywords: design of experiment, quantile regression, bootstrap, `R-CRAN`

1. Introduction

Design of experiments is a statistical tool used in agriculture, medicine and biology, but special attention should be paid to its application in statistical quality control.

2. The classical factorial design of experiments

Design of experiments is a part of the phase which precedes production processes, so its application should be carried out in accordance with the criteria formulated by D. C. Montgomery as follows [7, 8]:

- recognize and define the problem by determining all the aspects, circumstances and potential objectives of the experiment,
- appropriately select the factors, their levels and ranges and explore the possibility of considering them in the experiment,
- define the response variable,
- choose a proper design of experiment by determining the number of experimental trials and the possible randomization restrictions;
- perform the experiment,
- analyze the results using statistical methods,
- formulate conclusions and recommendations resulting from the analysis of the results.

The experiment should be understood as a sequence of n experimental trials, where a single experimental trial is a single result of random variable Y , with fixed values of factors $X_1, X_2, \ldots, X_m$. Let Let us define $\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_m}$, as the sets of all the possible values of factors $X_1, X_2, \ldots, X_m$, and then the experimental area is a set of points $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ where $x_i \in X_i$, for $i = 1, 2, \ldots, m$. The set of pairs $P_n = \{\mathbf{x}_j, p_j\}_{j=1}^n$ defines a design of experiment with n experimental trials, $\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{mj})$ and $p_j = \frac{n_j}{n}$ where n_j is the number of experimental trials in point $\mathbf{x}_j$ of the experimental area. Moreover $\sum_{j=1}^n n_j = n$ and $\sum_{j=1}^n p_j = 1$ for $j = 1, 2, \ldots, n$. The relationship between the set of random and non-random factors $X_1, X_2, \ldots, X_m$ and the response variable Y can be presented in the form of the following statistical model [12]

$$Y(X_1, X_2, \ldots, X_m) = y(X_1, X_2, \ldots, X_m) + \epsilon \tag{1}$$

where $EY(X_1, X_2, \ldots, X_m) = y(X_1, X_2, \ldots, X_m)$, $E\epsilon = 0$ and $V\epsilon = \sigma^2$. The function $y(x_1, x_2, \ldots, x_m)$ is named response surface function. The arguments of the response surface function are m realizations of non-random variables $X_1, X_2, \ldots, X_m$. The model (1) can be presented as a general linear model as $\mathbf{Y} = \mathbf{F}\beta + \epsilon$ where

$$\mathbf{Y^T} = (Y_1 \ Y_2 \ \ldots \ Y_n) \tag{2}$$

$$\epsilon^{\mathbf{T}} = (\epsilon_1 \ \epsilon_2 \ \ldots \ \epsilon_n) \tag{3}$$

$$\beta^{\mathbf{T}} = (\beta_1 \ \beta_2 \ \ldots \ \beta_n) \tag{4}$$

$$\mathbf{f^T}(\mathbf{x}) = (f_1(\mathbf{x}) \; f_2(\mathbf{x}) \; \ldots \; f_k(\mathbf{x})) \tag{5}$$

$$\mathbf{F} = \begin{bmatrix} f_1(\mathbf{x_1}) & \ldots & f_k(\mathbf{x_1}) \\ \vdots & \ddots & \vdots \\ f_1(\mathbf{x_n}) & \ldots & f_k(\mathbf{x_n}) \end{bmatrix} \tag{6}$$

The estimation of parameters of response surface function $\mathbf{y} = \mathbf{F}\beta$ is carried out with the use of least squares method.

In specialist literature ([7], [8], [12]), the following response surface function is considered for classical factorial designs of experiments

$$y(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \beta_{12} x_1 x_2 + \ldots + \beta_{12\ldots m} x_1 x_2 \ldots x_m \tag{7}$$

In justified cases, one ought to assume that the interactions between factors equal 0, and then the response surface function (7) can be presented by means of the following formula:

$$y(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m \tag{8}$$

The estimation of response surface function parameters consists in performing so many experimental trials that the estimation of all the parameters of response surface function is possible. In practice, the most frequently used designs of experiments are full and fractional designs of experiments.

3. Non-classical procedures in design of experiments

The least squares method is the most frequently used method of estimation of regression function presented as a general linear model. One of the assumptions of the least squares method is the Gaussian distribution of random component of the considered model. In practice, this assumption cannot frequently be ensured. Then, in accordance to the design of experiments methodology, the interpretation of estimated response surface function can lead to incorrect conclusions and recommendations for the manufacturing process. When the least squares method estimation is improper, M. Szerszunowicz suggests using, as an alternative, the quantile regression [10] and bootstrap method of estimation [11].

3.1. Quantile regression

The concept of quantile regression was introduced by Koenker and Basset in 1978 [4]. Currently, the quantile regression method is a statistical tool widely used in the modelling of economic phenomena. The aim of quantile regression is to determine a linear model – on the basis of an n-elemental sample of an unknown distribution F – in the following form

$$Q_\tau(Y \mid X) = X^T \beta_\tau \tag{9}$$

where $\tau \in (0, 1)$ is a fixed for sample quantile, and β_τ is a vector of unknown values of the parameters of a given linear model. In this case, the values of vector β_τ are estimated by means of an estimator in the form of

$$\hat{\beta}_\tau = \min_{b \in R^k} \sum_{i=1}^{n} \rho_\tau(y_i - x_i^T b) \tag{10}$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}_{(u<0)})$. Then the estimator of model (9) is

$$\hat{Q}_\tau(Y \mid X) = X^T \hat{\beta}_\tau \tag{11}$$

The procedure for the appointment of the estimator 10 may be written as a linear programming problem [4]. In design of experiments, the theory of quantile regression allows one to estimate response surface functions which describe the relationship between the high or low values of response variable and levels of factors, which may improve the analysis of production processes [10].

3.2. Bootstrap estimation

When the assumptions of classical statistical methods cannot be ensured, the bootstrap methods are used more and more frequently. Also, the bootstrap method has been introduced for the estimation of the parameters of regression functions [2, 3]. The bootstrap method for the statistical model in the form of

$$y_i = f(x_i, \beta) + \epsilon_i \tag{12}$$

can be presented in the following phases [3]:

- determination of the values of $\hat{\beta}$ estimator with the use of the least squares method,

- estimation of the vector of residuals according to the following formula $\hat{\epsilon}_i = y_i - f(x_i, \hat{\beta})$,
- generation of a bootstrap sample $(\epsilon_1^*, \epsilon_2^*, \ldots, \epsilon_n^*)$ according to the specified probability distribution $P(Z = \hat{\epsilon}_i) = \frac{1}{n}$,
- formation of a sample $(Y_1^*, Y_2^*, \ldots, Y_n^*)$ with realizations $y_i^* = f(x_i, \hat{\beta}) + \epsilon_i^*$,
- estimation of the vector of parameters for the specified model with the use of the least squares method on the basis of the obtained values $(y_1^*, y_2^*, \ldots, y_n^*)$,
- generation of a vector of the values $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_N)$ with N-time realization of points 3, 4 and 5 of the present algorithm.

The determined vector describes the bootstrap distribution of the estimator of the parameters of response surface function. This vector allows one to define the appropriate confidence interval for the parameters in question [6].

4. Empirical case

T.P. Ryan [9] takes into account the influence of two factors: X_1 - dilution rate - and X_2 - pH level - both on three levels, on response variable Y – biomass concentration. The design of experiment under consideration is presented in Table 1.

No	X_1	X_2	Y
1	-1	-1	3.8
2	-1	0	4.0
3	-1	+1	1.4
4	0	-1	5.1
5	0	0	4.867
6	0	+1	1.9
7	+1	-1	4.5
8	+1	0	3.8
9	+1	+1	0.8

Table 1: Experimental data

Let us assume the following response surface function:

$$\hat{y}(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{13}$$

According to the classical approach, in order to estimate the parameters of the response surface function which describes the dependence between the response variable and factors, one ought to use the least squares method. The procedure for the estimation of response surface function parameters and its results using the R software is presented below.

```
> model_lin= lm(Y~X[,1]+X[,2])
> summary(model_lin)
Call:
lm(formula = Y ~ X[, 1] + X[, 2])
Residuals:
     Min       1Q   Median       3Q      Max
-1.11852 -0.41852  0.09815  0.46481  1.51481

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.35185    0.31925  10.499 4.38e-05 ***
X[, 1]      -0.01667    0.39100  -0.043  0.96738
X[, 2]      -1.55000    0.39100  -3.964  0.00742 **

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.9577 on 6 degrees of freedom
Multiple R-squared: 0.7237,     Adjusted R-squared: 0.6316
F-statistic: 7.859 on 2 and 6 DF,  p-value: 0.02109
```

With reference to the incorrect specifications for manufacturing processes, which correspond to the least squares method, the non-classical methods of estimation of response surface parameters is presented below.

4.1. The use of quantile regression

In order to estimate response surface function parameters (13) with quantile regression, the appropriate function from the `R quantreg` package has been used. The notation of this procedure is presented below.

```
> beta.rq=coef(rq(Y~X[,1]+X[,2], tau=c(0.05,0.25,0.5,0.75,0.95)))
> beta.rq
            tau= 0.05 tau= 0.25 tau= 0.50 tau= 0.75  tau= 0.95
```

```
(Intercept)        2.3      2.95       3.5       3.9  4.8666667
X[, 1]            -0.3      0.35       0.3      -0.1  0.8666667
X[, 2]            -1.2     -1.20      -1.6      -1.2 -0.2333333
```

Then the response surface function (13) for $\tau = 0.5$ is determined by the equation

$$\hat{y}_{\tau=0.5}(x) = 3,5 + 0,3x_1 - 1,6x_2 \tag{14}$$

The plot of the presented response function (14) can be determined in the R software with following notation:

```
f.5rq=function(x,y) beta.rq[1,3]+beta.rq[2,3]*x+beta.rq[3,3]*y
z5=outer(x1,x2,f.5rq)
persp(x1, x2, z5,col = "grey20", xlab = "X_1", ylab = "X_2",
zlab = "Y", theta=315, phi=0)
```

The realization of the proposed notation is presented in Figure 1.

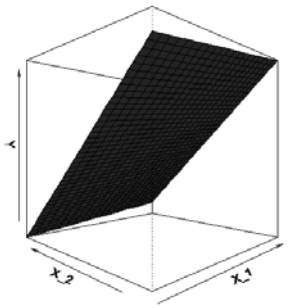


Figure 1: Response surface function for $\tau = 0.5$

Figure 2 presents the plots of response surface functions for $\tau \in \{0.05, 0.25, 0.75, 0.95\}$.

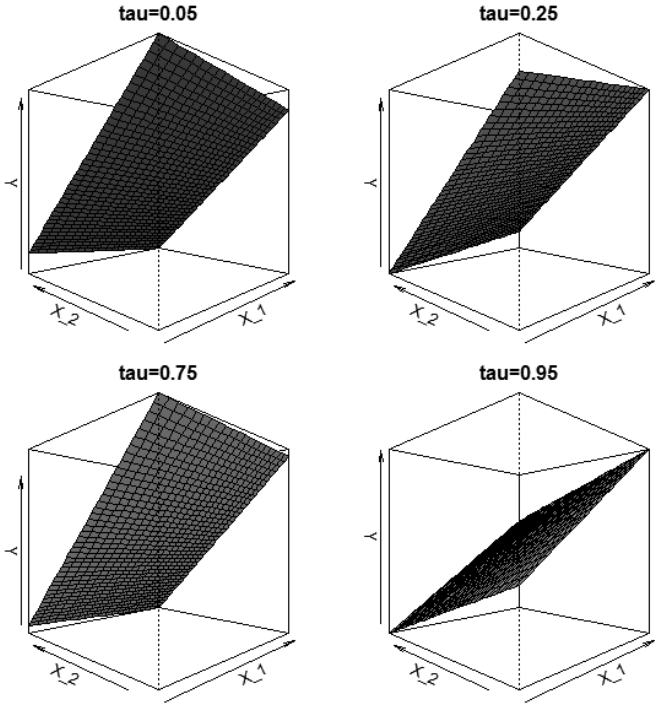


Figure 2: Response surface function for different values of τ

The estimation of the form of response surface function with quantile regression allows one to obtain a wide range of dependence between the response variable and the factors, which leads to the complementation of the analysis of the considered production process analysis.

4.2. The use of bootstrap estimation

For the considered experimental data, the response surface parameters (13) have also been estimated with the bootstrap method presented in subsection 3.2. The notation of this procedure is presented as follows:

```
> epsilon_0=summary(model_lin)$residual
> e=sample(epsilon_0)
> beta=beta_mlin
> y=c()
> n=1000
> betaB=matrix(0,n,length(beta))
> betaB[1,]=beta
```

```
>
> for (i in 2:n)
+ {
+ e=sample(epsilon_0)
+ y=beta[1] + beta[2]*X[,1]+beta[3]*X[,2]+e
+ model = lm(y~X[,1]+X[,2])
+ beta_=summary(model)$coef[,1]
+ betaB[i,]=beta_
+ }
> beta_B0=mean(betaB[,1])
> beta_B0
[1] 3.351852
> beta_B1=mean(betaB[,2])
> beta_B1
[1] -0.02519167
> beta_B2=mean(betaB[,3])
> beta_B2
[1] -1.539811
```

Thanks to the computational possibilities of the R program, the estimation of the response surface function has been obtained as the following formula:

$$\hat{y}_{\tau=0.5}(x) = 3,35 - 0,03x_1 - 1,54x_2 \tag{15}$$

Just like in the quantile regression case, the plot of the estimated function has been created and presented in Figure 3.

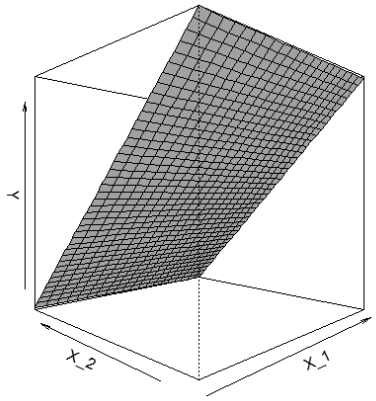


Figure 3: Response surface function - bootstrap estimation

The application of the bootstrap method in estimating the response surface function also allows one to determine the distribution of the parameters of the response surface function. Moreover, this method leads to the determination of confidence intervals for the considered parameters with the use of the percentile method, among other things [3, 6].

5. Conclusion

The procedures for preparing production processes in practice do not always allow one to verify the assumptions of classical statistical methods. If this is the case, one ought to search for alternative methods which produce reliable results. The R software, thanks to its computational capabilities, allows one to use alternative methods of estimation of response surface parameters, i.e. the quantile regression and bootstrap estimation. The use of the proposed estimation methods results in the extension of the manufacturing process analysis, which has an important impact on economical and technological results.

Acknowledgements

The research was supported by Polish Science Center grant DEC-2011/03/B/HS4/05630.

References

[1] Antony, J. (2003). *Design of Experiments for Engineers and Scientists.* Oxford: Butterworth-Heinemann.

[2] Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap.* New York: Chapman and Hall.

[3] Domański, C., & Pruska, K. (2000). *Nieklasyczne metody statystyczne.* Warszawa: PWE.

[4] Koenker, R., & Basset, G. (1978). Regression quantiles. *Econometrica*, 46, 33-50.

[5] Kończak, G. (2007). *Metody statystyczne w sterowaniu jakością produkcji.* Katowice: Wydawnictwo Akademii Ekonomicznej.

[6] Kończak, G. (2012). *Wprowadzenie do symulacji komputerowych.* Katowice: Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach.

[7] Montgomery, D.C. (2001). *Design and Analysis of Experiments.* New York: John Wiley & Sons, Inc.

[8] Montgomery, D.C. (1997). *Introduction to Statistical Quality Control.* New York: John Wiley & Sons, Inc.

[9] Ryan, T.P. (2007). *Modern Experimental Design.* New Jersey: John Wiley & Sons, Inc.

[10] Szerszunowicz, M. (2013). *On a Method of Estimation of Response Surface Function.* Mathematical Methods in Economics 2013, College of Polytechnics Jihlava, Jihlava 2013.

[11] Szerszunowicz, M. (2014). On the Bootstrap Method of Estimation of Response Surface Function. *Acta Universitatis Lodziensis - Folia Oeconomica*, 3(302), 101-107.

[12] Wawrzynek, J. (2009). *Planowanie eksperymentów zorientowane na doskonalenie jakości produktu.* Wrocław: Wydawnictwo Uniwersytetu Ekonomicznego.

ISBN: 978-83-65173-38-6 (printed version)
ISBN: 978-83-65173-39-3 (on-line pdf)