STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **97** (4) 2017

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

EXECUTIVE BOARD

MANAGING EDITOR

CONTENTS

ANALYSES

METHODOLOGY

INFORMATION

About Statistika

The journal of Statistika has been published by the Czech Statistical Office since 1964. Its aim is to create a platform enabling national statistical and research institutions to present the progress and results of complex analyses in the economic, environmental, and social spheres. Statistika is professional double-blind peer reviewed open access journal included (since 2015) in the citation database of peer-reviewed literature **Scopus (SJR** 2016 = **0.121, CiteScore** 2016 = **0.15)**, in the **Web of Science** *Emerging Sources Citation Index* (since 2016) and also in other international databases of scientific journals. Since 2011 Statistika has been published quarterly in English only.

Publisher

The Czech Statistical Office is an official national statistical institution of the Czech Republic. The Office´s main goal, as the coordinator of the State Statistical Service, consists in the acquisition of data and the subsequent production of statistical information on social, economic, demographic, and environmental development of the state. Based on the data acquired, the Czech Statistical Office produces a reliable and consistent image of the current society and its developments satisfying various needs of potential users.

Contact us

Journal of Statistika | Czech Statistical Office | Na padesátém 81 | 100 82 Prague 10 | Czech Republic
e-mail: statistika.journal@czso.cz | web: www.czso.cz/statistika_journal

Changes in Methodology for Assessing Performance of Research Organisations and Influence of Such Changes on Researchers' Behaviour[1]

Luboš Marek[2] | *University of Economics, Prague, Czech Republic*
Stanislava Hronová[3] | *University of Economics, Prague, Czech Republic*
Richard Hindls[4] | *University of Economics, Prague, Czech Republic*

Abstract

Assessing quality of research results on an international scale is a basis for evaluating the level of scientific activities pursued in research organisations. In the past 15 years, significant changes have occurred in the Czech Republic in research management and, in particular, the methodology of assessing research results. The methodology of assessment and its modifications should always be focused on increasing quality of research results; the rules of assessment have their effects on researchers' behaviour. This paper studies a question of whether the changes applied to the methodology of assessing research results in the Czech Republic have supported higher quality research results, i.e., results published in high-quality international journals. The authors have developed their own statistical test to measure significance of such changes, as well as other statistical tests of hypotheses. The main source is represented by the results of assessing public universities in the Czech Republic according to "Methodology for assessing results of research organisations" in 2010 and 2013. Our tests have not proven any statistically significant differences in the numbers of papers published in the journals monitored in the Web of Science and Scopus databases.

Keywords

Assessment methodology, test of significance of the changes, public universities

JEL code

C12, I20

[1] This article was processed with contributions from long-term institutional support of research activities by the Faculty of Informatics and Statistics, University of Economics, Prague.

[2] Faculty of Informatics and Statistics, Department of Statistics and Probability, W. Churchill Square 4, 130 67 Prague 3, Czech Republic. E-mail: marek@vse.cz.

[3] Faculty of Informatics and Statistics, Department of Economic Statistics, W. Churchill Square 4, 130 67 Prague 3, Czech Republic. E-mail: hronova@vse.cz.

[4] Faculty of Informatics and Statistics, Department of Statistics and Probability, W. Churchill Square 4, 130 67 Prague 3, Czech Republic. E-mail: hindls@vse.cz.

INTRODUCTION

Assessing research results undoubtedly has motivational effects. Emphasis on a higher quality of basic-research results should be manifested in a higher number of results passing internationally recognised criteria of evaluation, i.e., papers published in journals with nonzero impact factor values monitored by the Web of Science and Scopus databases. The motivational effects of assessment in research organisations in the Czech Republic have been widely discussed. A question arises: have qualitative changes in this methodology really led to a focus on better publication output in recent years?

1 ASSESSMENT OF R&D RESULTS IN THE CZECH REPUBLIC

Systemic evaluation of research results on the basis of strictly specified rules and procedures in the Czech Republic dates back to 2004. The reason for introducing the evaluation system based on a methodology approved by the government were the stagnating and, in certain fields, even decreasing numbers of R&D results in the Czech Republic while the expenses incurred on R&D from the state budget were increasing. The position of the Czech Republic in the international comparison was worsening. The goal was to motivate researchers to higher quality and quantity of research results via allocating the means provided to the research organisation from the state budget on the basis of the assessment results. The first "Methodology of assessment of R&D results" (hereinafter the "Methodology") was approved in 2004; the currently valid methodology is called Methodology 2017+.[5]

The basic general rules for the assessment (creating the database containing the information about the R&D results, definitions of the output types – papers, books, patents, applied results, evaluation for the five most recent years, and evaluation of R&D efficiency) were set out in the National R&D policy of the Czech Republic in the period of 2004–2008.[6] As a future plan, the policy mentions a relationship between the assessment results and allocation of financial means to research organisations. This methodology has been a set of measures and tools to assess R&D results. The formulation was changed every year, which fact has always been criticised by research organisations. However, the changes were implied by the effort to rectify the most serious errors and shortcomings of the preceding version. Problems concerning the concept of the methodology for assessing the R&D results were fully manifested when the assessment results were applied according to Methodology 2008; it was the first time when part of the means from the state budget was allocated according to the assessment results.

1.1 Development of assessment methodology for research organisations in the Czech Republic

The evolution of the Methodology in the Czech Republic can, from the viewpoint of principles, be divided into four stages. The first stage (Methodology 2004 – Methodology 2009) represents the beginnings of the assessment principles (unfortunately, sometimes by trial-and-error) and the modification of the rules every year. The second stage came with Methodology 2010, whose validity was first approved for two years (2010 and 2011), and later extended to 2012. Methodology 2010 brought a number of modifications directed at respecting specific features of different fields, but it did not rectify the fundamental shortcoming of all previous methodologies, namely, the focus on quantity. The third stage is represented by Methodology 2013 (valid for the period of 2013–2016); it brought a fundamental change of combining bibliometric parameters with peer-review, and a different assessment for applied research results. The last stage, Methodology 2017+, should be a transition from evaluation of mere results to that

[5] The official name "Methodology of assessment of R&D results" was valid in the period of 2004–2009. In 2010, the name was changed to "Methodology of assessing results of research organisations and results of completed projects". For the sake of brevity, we will use the simplified name Methodology, or Methodology with the year of validity, i.e., Methodology 2004, Methodology 2005, etc.

[6] Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=5580*>.

of each research organisation as a whole (not only on the basis of the results). Ideological theses of this Methodology have been published, and its full implementation is expected in the years 2019 and 2020.

The principles of the first Methodology of assessment of R&D results[7] (the so-called Methodology 2004) were very simple[8] and completely insufficient for assessing the quality of scientific results (all types of results were valued identically, by one point). The results of this assessment should lead to classifying all research organisations into three categories according to efficiency of the means incurred on R&D (above-average, average, and below-average). Consequently, the allocation of financial means in future years should have been related to that classification. However, this stage was not implemented due to the lack of a criterion of efficiency and the disputed assignment of the same values to all types of output.

Methodology 2005[9] (for assessing the results achieved in the period of 2000–2004) was just a more accurate update of Methodology 2004. It newly distinguished between different types of results – a paper in a journal with nonzero impact factor, a paper in another type of professional journal, a professional book, a chapter in a book, a contribution to conference proceedings, a patent, and an applied research result – and a higher number of points is always assigned to a publication in a world language. An index was set up for comparing the number of points assigned for the results achieved with the R&D means allocated to the given organisation from the state budget. On the basis of this index, all organisations were classified into four colour-coded categories. Such classification according to the efficiency level should have a positive/negative impact on the amount of means allocated from the state budget in future years. However, the results of this assessment turned out to be very disputable and a system for future allocation of financial means was not implemented.

In the introduction to Methodology 2006[10] (for assessing the results achieved in the period of 2001–2005), it is said on page 2 that "applications of principles given in Methodologies 2004 and 2005 did not bring the expected effects and, despite the ever-increasing R&D expenses from the state budget, many fields of science in the Czech Republic lag behind even more". Hence "SR index" (a ratio between the number of points obtained for results and the amount of the R&D means allocated to the given organisation from the state budget)[11] was defined as an indicator of efficiency. Similar to Methodology 2005, research organisations were again classified into four colour-coded categories according to their efficiency levels. Another modification in this Methodology was concerned with increasing the number of types of applied research results and increase of their point valuation compared to basic research results. This approach was criticised and later led to an "inflation" of these types of results.

Methodology 2007[12] (for assessing the results achieved in the period of 2002–2006) was an update of Methodology 2006 and emphasis was again – even if disputably – put on the efficiency level expressed by the SR index. For the first time, this Methodology admitted verbal descriptions of the results and points to social sciences and humanities were assigned differently from other sciences.

Methodology 2008[13] (for assessing the results achieved in the period of 2003–2007) brought many more changes. The SR index was abandoned with respect to results of research organisations (but remained for assessing results of completed programmes); and only humanities were set aside for assessment and social sciences were returned back to the other sciences. A group of Czech journals was defined so that only papers in those selected journals would pass for the assessment, and contributions

7 Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=18750*>.

8 Just for comparison: Methodology 2004 was a six-page text; Methodology 2013 was 59 pages.

9 Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=18751*>.

10 Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=21846*>.

11 SR index = index of the state budget; state budget = "Státní Rozpočet" in the Czech language.

12 Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=31543*>.

13 Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=503762*>.

to proceedings were restricted to ISI Proceedings (today CPCI). This Methodology was for the first time used to allocate the financial means for 2010 pursuant to the amended Act No. 130/2002 Coll. The intention to do so had already been announced in Methodology 2004; nevertheless, the direct calculations of financial R&D allocations from the state budget on the basis of obtained numbers of points a surge of disagreement came from research institutions.

Prior to approval of Methodology 20091 (for assessing the results achieved in the period of 2004–2008), there was a very sharp debate in the academic sphere about "what now" – the gradual improvements of the Methodology had not removed its basic shortcomings (motivation to quantity, not quality of results; no differentiation by fields; no peer-review; etc.). The resulting changes were, however, minor; e.g., a category of prestigious journals was introduced (Nature, Science) with a high assignment of points for results published in them.

Methodology 2010 and 20112 (for assessing the results achieved in the period of 2005–2009, or rather 2006–2010) under a new name of "Methodology of assessing results of research organisations and results of competed projects" tried to cope with the most glaring problems in the assessment process. That is why a chapter on allocation of financial means was, for the first time, included into the Methodology; in that chapter, an idea occurred that the means should be divided by fields and the points should be corrected with respect to the numbers of results. Results published in the journals monitored in the Scopus and ERIH databases were newly added to the results to be assessed.

Methodology 2012[16] (for assessing the results achieved in the period of 2007–2011) was, in principle, an extension of Methodology 2010–2011. Only the chapter on the allocation of financial means on the basis of the assigned numbers of points was modified (made more specific).

Preparations of Methodology 2013[17] had taken a lot of time; this Methodology introduced fundamentally different methods for result evaluation. In addition to bibliometric evaluation, exclusively applied to that date, peer-review evaluation of papers and books was to be applied, as well as evaluation of selected excellent results. Panels of reviewers were set up, in which experts from abroad also participated. The methods for assessing applied research results were also modified, but this concept was criticised. This evaluation process was originally planned to take place every year, which was too demanding; this fact led to delays and degradation of the originally good idea. This Methodology should have been valid for the period of 2013–2015, but it was later extended to 2016 (that is, results were evaluated for the periods 2008–2012, 2009–2013, 2010–2014, and 2011–2015).

Despite many year-to-year modifications, each Methodology was just a tool for calculating money from obtained points.[18] A comprehensive system for assessing research had been and still is missing, which would view a research organisation regarding not only the results achieved, but also other aspects of activities pursued in R&D. This approach should be implemented in the new "Methodology for assessing research organisations and targeted-support programmes in research, development and innovations"[19] (Methodology 2017+), whose roles are to be introduced in the period of 2017–2019; beginning 2020, the comprehensive assessment should be carried out in five-year cycles, not every year (the annual assessments turned out to be impossible to implement).

1.2 Motivational effects of Methodology

All the year-to-year modifications of the Methodology were motivated by the effort to improve the assessment of results and respond to criticism from research organisations. This criticism was mainly aimed

[14] Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=532412*>.
[15] Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=566918*>.
[16] Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=650022*>.
[17] Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=685899*>.
[18] For this reason, this methodology is often called the "coffee grinder" in the Czech academic environment.
[19] Cf. <*http://www.vyzkum.cz/FrontClanek.aspx?idsekce=799796*>.

at the lack of concept in creation of the Methodology (this lack was namely proved by the year-to-year changes), no regard to specific aspects of each field, subjective and erratic character of the point values assigned to individual results, and preferring quantity to quality. The last-mentioned aspect was the reason for the inflation of low-quality results and non-ethical behaviour of certain research organisations, which led to the necessity of sanctions for wrongly reported results.[20] The direct relationship between the assessment results and allocation of financial means was also criticised, because this relationship had negative impacts on management of some research organisations.

On the other hand, there was a positive effect of the mere fact that a methodology was created to implement the outcome of the discussion about possibilities in assessment of results achieved by research organisations. The awareness that research activities must be assessed was important. However, it is disputable whether the Methodology modifications always brought the expected impacts on increasing not only quantity, but also quality of research activities.

We asked whether qualitative changes in Methodology 2013 as compared with Methodology 2010 were reflected in a better quality of research results. For the purposes of this study, we deem high-quality results papers published in the Web-of-Science-monitored journals (denoted by Jwos) and the Scopus-monitored journals (denoted by Jsc). Papers published in such journals undergo an independent review process according to international standards and can, therefore, be viewed as a certain indicator of good quality of research activities.[21] If Methodology 2013 was to bring a new approach to result evaluation and motivate researchers to focus on high-quality results, numbers of the Jwos and Jsc papers should be higher within assessment according to Methodology 2013 (for the period of 2008–2012) than those according to Methodology 2010 (for the period of 2005–2009).[22] Even though it is clear that there is a two-year overlap, newer data could not be used due to the requirements for comparability of results – the assessments in 2014 and 2015 follow the principles of Methodology 2013, but only numbers of points assigned to the so-called assessment pillars are public, not the numbers of results by the type (papers, books, etc.). The results of the assessment which should have been made in 2016 are yet not available. Due to the incomparability of the assessment reports, neither older data (assessment according to Methodology 2009 and older) could be used.

2 METHODS OF ANALYSIS AND THE DATA USED

In order to verify the hypothesis that the changes in Methodology 2013 brought a fundamental change in quality, manifested by increased numbers of Jwos and Jsc results, we will apply our originally developed statistical test of significance of the changes, as well as standard hypothesis testing. The variables of interest are the numbers of papers published by twenty Czech public universities in the Web-of-Science-monitored (Jwos) and Scopus-monitored (Jsc) journals in two different periods of time and according to different methodologies for assessing research organisations' results – namely, M2010 (Methodology 2010, period of 2005–2009) and M2013 (Methodology 2013, period of 2008–2012). The source of the data was the R&D Information System of the Czech Republic.

To form a basic idea of the character of the data to be processed, we reviewed descriptive statistics, which may indicate some differences. The values of the descriptive statistics are shown in Table 1. The input data (numbers of articles Jwos and Jsc by universities) related to the partial calculations in the significance test are given in Table 1A in the Appendix.

[20] Nevertheless, this effort did not have the desired effect; the sanctions for incorrectly claimed results were only applied once.

[21] We are aware that this assumption is not exactly true: it is clear that not all fields have papers in professional journals as their main output, and not all journals monitored by these databases are particular about the high professional quality.

[22] Therefore our analysis takes in account only bibliometric data.

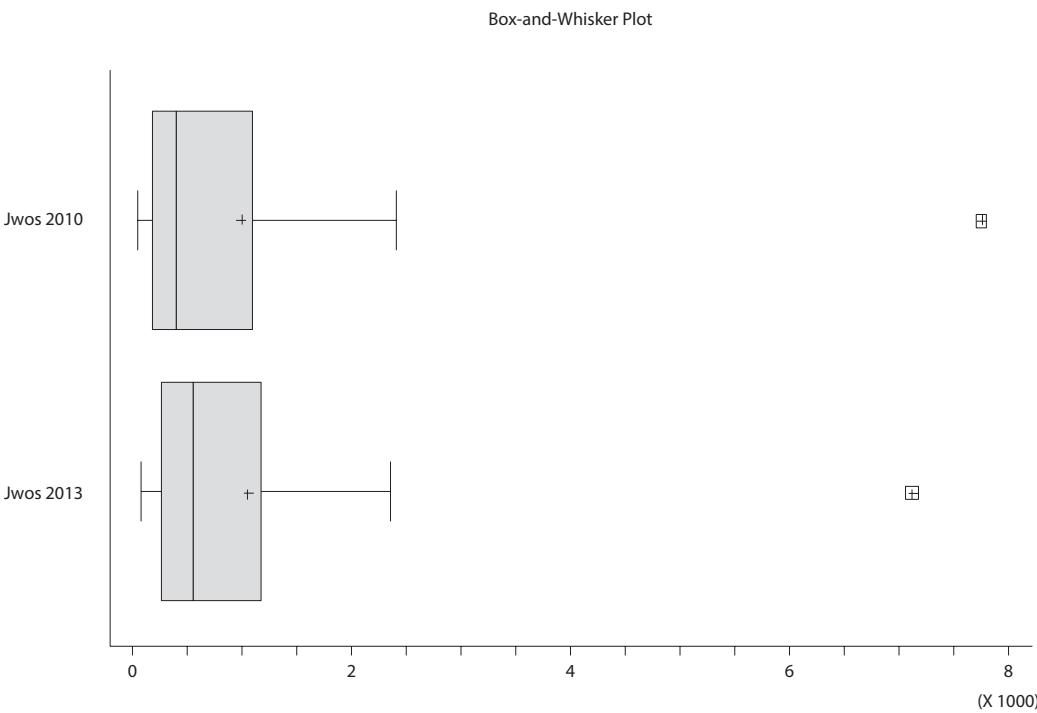
Table 1 Values of selected descriptive characteristics for Jwos and Jsc

	Jwos		Jsc	
	M2010	**M2013**	**M2010**	**M2013**
Average	982.3	1 052.5	445.9	529.3
Standard deviation	1 705.1	1 556.4	881.1	716.4
Coefficient of variation (%)	173.6	147.9	197.6	135.4
Minimum	30.0	60.0	48.0	58.0
Maximum	7 751.0	7 117.0	3 936.0	3 182.0

Source: <*www.rvvi.cz*>, authors' own results

This preview is confirmed by box-and-whisker plots. Regarding numbers of papers published in the Web-of-Science-monitored journals, no substantial differences are observed between the 2005–2009 and 2008–2012 periods (cf. Figure 1).

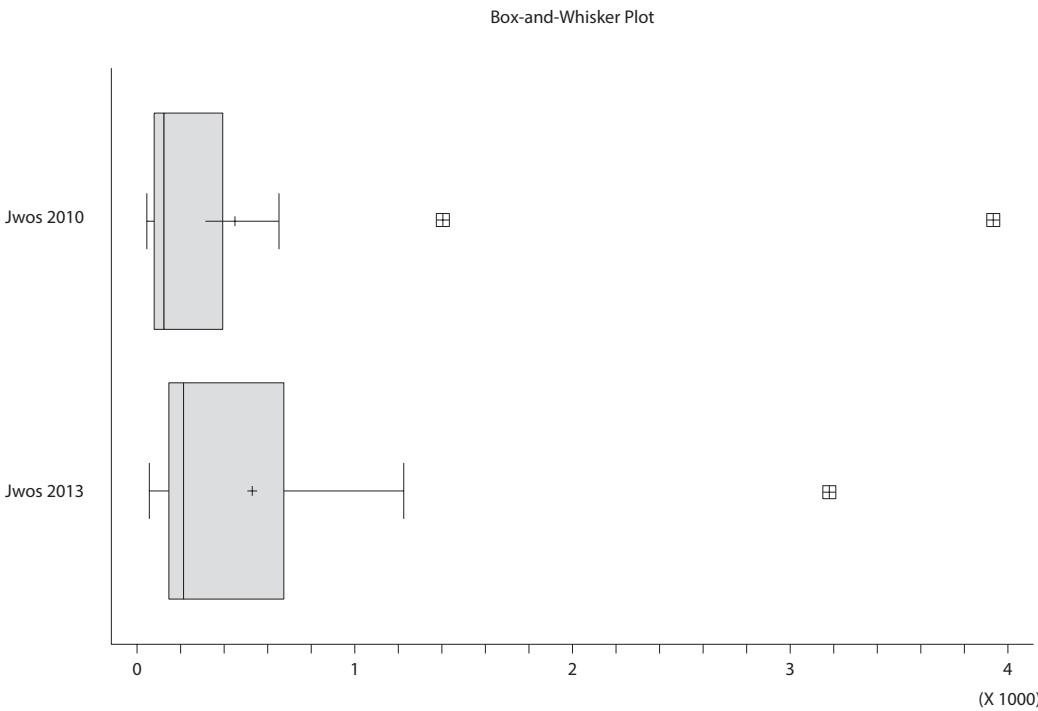
Figure 1 Box-and-whisker plot for Jwos



Source: <*www.rvvi.cz*>, authors' own results

On the other hand, changes in both levels and variability values are clearly seen for the numbers of papers published in the Scopus-monitored journals (cf. Figure 2).

Figure 2 Box-and-whisker plot for Jsc



Source: <*www.rvvi.cz*>, authors' own results

Selected statistical hypothesis tests will clarify whether the changes in the numbers of papers published in the Scopus-monitored journals (Jsc) outbalance the lack of such changes in the numbers of papers published in the Web-of-Science-monitored journals (Jwos), and whether, consequently, the overall changes in the numbers of papers published in internationally renowned journals (according to Methodology 2013 as compared with Methodology 2010) can be considered statistically significant.

2.1 Test for significance of changes

A test for measuring the significance of changes from one situation to another was proposed by two of the authors of the present paper (cf. Hindls and Hronova, 2007). It has turned out that this test is very well capable of identifying significance of changes from one situation in time to another. Here we try to establish a change in two variables (Jwos and Jsc publication numbers) for several units (public universities in the Czech Republic except for universities of arts) in two periods of time (according to M2010 and M2013). The hypothesis to be tested states that the numbers of the Jwos and Jsc publications were the same in both time periods of interest, while the alternative hypothesis denies the tested one in the sense that the numbers of the Jwos and Jsc publications in the period of 2008–2012 (i.e., according to M2013) was statistically significantly higher than in the period of 2005–2009 (i.e., according to M2010). An advantage of this test is the fact that, unlike the standard tests (cf. Section 2.2) it measures the overall significance of the changes in both variables at the same time.

Starting points and notation

Let us denote the first surveyed characteristic of two-criterion evaluation as x (number of articles Jwos), and the second one as y (number of articles Jsc). Further, we introduce the symbols 1 and 2 for the corresponding evaluation methodologies. We thus employ the following symbols:

- x_{1i} for the number of Jwos articles of i-th public university, $i = 1, 2, ..., n$, (in this case $n = 20$) in the first period (according to M2010);
- x_{2i} for the number of Jwos articles of i-th public university, $i = 1, 2, ..., n$, (in this case $n = 20$) in the second period (according to M2013);
- y_{1i} for the number of Jsc articles of i-th public university, $i = 1, 2, ..., n$, (in this case $n = 20$) in the first period (according to M2010);
- y_{2i} for the number of Jsc articles of i-th public university, $i = 1, 2, ..., n$, (in this case $n = 20$) in the second period (according to M2013).

The formulation for the test of significance of the changes in researchers' attitude over time using two-criterion evaluation

Let us denote:

- by K_1 the mean value of the aggregate two-criterion evaluation in the first period (the so-called *mean space localisation*); and
- by K_2 the mean value of the aggregate two-criterion evaluation in the second period (the so-called *mean space localisation*).

We test the null hypothesis H_0 about equality of the mean localisation in the space coordinate system, i.e.,

H_0: $K_1 = K_2$

against an alternative hypothesis:

H_1: $K_1 < K_2$

We will use the following statistic as the test criterion:

$$T = \frac{\frac{\sum_{i=1}^{n} k_i}{n}}{\sqrt{\frac{\sum_{i=1}^{n}\left[k_i - \frac{\sum_{i=1}^{n} k_i}{n}\right]^2}{n(n-1)}}},$$

where:

$$k_i = \text{sign}\{y_{2i}^2 + x_{2i}^2 - y_{1i}^2 - x_{1i}^2\}\sqrt{(x_{2i} - x_{1i})^2 + (y_{2i} - y_{1i})^2},$$

$\frac{\sum_{i=1}^{n} k_i}{n}$ is the point estimator of the statistic $K_2 - k_1$.

It can be proved that this T statistic has Student's distribution $t\,[n - 1]$ under the validity of the tested hypothesis H_0.

Comment: Using the sign {...} operator, the orientation of the aggregate "space" change ("±") is determined for the level of the two-criterion value in the second (later) period in comparison with the first one. This operator thus expresses whether the i-th space localisation (i.e., the localisation of the i-th university) in the 2nd period (i.e., M2013) has moved nearer to ("–") or farther from ("+") the origin of coordinates [0; 0], in comparison with the 1st period (i.e., M2010). For example, when the i-th space

localisation has moved farther from the centre, then the sign "+" expresses that the aggregate number (i.e., for both surveyed characteristics together) of the *i*-th university has been improved (it is a kind of a "geometric" summary of the surveyed characteristics "number of Jwos" and "number of Jsc"). The power of the test is sharply increased when not using the sign operator. The probability of the type II error β would be lower and the risk of the type I error α would be higher.

The critical region of the test W is defined by the following inequalities:

$$W = \{\ T \geq t_{1-\alpha}[n-1] \text{ for left-sided alternative}\},$$

where: $t_{1-\alpha}[n-1]$ is the quantile of Student's t-distribution.

According to the result and using the significance level $\alpha = 0.05$, we draw a conclusion about the statistical significance of the time change in researchers' attitudes on the basis of results of the two-criterion evaluation.

The value of the test criterion, $T = 1.555$, does not exceed the critical level $t_{0.95}\,[19] = 2.093$. We can therefore observe that, at the 5% level of significance, the changes in the assessment methodology did not cause a statistically significant change in researchers' behaviour in the sense of their stronger focus on the results published in the Web-of-Science- and Scopus-monitored databases. The input data and partial calculations are shown in Table 1A.[23]

2.2 Tests for equality of expectation values

We have applied additional tests to verify (or reject) hypotheses formulated in compliance with this particular problem. Namely, a parametric test for equality of averages, and the Mann-Whitney (Wilcoxon) median test have been carried out. Unlike our own test mentioned above, these standard tests only deal with equality of expectation values for each of the variables of interest, i.e., separately for Jwos and Jsc. All tests were carried out at the 5% significance level.

Let us first review the outcome of the *t*-test concerning the equality of averages.[24] Hypothesis H_0 states that the values of the average numbers of papers were the same in both time periods of interest; the alternative hypothesis, H_1, denies H_0 in the sense that the average number of papers in the period of 2005–2009 (Methodology 2010) is smaller than that in the period of 2008–2012 (Methodology 2013).

Concerning the equality of averages for Jwos, the value of the test criterion is $t = -0.136$, and P-value = 0.446. Hence, the difference between the numbers of Jwos papers in the 2008–2012 and 2005–2009 periods cannot be viewed as statistically significant at the selected significance level. We have applied the test without knowing the variance values for the samples, but assuming that they are equal to each other. We have further tested this equality of variance values by F-test[25] (the value of the test criterion $F = 1.200$, and P-value = 0.695; the hypothesis that the variance values are equal to each other cannot be rejected at the selected significance level).

Similar conclusions has been made for Jsc, for which the value of the test criterion is $t = -0.328$, and P-value = 0.372. Hence, the difference between the numbers of Jsc papers in the 2008–2012 and 2005–2009 periods again cannot be taken for statistically significant at the selected significance level. We have applied the test without knowing the variance values for the samples, but assuming that they are equal to each other. We have again further tested this equality of variance values by F-test with the value of the test criterion $F = 1.512$, and P-value = 0.375; the hypothesis that the variance values are equal to each other cannot be rejected at the selected significance level).

---

[23] The results of this test for significance of changes are usually easy to display in graphical form; however, for the data processed here the graphical presentation would be unclear due to the necessity to display outlying observations. That is why we only present this tabular form of the results.

[24] Cf. e.g., Hindls et al. (2007).

[25] Cf. e.g., Hindls et al. (2007).

[26] Cf. Blatná (1996).

We have also applied a nonparametric test, namely, the Mann-Whitney (Wilcoxon) median test.[26] Hypothesis H_0 states that the medians of the numbers of papers were the same in both time periods of interest; the alternative hypothesis, H_1, denies H_0 in the sense that the median of the number of papers in the period of 2005–2009 (Methodology 2010) is smaller than that in the period of 2008–2012 (Methodology 2013).

Concerning the equality of medians for Jwos, the value of the test criterion is $W = 224.0$, and P-value = 0.262. Hence, the difference between the numbers of Jsc papers in the 2008–2012 and 2005–2009 periods cannot be taken for statistically significant at the selected significance level.

The same test for Jsc comes to the same conclusions. The value of the test criterion is $W = 259.5$, and P-value = 0.055. The difference between the numbers of Jsc papers in the 2008–2012 and 2005–2009 periods again cannot be taken for statistically significant at the selected significance level.

CONCLUSION

Changes in the methodology for assessing results of research organisations have been present in the academic environment in the Czech Republic since 2004. The original effort to simply keep records of research results has been replaced with different forms of and rules for assessing the results. Importance of year-to-year changes in such rules was regarded just marginally in the beginning, but later such changes were viewed negatively by research organisations (in particular, public universities). Since 2009, financial means from the state budget for long-term conceptual development of research organisations have been allocated according to the assessment results. Never-ending changes in the methodology were motivated by an effort to respond to the quickly changing environment in research organisations (which very quickly adapted themselves to the methodology rules), certain negative phenomena occurring in applying the methodology to management of research organisations, and – of course – to justified criticism. Methodology 2013 was the latest version of the methodology according to which assessment of research organisations' results was actually carried out and completed. This version of methodology brought qualitatively new aspects in assessment of results achieved by both basic and applied research. It was the last version of the methodology which took that approach; the currently valid Methodology 2017+ views assessment of results (i.e., bibliometric assessment) as one of five modules to be applied within assessing activities of research organisations.

A question arose whether the important qualitative changes in Methodology 2013 as compared with Methodology 2010 were positively reflected in the behaviour of public universities, namely, whether they caused an increase in the numbers of papers published in internationally renowned journals, i.e., monitored by the Web of Science and Scopus databases. Our own test for significance of changes and other tests regarding the equality of levels were applied to verification of that assumption. None of the tests have proved that the changes in the assessment methodology would lead to statistically significant changes in researchers' behaviour in the sense of a stronger focus on results published in Web-of-Science- and Scopus-monitored journals in the period of 2008–2012 (Methodology 2013) as compared with 2005–2009 (Methodology 2010).

References

ARNOLD, E. et al. *Metodika hodnocení ve výzkumu a vývoji a zásady financování* [*R&D Evaluation Methodology and Funding Principles*]. Prague: MŠMT, 2015.

BLATNÁ, D. *Neparametrické metody* [*Non-parametric methods*]. Prague: University of Economic, 1996.

HINDLS, R. AND HRONOVÁ, S. How Much Are Changes in Attitudes Significant over Time? In: *ISI 2007*, Lisabon: International Statistical Institute, 2007.

HINDLS, R. AND HRONOVÁ, S. Odraz ekonomického vývoje vybraných zemí ve struktuře výdajů na konečnou spotřebu [Reflection of Economic Development of Selected Countries in the Structure of Final Consumption Expenditure]. *Politická ekonomie*, 2012, Vol. 60, No. 4, pp. 425–442.

HINDLS, R., HRONOVÁ, S., SEGER, J., FISHER, J. *Statistika pro ekonomy* [*Statistics for economists*]. 8th Ed. Prague: Professional Publishing, 2007.

JURAJDA, Š., KOZUBEK S., MUNICH, D., ŠKODA, S. *Mezinárodní srovnání kvality publikačního výkonu vědních oborů v České republice* [*International comparison of publication performance of sciences in the Czech Republic*]. Prague: CERGE-EI, Studie 12/2015.

MUNICH, D. AND ŠKODA, S. *Světové srovnání českých a slovenských časopisů podle indikátorů Impact Factor (IF) a Article Influence Score (AIS)* [*Worldwide comparison of Czech and Slovak journals according to the Impact Factor (IF) and Article Influence Score (AIS) indicators*]. Prague: CERGE-EI, Studie 19/2016.

VANĚČEK, J., FAŤUN, M., PAZOUR, M. *Srovnávací studie vybraných metodik hodnocení výzkumu a vývoje* [*A comparative study of selected methodologies for assessment R&D*]. Prague: Technologické cetrum AV ČR, 2008.

Závěrečná zpráva mezinárodního auditu výzkumu, vývoje a inovací v České republice [*Final report of international audit of research, development and innovations in the Czech Republic*]. Prague: MŠMT, 2012.

Table 1A Number of articles and results

University	Jwos M2010	Jwos M2013	Jsc M2010	Jsc M2013								
	x_{1i}	x_{2i}	y_{1i}	y_{2i}	x_{1i}^2	x_{2i}^2	y_{1i}^2	y_{2i}^2	h_i	V_i	k_i	$(k_i - k_{aver})^2$
Czech Technical University, Prague	1 426	1 685	325	499	2 033 476	2 839 225	105 625	249 001	312	949 125	312	39 823
Czech Univ. of Life Sciences, Prague	469	646	350	515	219 961	417 316	122 500	265 225	242	340 080	242	16 774
USB, České Budějovice	827	948	189	206	683 929	898 704	35 721	42 436	122	221 490	122	95
Masaryk University, Brno	2 413	2 351	1 406	1 226	5 822 569	5 527 201	1 976 836	1 503 076	190	–769 128	–190	91 713
Mendel University, Brno	341	471	507	992	116 281	221 841	257 049	984 064	502	832 575	502	151 833
University of Ostrava	183	328	96	207	33 489	107 584	9 216	42 849	183	107 728	183	4 920
Silesian University, Opava	125	154	80	58	15 625	23 716	6 400	3 364	36	5 055	36	5 786
Technical University, Liberec	162	202	77	186	26 244	40 804	5 929	34 596	116	43 227	116	13
University of Pardubice	778	740	168	211	605 284	547 600	28 224	44 521	57	–41 387	–57	28 848
University of Hradec Králové	30	60	48	148	900	3 600	2 304	21 904	104	22 300	104	65
J. E. Purkyně University, Ústí nad Labem	103	147	56	67	10 609	21 609	3 136	4 489	45	12 353	45	4 504
Charles University, Prague	7 751	7 117	3 936	3 182	60 078 001	50 651 689	15 492 096	10 125 124	985	–14 793 284	–985	1 204 701
Palacký University, Olomouc	1 350	1 760	654	1 009	1 822 500	3 097 600	427 716	1 018 081	542	1 865 465	542	184 788
Tomas Bata University, Zlín	249	374	61	291	62 001	139 876	3 721	84 681	262	158 835	262	22 293
University of VPS, Brno	553	549	109	108	305 809	301 401	11 881	11 664	4	–4 625	–4	13 592
Technical University, Ostrava	267	527	103	379	71 289	277 729	10 609	143 641	379	339 472	379	71 137
University of Economics, Prague	159	202	67	105	25 281	40 804	4 489	11 025	57	22 059	57	3 034
Univ. of Chemistry and Technology, Prague	1 410	1 386	125	143	1 988 100	1 920 996	15 625	20 449	30	–62 280	–30	20 296
University of Technology, Brno	702	941	438	835	492 804	885 481	191 844	697 225	463	898 058	463	123 149
West Bohemia University, Pilsen	348	462	123	219	121 104	213 444	15 129	47 961	149	125 172	149	1 338
Total	**19 646**	**21 050**	**8 918**	**10 586**	**74 535 256**	**68 178 220**	**18 726 050**	**15 355 376**	**4 783**	**–9 727 710**	**2 249**	**1 988 702**
$t_{0,95}(19)$	2 093											
T	1 555											

Valuation of Volunteer Work in Satellite Account of Non-Profit Institutions

Václav Rybáček[1] | *Jan Evangelista Purkyňe University, Ústí nad Labem, Czech Republic*
Jitka Fořtová[2] | *Czech Statistical Office, Prague, Czech Republic*
Šárka Skaláková[3] | *Czech Statistical Office, Prague, Czech Republic*

Abstract

Volunteer work constitutes an important input into the activities of non-profit institutions. However, in the core system of national accounts, volunteering falls outside the production boundary even if it leads to the production of services. By doing so, national accounts inevitably underestimates the contribution of non-profit institutions to the well-being. This shortcoming is overcome by the Satellite Account of Non-profit Institutions complementing and extending the concept of national accounts chiefly by incorporation of the value of volunteering and by full coverage of non-profit institutions classified in a number of economic sectors. This paper is an attempt to address the key issue that is the way of volunteer work´s valuation for analytical purposes. We will discuss different approaches to the valuation and their impact on key macroeconomic aggregates.

Keywords

Non-profit institutions, volunteer work, valuation

JEL code

E23, E24, J30

INTRODUCTION

The Satelite account of non-profit institutions (hereinafter "SANPI") has been enjoying a growing attention of social and economic policy interest. One of the major reason is that non-profit institutions may be seen as a supplement to general government in pursuit of economic and particularly social policies; they are often reffered to as "third sector" or "civil society". In general, activities of non-profit institutions represent a form of social entrepreneurship (Boettke and Coyne, 2009) whose aim is to create a social value (Boschee, 1997). Depending on the number and economic strenght, non-profit institutions (hereinafter "NPI´s") may be a significant economic force in supplying private as well as public goods and services throughout the world.

The institutional background of NPI´s deserves a special attention as they operate in a different incentive schemes compared to private market producers or a majority of government institutions. First of all, non-profit institutions are not legally permitted to distribute profits which must be retained and used for their activity. An owner or founder is thus not motivated to maximise profit for the purpose

[1] Also the Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: vaclav.rybacek@czso.cz.
[2] Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: jitka.fortova@czso.cz.
[3] Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: sarka.skalakova@czso.cz.

of paying dividends. On the other hand, NPI´s might be closely linked to a profit-oriented organisations, as is the case of many foundations or charities established by market producers (Boettke and Coyne, 2009).

The impossibility to distribute a profit might pose a limit on the ways of funding as main objectives are assisting needy people or acquiring reputation. NPI´s are thus in a measure funded by voluntary contributions having forms of monetary payments, gifts in kind, but also volunteer time. Time contribution brings us to the key issue which is the labour input employed into the operation of NPI´s. To a large extent, the labour input takes the form of unpaid volunteer work, it is not, by convention, considered as source of value added in the core accounts of national accounts. Such conventions gave raise to the need for a supplementary datasets reflecting specific features, as is the case of satellite accounts.[4]

To overcome the simplified convention in the core national accounts, the Satellite account regards unpaid work as a source of value added.[5] Obviously, a number of important questions remains open when it comes to the way of volunteer work valuation. This issue has been widely adressed in economic research. The wage-based valuation has become the most-commonly used technique when valuating volunteer work (Brown, 1999). For example, Salamon, Sokolowski and Haddock (2011) apply the replacement cost method using observed market wages. From available resources, the same method is used for practical compilation by the statistical office in New Zeland (Statistics New Zealand, 2007), etc. However, a number of objections can be raised.

As Brown (1999) claims it does not reflect the willingness of recipient to pay for this service if not donated. According to Brown, this makes the value for recipient overstated while it understates the gains of volunteers themselves (Brown, 1999). Similarly to Brown, Bowman (2009) is critical of the application of replacement costs or demand price. As the author argues, the value should be rather expressed through its impact on the revenues of an organization (Bowman, 2009). To provide an exhaustive discussion goes beyond the scope of this text. The purpose was purely a brief demonstration of differentiating views which may have a significant impact on macroeconomic indicators. To illustrate this, Pho (2008) found out that the value of volunteer work in the US varied between 0.9 and 1.3 of the gross domestic product (thereinafter "GDP") in 2005.[6] It is apparent that quite a lot of research remains to be done.

Valuation of volunteer work is, of course, a subject of interest of not only researches but also organisations promoting the general interests of its members. E.g. the Independent Sector operating in the US publishes the historical time series of the generally accepted value of volunteer work currently standing at 24.14 dollars per hour.[7] The estimation basis is the hourly earnings of workers on private non-farm payrolls average published by the Bureau of Labor Statistics. According to the Financial Accounts Standard Board (1993), the value of volunteer work can be recognized in the financial statement under the condition that

---

4 Satellite accounts are recognised internationally as a way of rearranging existing informations in national accounts. They can cover a wide range of areas as agriculture, enviroment, health, etc. The new generation of the manuals (SNA2008 and ESA2010) gives far more space to this areas of macroeconomic statistics compared to their predecesors (SNA1993 and ESA1995). In the ESA2010, Chapter 22 is exclusively devoted to the issue of the satellite accounts compilation. For the purpose of the SANPI compilation, the United Nations (UN) published the Handbook of Non-Profit Institutions in the System of National Accounts recommending appropriate statistical procedures for the data compilation.

5 Further important contribution of the Satellite account is the coverage since NPI´s are not covered in the basic sector schemes in their entirety. In other words, there is no single sector containing all NPI´s operating in the economy.

6 In 2012, 64.5 million Americans volunteered nearly 7.9 billion hours with an estimated value of nearly $175 billion, it is 1.08% of the GDP. In New Zealand, non-profit institutions´economic contribution is $6 billion (2.7% of the total) to the GDP for the year ended March 2013. When the value from the labour of volunteers ($3.5 billion) is included, non-profit institutions contributed $9.4 billion (4.4%) to total GDP. In Norway, non-profit institutions are estimated to have contributed NOK 53 billion (1.7%) to the GDP. Including the value of unpaid work, the total value added in the non-profit sector accounted for around NOK 125.5 billion (3.9%) of the GDP in 2014. In the Czech Republic, non-profit institutions contributed 69,5 mil CZK (1.5%) of the GDP in 2014. Including the value of unpaid work CZK 5.8 mil the total contribution is CZK 75.3 mil (1.6%) to the GDP.

7 *<https://www.independentsector.org/resource/the-value-of-volunteer-time>* [downloaded: 1.8.2017].

the service would have been purchased if it had not been donated. The valuation technique should use relevant market prices of labor by occupation.

The purpose of this paper is to bring a modest constribution to the aforementioned discussion on the volunteer work valuation. We will address the impact of different approaches on the final figures and the relevance of several methods from the theoretical and the practical point of view. The text is based on the practical experience with the compilation of the SANPI in the Czech Republic. The Czech Statistical Office publishes the SANPI in the autumn of each year following the publication of the sector accounts at the end of June. The release covers the year before last, i.e. in October 2016, the SANPI published covers the year 2014 and previous years.

1 METHODOLOGY

To begin with, we present the relevant definitions and the classification issue. The restated structural-operational definition of non-profit institutions is used for the purposes of statistical monitoring of non-profit institutions in the Satellite Account of Non-profit Institutions. According to this definition from the Handbook of Non-Profit institutions (UN, 2003), the Satellite Account of Non-profit Institutions consists of economic units which are:

- *organizations*, i.e. they have a certain institutionality, are legal entities with a certain degree of internal organizational structure,
- *non-profit or non-profit-distributing*, i.e. any generated surpluses are used for the main object of activities which the non-profit institution was established for,
- *institutionally separated from government institutions*, i.e. they are not part of the government apparatus or delegated to exercise state power,
- *self-governing*, i.e. they are able to manage their activities and create their organizational structure,
- *optional, i.e.* their formation, activity and membership in them is voluntary.

The main reason for establishing non-profit institutions is either voluntary or charitable activity, or the effort to support certain groups of people in business, politics, or other areas of social life. Compilation of national accounts requires working with a number of classifications in the Business Registry such as sector classification, classification of branches or, most importantly, so called legal forms. Legal forms reflect the mode of operational functioning of different kinds of units. The delimitation of the sphere composed of NPI´s is thus based on the legal forms codes. In 2014, the definition of non-profit institutions meet the following legal forms:

Table 1 Number of NPIs included into satellite account by legal form for the year 2014

Code	Title	NPI´s	NPI´s - nonfinancial and financial institutions	NPI´s - government institutions	NPISH´s
TOTAL		**129 061**	**776**	**28**	**128 257**
117	Foundation	490	:	:	490
118	Endowment Fund	1 331	:	:	1 331
141	Generally beneficial company	2 867	157	:	2 710
161	Institute	142	:	:	142
601	Public university	26	:	26	:
641	School corporation	236	4	:	232
703	Trade union and employers' organizations	701	6	:	695
704	Special organization[8]	16	:	:	16

[8] For representation of Czech interests in international non governmental organizations.

Table 1 Number of NPIs included into satellite account by legal form for the year 2014 (continuation)

Code	Title	NPI´s	NPI´s - nonfinancial and financial institutions	NPI´s - government institutions	NPISH´s
706	Society	82 778	181	:	82 597
711	Political party, political movement	233	:	:	233
721	Church organisation	4 117	:	:	4 117
733	An organizational unit of a trade union and employers' organizations	5 777	:	:	5 777
736	Branch of society	24 761	22	:	24 739
741	Professional organization/chamber	22	:	:	22
745	Other chamber (excl. professional ones)	207	207	:	:
751	Association of legal persons	1 201	199	2	1 000
761	Hunting community	4 156	:	:	4 156

Source: Czech Statistical Office, <*www.czso.cz*>

The structure is considerably impacted by existing legislation. In 2014, a massive wave of transformation of non-profit institutions took place following the new civil code entering into force. Among others, associtions were replaced by the legal form of "societies", trade unions and employers´organizations were separated, generally beneficial societies can be no longer founded and existing entities should be transformed into institutes, foundations of endowment funds. The same holds true for associations of legal persons, even if they continue its activity.

Societies and branches of societies represent the largest group of non-profit institutions. According to relevant legislation which entered into force in 2013, NPI´s are not obliged to report the cessation of their activity. Total numbers of units is thus inevitably overestimated whereas the extent of an overestimation might range from one third to one half of total number. Following the System of National Accounts, the SANPI presents the national economy classified by individual institutional sectors according to the producer type, and by individual industries according to the product type. As shown in Table 1, non-profit institutions are included not only in the sector of Non-profit institutions serving households (S.15 – thereinafter "NPISH´s") but a number of them is included in the institutional sector of non-financial corporations (S.11), financial corporations (S.12) and in the general government sector (S.13), which is the case of public universitites and health insurance companies.

Concerning the data sources for the NPI´s, an exhaustive annual statistical survey is conducted for units with 10 and more employees. Units with 0–9 employees are surveyed once in five years, whereas each year a certain legal form is picked to be a subject of survey (or group of legal forms). Data for units with 0–9 emploees which are not surveyed in given year are grossed up.

2 FUNDING OF NPISH´S AND VOLUNTEER WORK

NPI´s are usually funded differently from other economic sectors. Except for the monetary revenues, volunteer work constitutes an important input. For the sake of the argument, we will firstly take a look at the structure of the monetary resources and the sectoral structure of contributors or donators. Since the relevant breakdown is available for the NPISH´s sector only, we will concentrate on this sector for now. Though, the explanatory power is not much undermined by doing so, because the NPISH´s sector plays a crucial role in the SANPI.

NPISH´s (S.15), as well as NPI´s in their entirety, are funded differently from other economic sectors. While NPISH´s can similarly as other sectors raise revenues from selling its own products or from property

income, about 61.7% of the total income comes from other sectors in form of current transfers. These transfers are recorded under the item D.751 (Current transfers to NPISH´s); for other sectors, given transfers are covered by the item D.759 (Other current transfers).The largest transfers came from the government sector (about 50%), the contribution of households to non-profit institutions reached 32% in 2014. On the top of these, the NPISH´s collect membership fees and they normally receive donations from other economic sectors, including non-financial and financial corporations. NPISH´s may obtain funds from non-profit organisations themselves (especially foundations). Because the sector is consolidated, the amount of these revenues can not be determined.

A very specific source of input into operation of NPI´s recorded on the resources side is a contribution of volunteering. Volunteering concerns not only households, but also corporations. Mentioning the work of volunteering brings us to the key question, how the work of volunteers should be valued? The evaluation of volunteer work and its inclusion into the accounts represents an important step beyond the standard framework of national accounts. Unpaid volunteer work does not fall within the production border as defined by the methodology, however, it is unquestionably an important input into the activities of non-profit institutions. Disregarding the volunteer work leads to underestimation of the actual contribution of non-profit institutions to the welfare of the society.

Here, the term volunteer means a person who is not in an employment relationship with an economic entity as regards the respective voluntarily done activity and performs his or her activity without any financial or other remuneration or legal entitlement (including any entitlements arising from obligations of the entity's members according to the statutes or other resolutions adopted by the economic entity). Voluntary workers may be volunteers performing work for an economic entity, on volunteer service, as well as other persons performing work in an organisation without entitlement to remuneration (unpaid members of administrative and control bodies, members of an economic entity and other persons).

It remains valid that it is not possible to establish the number of inhabitants of the Czech Republic performing volunteer work for non-profit institutions on the basis of source data. This is due to the fact that one person can perform volunteer work for several non-profit institutions. Therefore, the number of volunteers is given as a number of natural persons converted on the basis of the number of hours worked by volunteers (full-time equivalent approach, thereinafter "FTE"), it means 26414 FTE in 2014. The final surveyed figures are presented in the following table:

Table 2 Number of volunteers and hours worked in particular kind of NPI (2011–2014)

	Non-profit institutions according to number of employees	Year	Number of volunteers	Hours worked by volunteers	Average of hours worked per 1 person
0–9 employees	Churches	2011	20 207	1 309 148	65
	Society/association	2012	608 693	30 465 085	50
	Branch of society/ Organizational components of associations	2012	300 958	8 633 959	29
	Generally beneficial societies/ institutes	2013	3 820	215 883	57
	Foundation, Endowment fund	2014	4 243	241 681	57
	Others	2015	8 860	201 851	23
More than 10 employees	Churches	2014	11 012	198 108	18
	Society/association	2014	194 615	15 538 884	80
	Branch of society/ Organizational components of associations	2014	1 440	52 520	36
	Generally beneficial societies/ institutes	2014	4 663	209 367	45
	Foundation, Endowment fund	2014	19	1010	53
	Others	2014	633	40 325	64

Source: Fořtová (2017)

3 VALUATION OF VOLUNTEER WORK IN THE CZECH REPUBLIC

In December 2005, the Working Group for the Implementation of the Satellite Account of Non-profit Institutions in the Czech Republic held a seminar on the issues related to the valuation of volunteer work. After having heard various views and proposals, they decided to adopt the method of valuation by means of the median determined on the basis of the results obtained from the Average Earnings Information System (thereinafter "ISPV"), which is carried out by the Statistical Services Department of the Ministry of Labour and Social Affairs, on salary and remuneration for stand-by duty in budgetary and certain other organizations and bodies. In 2007, the Satellite Account of Non-profit Institutions was first compiled for 2005, including the imputed value of volunteer work. For the year 2012, the valuation of volunteer work was reassessed.

The valuation method using the median determined on the basis of results from the ISPV has been preserved. After having heard different opinions of users on the level of the median and processing of analyses themselves, the decission of the Czech Statistical Office was to use the median value of wages in the Czech Republic which corresponds to the salaries in the non-profit sector more than the median value of the salary in the Czech Republic.

For the year 2014, the median value of salaries in the Czech Republic according to the ISPV reached CZK 127.24 /hour. The number of hours worked by volunteers, that the Czech Statistical Office obtained from the statistical surveys by means of the questionnaires NI 1–01 (a), was multiplied by this median. The following table lists the valuation of volunteer work for non-profit institutions and the median value of wages in the Czech Republic for the years 2005 to 2014.

Table 3 The valuation of volunteer work for non-profit institutions in total from 2005 to 2014

Year/ sector	Number of hours volunteered				Valuation of volunteer work (CZK million)			
	in S.11, S.12	in S.13	in S.15	Total	in S.11, S.12	in S.13	in S.15	Total
2005	168 872	0	62 819 667	**62 988 539**	17	0	6 219	**6 236**
2006	233 680	0	48 650 387	**48 884 067**	25	0	5 152	**5 177**
2007	288 483	0	82 937 006	**83 225 489**	33	0	9 396	**9 429**
2008	530 707	0	46 674 947	**47 205 654**	64	0	5 602	**5 666**
2009	287 072	0	46 890 116	**47 177 188**	35	0	5 734	**5 769**
2010	299 531	0	44 021 402	**44 320 933**	37	0	5 479	**5 516**
2011	292 602	0	44 892 904	**45 185 506**	37	0	5 634	**5 671**
2012	180 088	116	44 686 130	**44 866 334**	23	0	5 648	**5 671**
2013	184 448	2 436	43 579 217	**43 766 101**	24	0	5 509	**5 533**
2014	106 118	2 132	45 499 777	**45 608 027**	14	0	5 789	**5 803**

Source: Notes on Satellite account of NPI´s, CZSO (2016a)

The evaluated volunteer work is recognized in the Satellite Account of Non-profit Institutions as part of the Wages and Salaries (D.11) item. The increase in item D.11 is reflected in the change to the total remuneration of employees (D.1) and in the balance items (the operating surplus, disposable income, net savings, net loans, and other). For S.15, the non-market output (P. 132), which is calculated using the cost method, increases correspondingly.

4 ALTERNATIVE APPROACHES TO THE VALUATION OF VOLUNTEER WORK

As mentioned above, the wage-based valuation is widely used but not a single method. A practical appropriatness and readily accesible data for valuation do not imply a conceptually correct and the most appropriate measure. At least two questions should raise in this case. What kind of data is available

for the estimation of volunteer work? Will using alternative method lead to rather different outcomes? To bring the evidence from the Czech economy, Table 4 represents the results of alternative methods.

Table 4 Alternative approaches to valuation of volunteer work, the Czech Republic, CZK mill.

	Valuation of volunteer work (CZK mill)				
Institutional sector	**1. Hour wage median for the CR (ISPV)**	**2. Minimum wage**	**3. Hour wage median for the CR (ISPV) - by NACE**	**4. Hour wage median for the CR (ISPV) - by type of work**	**5. Hour wage median for the CR (ISPV) - salaries**
Non-financial corporations	14	5	10	16	15
Financial corporations	0	0	0	0	0
Government institutions	0	0	0	0	0
Non-Profit institutions serving households	5 789	2 302	5 058	6 079	6 629
Total	5 803	2 307	5 068	6 095	6 644

Source: <*www.czso.cz*>, <*www.ispv.cz*>, own calculation

Table 4 shows the results of five different methods. The first method, which uses the hour wage median of wages according to the IPSV, represents the currently used approach in the Czech Republic as was described above. The second method uses the minimum wage set by government. Using the legally set minimum wage, we arrived at much lower value of volunteer work. The evident disadvantage of this approach subsists in the general nature of the minimum wage which is oriented on manual work. However, volunteer work consists largely in expert work as social assistence, accounting, etc.

The third method incorporates data on volunteers work by the statistical Classification of Economic Activities in the European Community (thereinafter "NACE") in combination with the median of hour wage by the NACE codes from the ISPV. The advance is the availabilty of data by the NACE classification, however, the NACE classification is not entirely appropriate as the units are classified by their prevailing activity here. Nevertheless, the NPI´s normally carry out more than one activity, as well as workers volunteeering in these institutions. The classification COPNI[9] (Classification of Services of Non-profit Institutions Serving Households by Purpose) seems to be much more suitable for this purpose. However, the problem with data availability usually occurs, i.e. structure of wages by COPNI, which is also currently the case in the Czech Republic.

The fourth method using hour wage median by type of work seems to be very appropriate; it is used in some countries (Poland, New Zealand) depending on the data availability. In the Czech case, the ISPV publishes data on median wage by type of work, even in more detailed breakdown. However, the data for NPI´s are not very detailed providing very rough structure of work types (three groups), as shown in Table 5.

Table 5 Number of hours and valuation of volunteer work by type, the Czech Republic, 2014

Value of volunteer work by type	**Hours worked**	**Valuation (CZK mill)**
Managers and Professional mental work	14 994 999	2 729
Clerical support workers, Services and Sales workers, Craft and related trades workers	24 731 614	2 703
Plant and machine operators and assemblers, Elementary occupations	5 879 282	663

Source: <*www.czso.cz*>, <*www.ispv.cz*>, own calculation

[9] COPNI is surveyed since 2012 in the Czech Republic. It is intended for the monitoring of the purpose which the funds of NPI´s were spent on. This data is supposed to better describe actual spheres in which NPI´s are active.

The last method incorporates hour wage median by salaries as provided by the ISPV. As mentioned above, this method does not correspond to wages in the sector of NPISH´s because the salaries, i.e. the compensations of civil servants, are bigger than wages of workers in the corporate sectors. This calculation provides the total value of volunteer work by CZK 800 bill. higher compared to the first method, which is used now.

To compare the results in relative terms, the following table summarises the impact of alternative method on the gross value added (GVA) of the NPISH´s (S.15) and in the SANPI. The currently applied method led to an increase in GVA of NPISH´s by 17% and of SANPI 7.2% in 2014. Except for the second method, which seems to be clearly inappropriate, the other method working with structural indicators of work or salaries gave results not significantly different from the method applied in the compilation of the SANPI in the Czech Republic. For the third method, the impact on GVA would be only by 1.8 in NPISH and 0.9 in SANPI percentage point lower. In case of the fourth method, which seems to be the most appropriate from the theoretical point of view, the impact on GVA would go up by 0.7 in NPISH and 0.3 in SANPI percentage point.

Table 6 An impact of alternative approach to GVA, differences in the average wage, the Czech Republic, 2014

	1. Hour wage median for the CR (ISPV)	2. Minimum wage	3. Hour wage median for the CR (ISPV) - by NACE	4. Hour wage median for the CR (ISPV) - by type of work	5. Hour wage median for the CR (ISPV) - salaries
Impact to GVA of NPISH (S.15)	17.0%	7.6%	15.2%	17.7%	19.0%
Impact to GVA of SANPI	7.2%	3.0%	6.3%	7.5%	8.1%

Source: <*www.czso.cz*>, <*www.ispv.cz*>, own calculation

CONCLUSION

The paper discussed the issue of the volunteer work valuation for the sake of the compilation of the SANPI. The specific nature of the NPI´s requires to add the value of volunteer work in the relevant aggregates describing the economic behaviour of this segement of the economy. We discussed alternative method which can be used for this purpose. Final choice is usually affected by data availability. As shown in the table 6, except for one method, the methods did not result in significantly different outcomes. At the same time, the valuation of work by type seems to be the most appropriate but data-intensive method. However, using this method based on rough structure as provided by the ISPV, we arrived at the results not significantly different compared to the currently applied method which seems to be sufficient. But it does not rule out further improvements when it comes to relevant data sources and a choice of method to make the result more precise.

References

BOETTKE, P. AND COYNE, CH. Context Matters: Institutions and Entrepreneurship. *Foundations and Trends in Entrepreneurship*, 2009, Vol. 5, No. 3, pp. 135–209.

BOSCHEE, J. What does it take to be a social entrepreneur? *Not-For-Profit CEO Monthly Letter*, 1997, 4(6), pp. 1–3.

BOWMAN, W. The economic value of volunteers to nonprofit organizations. *Nonprofit management and leadership*, 2009, Vol. 19, Iss. 4, pp. 491–506.

BROWN, E. Assessing the value of volunteer activity. *Nonprofit and Voluntary Sector Quarterly*, 1999, 28(1), pp. 3–17.

Europan System of National and Regional Accounts (ESA2010). Eurostat, 2014.

FINANCIAL ACCOUNTING STANDARD BOARD. *Statement of Financial Accounting Standards No. 116* [online]. June 1993. <http://www.fasb.org/jsp/FASB/Document_C/DocumentPage?cid=1218220124001&acceptedDisclaimer=true>.

FOŘTOVÁ, J. Dobrovolníci v Česku [online]. *Statistika&My*, Prague: Czech Statistical Office, March 2017. <http://www.statistikaamy.cz/2017/03/dobrovolnici-v-cesku>.
INDEPENDENT SECTOR. *The value of Volunteer Time* [online]. May 2016. <https://www.independentsector.org/resource/the-value-of-volunteer-time/>.
ISPV. *Wage sphere – year 2014* [online]. May 2016. <ispv.cz/getattachment/c81deb46-7d3f-4d37-acc8-98ee05ae3ed3/CR_144_MZS-pdf.aspx?disposition=attachment >.
PHO, Y. The value of volunteer labor and the factors influencing participation: evidence for the United States from 2002 through 2005. *The Review of Income and Wealth*, Vol. 54, Iss. 2, pp. 220–236.
CZSO. *Notes on Satellite account of NPI´s* [online]. Prague: Czech Statistical Office, 2016a. <https://apl.czso.cz/nufile/SUNI_2005_2014EN.pdf>.
CZSO. *Structure of satellite account of non-profit institutions* [online]. Prague: Czech Statistical Office, 2016b. <http://apl.czso.cz/pll/rocenka/rocenkavyber.satelit_en>.
SALOMON, L., SOKOLOWSKI, W., HADDOCK, M. Measuring the Economic Value of Volunteer Work Globally: Concepts, Estimates, and a Roadmap to the Future. *Annals of Public and Cooperative Economics*, 2011, Vol. 82, Iss. 3, pp. 217–252.
STATISTICS NEW ZEALAND. *Non-Profit Institutions Satellite Account* [online]. 2007. <http://www.stats.govt.nz/~/media/Statistics/browse-categories/industry-sectors/non-profit-institutions-satellite-account-2004/npisa-04.pdf>.
STATISTICS NEW ZEALAND. *Non-Profit Institutions Satellite Account* [online]. 2007. <http://www.stats.govt.nz/browse_for_stats/economic_indicators/NationalAccounts/non-profit-2013-mr.aspx>.
STATISTICS NORWAY. *Volunteer work reaches nearly 148 000 full-time equivalents* [online]. 2016. <https://statbank.ssb.no/en/nasjonalregnskap-og-konjunkturer/artikler-og-publikasjoner/volunteer-work-reaches-nearly-148-000-full-time-equivalents>.
STATISTICS USA [online]. 2014. <https://www.census.gov/newsroom/cspan/2014/volunteer.html>.
System of national accounts 2008. Washington, 2009.
UNITED NATIONS. *Handbook on Non-Profit Institutions in the System of National Accounts.* New York: UN, 2003.

The Problem of the SARIMA Model Selection for the Forecasting Purpose

Josef Arlt[1] | *University of Economics, Prague, Czech Republic*
Peter Trcka[2] | *University of Economics, Prague, Czech Republic*
Markéta Arltová[3] | *University of Economics, Prague, Czech Republic*

Abstract

The goal of the work is to assess the ability to identify the proper models for the time series generated by SARIMA processes with different parameter values and to analyze the accuracy of the forecasts based on the selected models. The work is based on the simulation study. To this end, a new automatic SARIMA modelling method is proposed. Other competing automatic SARIMA modelling procedures are applied as well and the results are compared. The important question to which the reference should be made is the relation of the magnitude of the SARIMA process parameters i. e. the size of the systematic part of the process and the ability to identify a proper model. Another issue addressed herein is the relationship between the quality of the identified model and the accuracy of forecasts achieved by its application. The simulation study leads to the results that can be generalized to most empirical analyses in various research areas.[4]

Keywords

SARIMA, simulation, identification of model, forecasting

JEL code

C15, C22, C63

INTRODUCTION

The principle and the application of the SARIMA models in the time series modelling has been well known for many years. Its practical applications can be found in many areas where empirical analyses are needed and it has become a basis standard tool of modern econometric analysis. The crucial phase of the practical application of the Box-Jenkins methodology is the identification and verification of the suitable model.

The goal of this article is to find the time series for which it is relatively easy to identify the proper model and the time series for which it is difficult. Another goal is to analyze the forecasting abilities of the SARIMA models for different kinds of time series. A convenient way to verify the aforementioned is the simulation study and the application of the automatic SARIMA procedures.

[1] Department of Statistics and Probability, nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. Corresponding author: e-mail: josef.arlt@vse.cz
[2] Department of Statistics and Probability, nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: trcp00@vse.cz
[3] Department of Statistics and Probability, nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: marketa.arltova@vse.cz
[4] This article is based on contribution at the conference *ASMDA 2017* – 17th conference of the Applied Stochastic Models and Data Analysis International Society and the Demographics 2017 Workshop.

The article is divided into four sections (excluding the Introduction). In the first section the SARIMA models are briefly described. In the second section, the simulation study as well as the Auto.SARIMA and Auto.AIC procedures for automatic model selection are explained. The results of the simulation study are the subject of the third section. The fourth section contains the conclusion, along with the summary of the work.

1 SARIMA MODELING AND FORECASTING

The ARMA(p, q) proces (Auto-Regressive-Moving-Average proces of orderes p, q) is defined as $\phi(B)y_t = c + \theta(B)a_t$, where B ($B^j y_t = y_{t-j}$) is the backshift operator and $\phi(B)$ and $\theta(B)$ are the polynomials in the lag operators of the order p and q respectively, $\{a_t\}$ is the white noise process. It is stationary, if the roots of the autoregressive polynomial $\phi(B)$ lie outside of the unit circle and it is invertible if the roots of the moving average polynomial $\theta(B)$ lie outside of the unit circle.

The SARMA(p, q)(P, Q)s proces (Seasonal ARMA process of orders p, q, P, Q) can be written in the form $\phi(B)\Phi(B^s)y_t = c + \theta(B)\Theta(B^s)a_t$, where s is the number of seasons (usually 4 or 12) and $\Phi(B^s)$ and $\Theta(B^s)$ are seasonal polynomials in the lag of the order P and Q respectively. It is denoted as SARMA(p,q)(P,Q)s. If the roots of all polynomials lie outside of the unit circle, the proces is stationary and invertible.

The special form of the non-stationary proces is the so called integrated process („I" in acronym). Such a proces is stationary after some degree of differencing. The SARIMA(p,d,q)(P,D,Q)s proces is the general form of the integrated proces and can be written as $\phi(B)\Phi(B^s)\Delta^d\Delta_s^D y_t = c + \theta(B)\Theta(B^s)a_t$, where $\Delta^d = (1 - B)^d$ is the nonseasonal difference of the order d and $\Delta_s^D = (1 - B^s)^D$ is the seasonal difference of the order D.

The forecasting of the future values of the time series is an important role of the SARIMA modelling. The optimal forecast, i. e. the forecast with the minimum mean square error, is the conditional mean of the future random variable, which is conditioned on the historical information available in the observed values of the applied time series.

The SARIMA time series modelling methodology has been well known for many years and there exists a vast amount literature devoted to this topic, *inter alia*, Box, Jenkins, Reinsel and Ljung (2015), Brockwell and Davis (2010), Wei (2005), Hamilton (1994), Enders (2014), Pesaran (2016).

2 SIMULATION STUDY

The goal of the simulation study is to analyze the relationship of the magnitude of the SARIMA process parameters; i. e. the size of the systematic part of the process, which is used for time series generation and the ability to select the proper model for the generated time series. This question is general in scope, and the qualified and substantiated answers can be important for the empirical analyses in the different fields of the research. Another goal is to analyze the quality of the forecasts for the time series generated by the processes with different systematic parts. Important is also the analysis of the ability to select suitable model and reach the relatively accurate forecasts for the time series generated by the near non-stationary and the non stationary processes.

In the simulation study the results of the two automatic procedures for SARIMA model selection and forecasting are presented. The first one is based on the classic model selection process, i.e. the model identification, the parameters estimation, the diagnostic controll (on the basis of the residual time series, the autocorrelation, the heteroscedasticity as well as the normality are tested). The second one is based on the minimization of the AIC criterion (Akaike, 1974). Both procedures were implemented in the R software (2008).

2.1 Procedure Auto.SARIMA

The Auto.SARIMA is fully automated procedure, whose goal is to find the best model with respect to predefined parameters for the analyzed time series. In the first stage, the order of the nonseasonal

and the seasonal differencing, i. e. the numbers d and D, after which the analyzed time series is stationary, has been found. For the nonseasonal unit root testing, the ADF (Dickey and Fuller, 1979), the PP (Phillips and Perron, 1988) and the KPSS (Kwiatovski, Phillips, Schmidt and Shin, 1992) tests are used. The seasonal unit root is tested by the CH test (Canova and Hansen, 1995).

The procedure will analyze the quality of the SARIMA(p,d,q)(P,D,Q)s models for the given order of the nonseasonal differencing d, as well as the seasonal differencing D, and for all possible combinations of values p, q, P, Q. It is therefore possible to skip the identification stage and to estimate the parameters for all the possible model forms. After the parameters estimation, the procedure continues with the diagnostic checking, which is mainly based on the residual analysis. The statistical significance of the parameter estimates is verified by the standard t tests. The autocorrelation is assessed by the residual autocorrelation function, and the Ljung-Box test (Ljung and Box, 1978). The conditional heteroscedasticity is tested by the ARCH test (Engle, 1982). The normality is tested by the Jarque-Bera test (Jarque and Bera, 1980).

If the parameter estimates are statistically significant and the null hypotheses of no autocorrelation, no conditional heteroscedasticity and normality are not rejected, then the value 1 is assigned to the particular property (autocorrelation, heteroscedasticity, normality, parameter significance). Otherwise, the value 0 is assigned. The suitability criterion of the model is computed as the weighted average of the results of the individual tests, where the individual properties have specific weight. The final value of each model is computed as a function of the value of the model suitability criterion and the value of the AIC criterion. The system mentioned above has been proposed by Trcka (2015).

2.2 Procedure Auto.AIC

The model selection on the basis of the AIC criterion is the content of the Auto.AIC procedure. The course of the procedure can be divided into four steps. In the first step, the stationarity of the time series is analyzed. The order of differencing is determined by the same methods as in the Auto.SARIMA procedure (see part 2.1). According to the order of differencing and the SARIMA model maximal orders, the set of the possible models is generated. Furthermore, the optimization criterion is set to such value which the AIC criterion cannot reach. In the third step, the adjustments are made so that all the models lead to the same number of residuals. On the basis of the adjusted time series, the model parameters are estimated, and the value of the AIC criterion is computed. In the following step, the actual value of the AIC criterion is compared with the value of the optimalization criterion. If the model is better that the last one, i. e. if its value of the AIC criterion is smaller than the value of the optimalization criterion, then it is denoted as the optimal model and the value of the optimization criterion is updated. In this manner the whole set of possible models is checked.

2.3 Data generation

In the simulation study, the time series generated by the SARIMA proces of the first order are analyzed. This process has the following form:

$$(1 - \phi_1 B)(1 - \Phi_1 B^{12})y_t = (1 - \theta_1 B)(1 - \Theta_1 B^{12})a_t. \tag{1}$$

The basic elements for the simulations are the time series generated by the normal white noise process with the variance $\sigma_a^2 = 1$. The parameters ϕ_1, θ_1, Φ_1, Θ_1 take all possible combinations of the following values: 0.0, 0.1, 0.2, 0.3, 0.4, 0,5, 0.6, 0.7, 0.8, 0.9, 1.0 (only positive parameters are used because in the economic practice, the SARIMA models with negative parameters occur rarely). When $\phi_1 = 1$, the process is non-seasonally non-stationary, when $\Phi_1 = 1$, the process is seasonally non-stationary, when $\phi_1 = 1$ and $\Phi_1 = 1$, the process is both non-seasonally and seasonally non-stationary. When $\theta_1 = 1$, the process is non-seasonally noninvertible, when $\Theta_1 = 1$, it is seasonally

noninvertible or both, when $\theta_1 = 1$ and $\Theta_1 = 1$. Overall, the time series from 14 641 different generating processes are analyzed. Each process generates 100 time series with a length of 150 values. The time series generator was created in the R software.

3 RESULTS

The results of the simulation study are presented in a two-dimensional space, whose structure is shown in Table 1. The possible values for p, d, q, P, D, Q of the selected models are 0 or 1. The rows of table represent an ordered combination of values of the seasonal parameters Φ_1 and Θ_1 and the columns of table represent an ordered combination of values of the nonseasonal parameters ϕ_1 and θ_1. In this way the whole set of the all possible generating processes is arranged.

The table is conditionally formatted to be able to visually evaluate the results and success of the individual automatic procedures when comparing their ability to find a suitable model. This feature is referred to as quality criterion. The quality criterion can take the values in the interval from 0 to 100 and it represents the percentage success rate of the selection of the correct model by the given procedure.

The forecasts are computed as the point estimates of the conditional expectations of the future random variables. The analyzed time series with a length of 150 values, which is about 24 observations longer than the series used for model selection, is the input of this function. In the first step, the forecasts with the horizon $h = 24$ values are computed on the basis of the model estimated from the first 126 values. In the second step, the RMSE criterion is computed. The resulting RMSE value is computed as the average from the all partial RMSE values of 100 time series forecasts with a horizon of 24 values. This criterion is presented in the same way as the quality criterion.

Table 1 The Detail of Arrangement of Values in Table

	A	B	C	D	E	F	G
1							
2		XX	AR	0	0	0	0
3		SAR	SMA/MA	0	0,1	0,2	0,3
4		0	0	0,985798631	0,978848707	1,011897837	1,057034876
5		0	0,1	1,002931047	1,030341625	1,005169594	1,055862883
6		0	0,2	1,009372642	1,017211208	1,039043504	1,046414115
7		0	0,3	1,062359565	1,049209862	1,076806091	1,090812811
8		0	0,4	1,039299045	1,071779671	1,096205459	1,084753008
9		0	0,5	1,045760435	1,06291802	1,089929381	1,118008903
10		0	0,6	1,104431568	1,139155391	1,10716083	1,146048226
11		0	0,7	1,165214121	1,16525288	1,181893295	1,216899106
12		0	0,8	1,181725599	1,187485838	1,218323404	1,232730082
13		0	0,9	1,229567078	1,244361277	1,300435194	1,25567936
14		0	1	1,29739229	1,309477773	1,335482921	1,320790968
15		0,1	0	1,037194205	1,019586911	1,004268888	1,036835566
16		0,1	0,1	1,01324022	1,027804063	1,030623177	1,053817914

Source: Own construction

3.1 Quality of the selected model

First, the results of the Auto.SARIMA and the Auto.AIC procedures from the point of view of the quality criterion are presented. In the case of the time series generated by the ARIMA(1,0,1) or the SARIMA(1,0,1)$(1,0,1)_{12}$ models the conditions of "quality" for the non-seasonal parts are the following:

$$\left|\frac{\hat{\phi}_1 - \phi_1}{s(\hat{\phi}_1)}\right| < t_{0.975} \quad \text{and} \quad \left|\frac{\hat{\theta}_1 - \theta_1}{s(\hat{\theta}_1)}\right| < t_{0.975}, \tag{2}$$

where $t \sim t(T-1)$, T is length of time series. If the selected models fulfill the above mentioned conditions they are denoted as "valid" models. The analogous criteria are applied to the seasonal parts of the models.

The results with the percentage quantifications are shown in Figure 1.

Figure 1 Quality Comparison of AIC, SARIMA

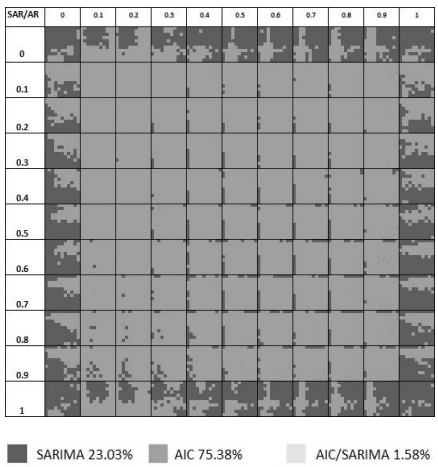
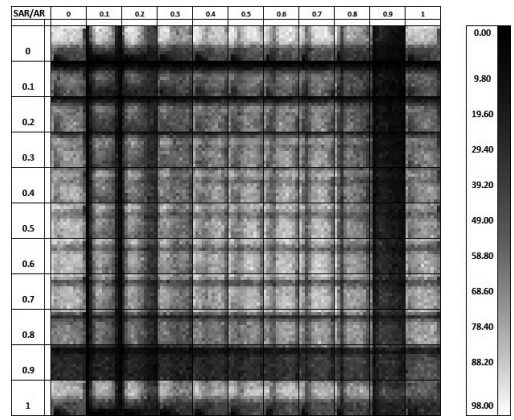


Figure 2 The Quality – Auto.SARIMA



Source: Authors' calculations

It is obvious that the Auto.AIC is better than the Auto.SARIMA in 75.4% of cases. The Auto.SARIMA achieves better results in 23% of cases. Identical results are found in 1.6% of cases. But it is clear that there is a general group of the generating processes for which the Auto.SARIMA is better than the Auto. AIC. They are mainly the seasonal and the non-seasonal non-stationary (integrated) processes, and those processes that do not contain the non-seasonal and the seasonal autoregressive parts (AR respectively SAR). Furthermore, this procedure is superior to the processes that partly do not contain the nonseasonal and the seasonal moving average parts (MA, SMA). All these processes can be denoted as marginal. The results show that, mainly there, the "classical" model identification analysis represented by the Auto.SARIMA procedure (unit root testing, residual autocorrelation testing, normality and conditional heteroscedasticity testing and parameters estimate testing) has considerable importance.

Figure 2 shows the quality criterion (the percentage of the correct model selections) for the Auto. SARIMA procedure. It can be seen that this procedure has problems with the near nonseasonal and the near seasonal non-stationary processes; i. e., for the processes with the parameters $\phi_1 = 0.9$ and $\Phi_1 = 0.9$. In the first case, the success rate is 29%, and in the second it is 22.5%. The processes with the low values of the parameters; i. e., when the parameters ϕ_1 and Φ_1 lie between 0.1 and 0.2 together with the parameters ϕ_1, and Φ_1 between 0 and 0.2, while on the contrary, the seasonally non-stationary processes, when $\Phi_1 = 1$, create more problem areas. For the proceses with parameters ϕ_1 and Φ_1 between 0.3 and 0.7, the Auto.SARIMA gives good results regardless of the values of θ_1 and Θ_1. The success rate in this area is 66.1%. The average overall success rate of this procedure is 50.6%.

Figure 3 shows the quality criterion for the Auto.AIC procedure. Also, this procedure has problems with the near nonseasonal and the near seasonal non-stationary processes. In the case of $\phi_1 = 0.9$, the success rate is 37.8%, and when $\Phi_1 = 0.9$, the rate is 34%. The problematic areas are also for $\phi_1 = 0, 1$ and $\Phi_1 = 0, 1$, together with practically any values of parameters θ_1 and Θ_1. For the processes with parameters ϕ_1 and Φ_1 between 0.1 and 0.8, the Auto.AIC gives good results regardless of the values of θ_1 and Θ_1. The success rate in this area is 82.8%. The average overall success rate of this procedure is 66.7%. In comparison with the Auto.SARIMA, the Auto.AIC procedure is better.

Figure 3 The Quality – Auto.AIC

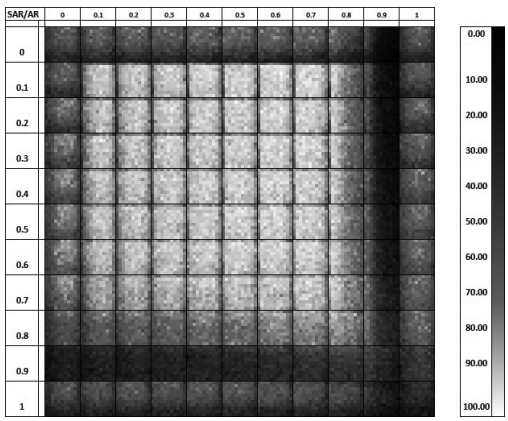
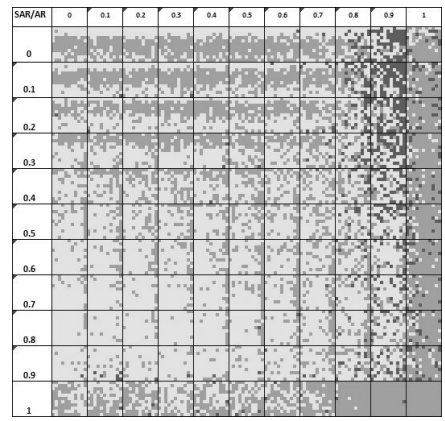


Figure 4 RMSE –1% tolerance



Source: Authors' calculations

3.2. Forecasts

The forecasts RMSE criterion is presented in the same way as the quality criterion. For each generating process, the procedure, which gives the the minimal value of the forecast RMSE, has been selected.

Figure 5 RMSE – Auto.AIC

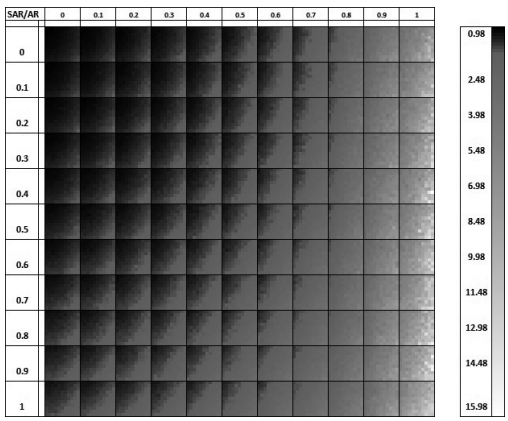
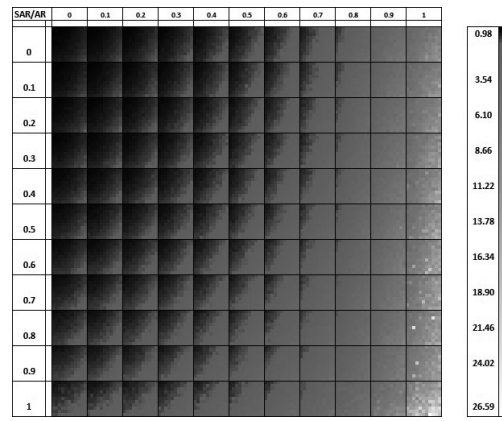


Figure 6 RMSE – Auto.SARIMA



Source: Authors' calculations

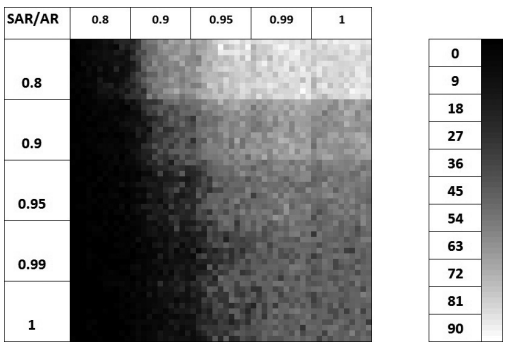
As the differences in the RMSE for the Auto.AIC and the Auto.SARIMA procedures are often very small, and the forecasts are very similar, it is suitable to compare them based on the tolerance limit of 1.0%. It means that the forecasts which are different in the RMSE up to 1.0% will be considered to be the same. Figure 4 ilustrates the results according which the Auto.SARIMA procedure gives the best forecasts in 5% of cases; the Auto.AIC in 36.7% cases. There are similar forecasts by both procedures in 58.3% cases. The Auto.AIC is better mainly for the non-seasonally non-stationary processes and the Auto.SARIMA for the near non-seasonally non-stationary processes.

Figure 5 shows the RMSE of the forecasts computed by the Auto.AIC procedure for the individual processes. It can be seen that along with the growing parameter values, the RMSE grows as well. The best results are either for the processes with zero or small values of the parameters. The worst results are for the nonseasonal non-stationary processes. It is interesting that the seasonal nonstationarity does not have such a strong influence on the forecasts RMSE like the nonseasonal nonstationarity. Figure 6 shows the RMSE of the forecasts computed by the Auto.SARIMA. The pattern is similar to that in Figure 5.

3.3 Forecasting of the nearly integrated time series

In this part we will extend the above analysis about the situation of so called near integrated, but still stationary processes. Figure 7 depicts the forecasting success of the nonseasonal integrated model of the SARIMA$(0,1,1)(1,0,1)_{12}$ type for this type of process, irrespective of the forecasting procedure. It can be seen that even for the non-seasonally stationary processes with ϕ_1 from 0.90 to 0.95, the integrated model is more suitable for forecasting than the correct stationary model. This result is consistent with the result for the example of Pincheira and Medel (2016). The possible explanation is that the estimates of the parameters of the correct models for the time series generated by the nearly non-stationary processes have greater variability and are thus less accurate.

Figure 7 The Forecasting Success of SARIMA(0,1,1)(1,0,1)



Source: Authors' calculations

CONCLUSION

The goal of the simulation study was to analyze the relationship of the size of the systematic part of the process (it is given by the magnitude of the SARIMA parameters, bigger values of parameters mean stronger systematic part), which is used for time series generation and the ability to select the proper model for the generated time series. The second goal was to analyze the quality of forecasts for the time series generated by the processes with different systematic parts. In this connection the analysis of the ability to select suitable model and reach the relatively accurate forecasts for the time series generated by the near non-stationary and the non-stationary processes was also important.

As a results of the simulation study, the following facts have been found:

1. The Auto.AIC procedure is better for the selection of models for the time series generated by the stationary and invertible processes. The Auto.SARIMA procedure is better for the modelling the time series from so called marginal processes; i. e. mainly from the non-stationary processes and the processes that do not contain the non-seasonal and the seasonal autoregressive parts.
2. For both procedures it is difficult to find the correct model for the time series generated by processes with low values of the autoregressive parameters, and by the near non-stationary processes.

In the first case, the systematic part of the time series is very weak and the property which we are looking for does not show sufficient transparency, so it is possible to overlook it. In the second case, the two different and incompatible situations have the same, or very similar effects, so it is difficult to distinguish between them.

3. The Auto.AIC procedure leads to the better forecasts, but for near to non-stationary processes the Auto.SARIMA procedure is better. The differences in the accuracy between the Auto.SARIMA and Auto.AIC procedures are relatively small. With the growing magnitude of parameters, the accuracy of forecasts decreases in the case of both procedures.
4. For the forecasting of the time series generated by the non-seasonally nearly integrated processes, the non-seasonally integrated models are more suitable than the correct stationary ones.

ACKNOWLEDGEMENTS

This paper was written with the support of the Czech Science Foundation project No. 402/12/G097 DYME – *Dynamic Models in Economics.*

References

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 1974, 19(6), pp. 716–723.

BOX, G. E. P, JENKINS, G. M., REINSEL, G. C., LJUNG, G. M. *Time Series Analysis: Forecasting and Control.* Wiley, 2015.

BROCKWELL, P. J. AND DAVIS, R. A. *Introduction to Time Series and Forecasting.* Springer, 2010.

CANOVA, F. AND HANSEN, B. E. Are Seasonal Patterns Constant Over Time? A Test for Seasonal Stability. *Journal of Business and Economic Statistics*, 1995, 13, pp. 237–252.

DICKEY, D. A. AND FULLER, W. A. Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Stat. Association*, 1979, 74, pp. 427–431.

ENDERS, W. *Applied Econometric Time Series.* Wiley, 2014.

ENGLE, R. Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. *Econometrica*, 1982, 50, pp. 987–1008.

HAMILTON, J. D. *Time Series Analysis.* Princeton University Press, 1994.

HANNAN, E. J. AND QUINN, B. G. The Determination of the Order of an Autoregression. *Journal of Royal Statistical Society*, 1978, 41, pp. 190–195.

JARQUE, C. AND BERA, A. Efficient Tests for Normality, Heteroscedasticity, and Serial Independence of Regression Residuals. *Economics Letters,* 1980, 6, pp. 255–259.

KWIATOVSKI, D., PHILLIPS, P. C. B., SCHMIDT, P., SHIN, Y. Testing the Null Hypothesis of Stationarity Against the Alternative of a Unit Root. *Journal of Econometrics*, 1992, 54, pp. 159–178.

LJUNG, G. M. AND BOX, G. E. P. On a Measure of Lack of Fit in Time Series Models. *Biometrika*, 1978, 65, pp. 297–303.

PESARAN, M. H. *Time Series and Panel Data Econometrics.* Oxford University Press, 2015.

PHILLIPS, P. C. B. AND PERRON, P. Testing for a Unit Root in Time Series Regression. *Biometrika*, 1988, 75, pp. 335–346.

PINCHEIRA, P. M. AND MEDEL, C. A. Forecasting with a Random Walk. *Finance a úvěr – Czech Journal of Economics and Finance*, 2016, 66(6), pp. 539–564.

R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing* [online]. R Foundation for Statistical Computing, Vienna, Austria, 2008. <*http://www.R-project.org*>.

SCHWARTZ, G. Estimating the Dimension of a Model. *Annals of Statistics*, 1978, 6, pp. 461–464.

TRCKA, P. *Výstavba lineárnych stochastických modelov časových radov triedy SARIMA – automatizovaný postup.* Dipl. thesis, University of Economics, Prague, 2015.

WEI, W. W. S. *Time Series Analysis: Univariate and Multivariate Methods.* 2nd Ed. Pearson Education, 2006.

Use of Discriminant Analysis of Data from the Fluorescence Spectometry Analysis of Archaeological Metal Artefacts

Ján Tirpák[1] | *Constantine the Philosopher University in Nitra, Slovakia*
Anna Tirpáková[2] | *Constantine the Philosopher University in Nitra, Slovakia*
Jozef Zábojník[3] | *Archaeological institute SAS in Nitra, Slovakia*

Abstract

This paper aims to present application of methods of mathematical statistics of performed on archaeological metal artefacts, in particular bronze ferrules dated to the period of Avar Khaganate (8^{th}–9^{th} century), which were found at burial site in the municipality of Obid, Slovakia. Based on the results of X-ray fluorescence spectrometry, which was applied for determination of the proportional content of chemical elements in archaeological metal findings, three types of bronze alloys were recognized that the ferrules were made of. In order to identify the ability of variables (chemical elements) to discriminate the bronze ferrule types and also in order to categorise the non-classified bronze ferrules in the three groups the method of canonical discriminant analysis was employed.

Keywords

Bronze ferrules, discriminant analysis, X-ray fluorescence spectrometry, Avar bronze ferrules

JEL code

C38, B19, Z13

INTRODUCTION

X-ray fluorescence spectrometry of archaeological metal artefacts has been known for several decades. The objective of these analyses is to determine the composition of alloys and, thus, contribute to, *inter alia*, the understanding of the production technology of the historical artefacts. This issue has been addressed by many researchers, such as J. Condamin and S. Boucher (1973), J. Riederer and E. Briesse (1974), P. T. Craddock (1978), L. Költő (1982), J. Bayley (1985, 1989), F. Beck et al. (1988), J. Frána and A. Maštalka (1992), B. Tobias (2007), E. Horváth et al. (2009), P. Craddock et al, (2010) and N. Profantová (2010). Liritzis, I. and Zacharias, N. (2011) wrote about portable X-ray devices (PXRF) as instruments that

[1] Gemological institute, Constantine the Philosopher University in Nitra, Nábrežie mládeže 91, 949 74 Nitra, Slovakia.
[2] Department of Mathematics, Constantine the Philosopher University in Nitra, Andrej Hlinka 1, 949 74 Nitra, Slovakia. Corresponding author: e-mail: atirpakova@ukf.sk.
[3] Archaeological institute SAS in Nitra, Akademická 2, 949 21 Nitra, Slovakia.

are aligned along this leading research trend. Issues of performance and reliability of portable X-ray fluorescence (pXRF) instrumentation in archaeological investigations have been studied also by Goodale, N. et al. (2012) and Speakman, R. J. and Shackley, M. S. (2013).

Many of these researchers named the historical bronze or brass artefacts by their principal component, such as tin bronze, leaded bronze, tin lead bronze, leaded brass. In order to determine the chemical composition of historical artefacts the methods of non-destructive X-ray fluorescence analysis were applied.

In this paper we focused on the issue of employment of spectral analysis, more specifically the X-ray fluorescence, in investigation of bronze archaeological artefacts found at the burial site in Obid. In the analysis of the objects, classical names of alloys, namely bronze (copper alloy with tin) and brass (zinc copper alloy), were added along with the addition of other elements such as, for example, bronze containing lead, silver and zinc, respectively brass containing lead, tin, gallium. For the statistical evaluation of the bronze artefacts the method of discriminant analysis was applied.

1 METHODOLOGY OF RESEARCH

In 1963 supervised by Z. Liptáková and later on in 1981 till 1984 supervised by J. Zábojník an in-town archaeological investigations were conducted in the municipality of Obid, locality Fenyes árok. Their results are known solely from papers in the yearbook AVANS (Zábojník, 1982; 1983; 1984; 1985). Altogether 195 graves were excavated.

At the burial site dated to the period of Avar Khaganate various 8^{th}–9^{th} century artefacts were found, such as women's jewellery, parts of belts and decorative parts of horse harnesses. The composition of alloys that the artefacts had been made of was studied by means of fluorescence method. Since most of the findings were parts of bronze ferrules of belts (54 pieces), our analyses were focused on belt bronze ferrules. Analyses of the artefacts were performed in 2016 with the use of manual X-ray fluorescence spectrometer DELTA CLASSIC+ produced by Olympus, USA. DELTA CLASSIC + is an energy-dispersive X-ray fluorescence spectrometer used for the analysis of small objects or heterogeneous materials (detailed technical information: 4 watt RTG lamp with current up to 200 uA; detector: Si-PIN; integrated full VGA camera; the possibility of narrowing the X-ray beam from 9 to 3 mm).

One of the drawbacks is that the spectrometer DELTA CLASSIC + measures only the surface of the examined material and therefore the choice of the location of the measurements on the studied subject is very important. In the actual measurements it should be noted that if the material is gold-plated or otherwise surface-treated, the chemical composition may not correspond to the weight percentages of the whole artefact volume but only to the weight percentages of the measured surface layer at the measuring site. In addition, as far as the location of a measurement is concerned, the reliability and the calibration of the measuring instrument is also essential. This issue was studied in detail by Hunt, A. M. W. and Speakman, R. J. (2015). In our case, for measuring the content of chemical elements in the bronze ferrules, we used a hand-held X-ray fluorescence spectrometer DELTA CLASSIC +, which is calibrated annually by certified company Olympus Industrial Systems Europe, the Czech Republic.

The spectrometer was employed for determination of the proportional content of seven chemical elements (Cu, Sn, Pb, Zn, Ag, As and Ga) in each of the bronze ferrules. Based on the determined content of the elements the bronze ferrules were divided into three groups, i.e. we have created the following three types of bronze alloys which were recognized in the ferrules – tin bronze group (bronze alloy containing less than 5% of Pb), tin lead bronze ferrules (bronze alloy containing from 5 to 10% of Pb) and leaded bronze group (bronze alloy containing more than 10% of Pb). The first group of bronze ferrules is a bronze alloy containing lead and other elements that were probably a natural part of the copper ore or tin ore used for the alloy production. The other two groups consist of a bronze alloy with an intentional addition of lead in the manufacture of the casts of bronze Avar artefacts for their better formability.

For each group, the arithmetic mean of the proportional contents of each of the chemical elements was calculated (Table 1).

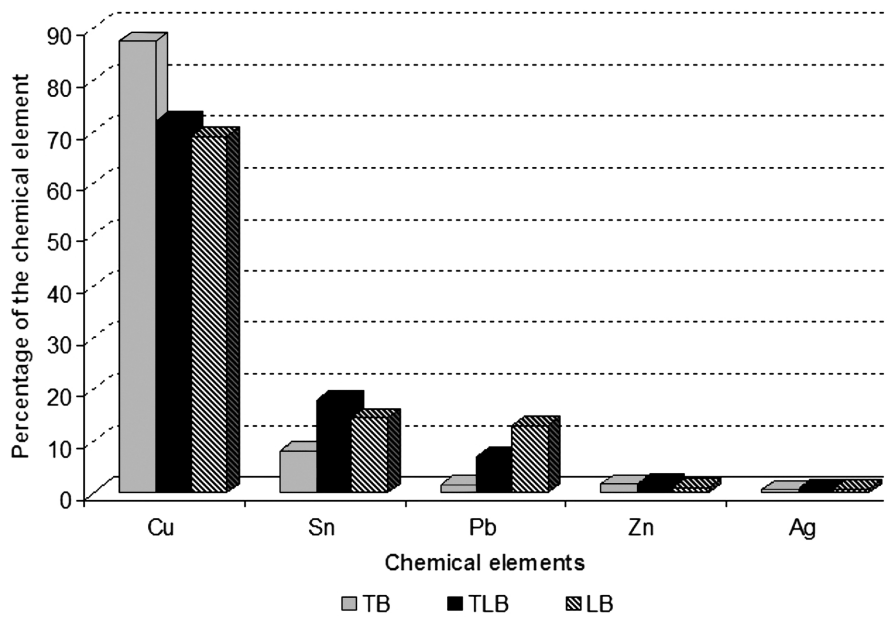
Table 1 Mean proportional content of chemical elements (%)

	Cu	Sn	Pb	Zn	Ag	As	Ga
Tin bronze	87.55	8.28	1.55	1.71	0.64	0.11	0.00
Tin lead bronze	72.06	18.16	7.07	1.90	0.73	0.00	0.00
Leaded bronze	69.17	14.73	13.26	1.09	0.73	0.11	0.69

Source: Own construction

The mean proportional content of chemical elements in the three bronze alloy groups are also displayed in Figure 1.

Figure 1 Mean proportional content of chemical elements in the three types of bronze alloys



Source: Own construction

As shown in Table 1 and Figure 1, each type of bronze alloy (tin bronze, tin lead bronze, and leaded bronze) contains different mean values of the chemical elements, which indicates that the proportional content of the elements varies in the three groups.

Our aim was to find a classification criterion for sorting bronze ferrules into the proposed three groups according to the chemical composition of these ferrules.

In order to verify the rightness of the categorization of the bronze ferrules in the three groups created by us the statistical method of discriminant analysis (DA) was used (Hebák, P., Hustopecký, J. et al., 2007). By means of DA the discriminative ability of the observed variables can be identified, and thus, the existing groups and the groups of statistical units of the population known in advance can

be discriminated. In addition, based on these variables it can be predicted which group a previously not categorized unit belongs to.

In DA, the process of classification follows various rules depending on the observed variables. An often used method is the canonical DA. By means of canonical DA it can be traced which variables (Cu, Sn, Pb, Zn or Ag) in the best way predict/determine the categorization of the bronze ferrules in the groups.

The main idea and the procedure of the canonical DA application are described below.

Suppose that there are n_i statistical units (bronze ferrules) which are divided into K groups ($K > 1$), and for every unit the values of p quantitative variables X_j, $j = 1, 2, ..., p$ were detected (in this case $p = 5$, the percentage of the chemical element (Cu, Sn, Pb, Zn and Ag) in the bronze artefact. Then, the i^{th} unit of the k^{th} group is characterized by the vector of values:

$$\mathbf{x}_{\boldsymbol{ki}}^{T} = (x_{ki1}, x_{ki2}, ..., x_{kip}), \quad k = 1, ..., K. \tag{1}$$

In canonical DA in order to discriminate the groups we look for so-called canonical discriminant functions, which present linear combinations of the studied variables X_j, $j = 1, 2, ..., p$. First, we calculate the matrix expressing the within-group variability:

$$\mathbf{E} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)(\mathbf{x}_{ki} - \overline{\mathbf{x}}_k)^T \tag{2}$$

and the matrix expressing the between-group variability:

$$\mathbf{B} = \sum_{k=1}^{K} \sum_{i=1}^{n_k} (\overline{\mathbf{x}}_k - \overline{\mathbf{x}})(\overline{\mathbf{x}}_k - \overline{\mathbf{x}})^T, \tag{3}$$

where $\overline{\mathbf{x}}_k$ is the vector of means in the k^{th} group and $\overline{\mathbf{x}}$ is the vector of means of variables in the whole sample.

The discriminant functions can be expressed as follows:

$$Y = v_1 x_1 + v_2 x_2 + ... + v_p x_p . \tag{4}$$

The objective of the analysis is to find such a vector $\mathbf{v}^T = (v_1, v_2, ..., v_p)$ that the ratio of the between-group and the within-group variability of the variable Y were the greatest possible, in other words, the discriminant function would discriminate the groups of statistical units in the best possible way. This ratio, denoted by F, is referred to as Fisher discriminant criterion and is expressed as follows:

$$F = \frac{\mathbf{v}^T \mathbf{B} \mathbf{v}}{\mathbf{v}^T \mathbf{E} \mathbf{v}}. \tag{5}$$

The solution (5) that maximizes values F are the eigenvalues of the matrix $\mathbf{E}^{-1}\mathbf{B}$ and their corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_r$. The matrix $\mathbf{E}^{-1}\mathbf{B}$ has r eigenvalues (nonzero), where $r = \min(p, K-1)$. Suppose that eigenvalues arranged in descending order are $\lambda_1, \lambda_2, ..., \lambda_r$. Their corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_R$ are not unique, since e.g. if $\mathbf{v}_1$ is an eigenvector of the matrix $\mathbf{E}^{-1}\mathbf{B}$, then also $a\mathbf{v}_1$ is an eigenvector of the matrix $\mathbf{E}^{-1}\mathbf{B}$ for any real number a. If the eigenvector $\mathbf{v}_1 = (\mathbf{v}_{11}, \mathbf{v}_{21}, ..., \mathbf{v}_{p1})$ is such that $\mathbf{v}^T\mathbf{E}\mathbf{v} = 1$, i.e. the within-group variability of the variable $Y_1 = v_{11}x_1 + v_{21}x_2 + ... + v_{pr}x_p$ equals 1, then it holds that $\lambda_1 = \mathbf{v}^T\mathbf{B}\mathbf{v}$, i.e. the between-group variability of the variable Y_1 equals the first eigenvalue. Consequently, if $\mathbf{v}_1$ is such that the within-group variability of the variable Y_1 equals 1, i.e.

$$\frac{1}{n-K}\mathbf{v}^T\mathbf{B}\mathbf{v}=1, \tag{6}$$

then it holds that:

$$\lambda_1=\frac{1}{n-K}\mathbf{v}^T\mathbf{B}\mathbf{v}, \tag{7}$$

i.e. the within-group variability of the variable Y_1 represented by the variance equals the first eigenvalue. Then, the total variance of the variable Y_1 equals $1 + \lambda_1$. The function Y_1 is referred to as the first discriminant or the first canonical variable. Analogously we calculate all the discriminant functions Y_1, Y_2,..., Y_R.

For every statistical unit (bronze ferrule) a so-called discriminant score can be computed when the values of variables found for this unit are taken for corresponding variables in the discriminant function which is further modified by such a constant that the mean discriminant score in the set of all units equals zero. Thus, the score of R^{th} discriminant variable for the i^{th} unit of the k^{th} group is computed by the following formula:

$$y_{kir}=-(v_{1r}\overline{x}_1+v_{2r}\overline{x}_2+\ldots+v_{pr}\overline{x}_p.)+v_{1r}x_{ki1}+v_2x_{ki2}+\ldots+v_px_{kip}, \tag{8}$$

where $\overline{x}_j$ is the arithmetic mean of values of the j^{th} variable ($j=1,2,\ldots,p$) detected in all statistical units. For every group k ($k=1,2,\ldots,K$) the mean values $\overline{x}_{k1}, \overline{x}_{k2},\ldots, \overline{x}_{kp}$ for all p variables are calculated first, and these are then put into the canonical variables. This way we obtain the vector of mean values of discriminants for the group k, i.e. the vector of group centroids $\overline{\mathbf{y}}_k^T=(\overline{y}_{k1}, \overline{y}_{k2},\ldots, \overline{y}_{kR})$. By comparison of the group centroids it can be found which groups are separated by the first discriminant function, which groups are discriminated by the second discriminant etc.

The coefficient v_{jr} describes the individual effect of the j^{th} variable X_j on the r^{th} canonical variable Y_r (in case the rest of them are constant). It is preferable to have the variables $X_1, X_2, \ldots, X_n$ standardized. In case the variables have not been previously standardized, the discriminants are standardized by the following formula:

$$\mathbf{v}_r^*=\frac{1}{\sqrt{n-K}}\mathbf{F}\mathbf{v}_r, \tag{9}$$

where **F** is a diagonal matrix, the non-zero elements in the diagonal being the square roots of the entries of the matrix **E**. In order to see which variable is characteristic for the r^{th} discriminant, the canonical correlation coefficient can be computed. The vector of the correlation coefficients of the canonical variable and variables $X_1, X_2, \ldots, X_n$ is obtained by the formula:

$$\mathbf{r}_r=\frac{1}{\sqrt{n-K}}\mathbf{F}^{-1}\mathbf{E}\mathbf{v}_r. \tag{10}$$

Up to this point the DA was applied just to describe the between-group differences. Next, we focus on the question to what extent the particular variables affect the categorization, and whether it is necessary to use all of the variables for the discrimination purposes. Whether the chosen statistical method is suitable and which variables (Cu or Sn, Pb, Zn, Ag) are useful for discrimination of the groups, it is necessary to test the null hypothesis H_0: *All eigenvalues equal zero*, i.e. *none of the discriminants is useful for discrimination of belt bronze ferrules groups.*

The above mentioned hypothesis is equivalent to the hypothesis that the vectors of mean values corresponding to the discriminant functions are mutually equal in all K groups. To test the null hypothesis we can use Bartlett's statistic V which has χ^2 ($p(K-1)$) distribution and is obtained by the formula:

$$V = (\sum_{i=1}^{K} n_i - 1 - \frac{p+K}{2})(-\ln \Lambda), \tag{11}$$

where Λ is Wilks statistic expressed as:

$$\Lambda = \Pi_{i=1}^{r} \frac{1}{1+\lambda_i}. \tag{12}$$

Hypothesis is rejected at significance level α, if V exceeds the critical value $\chi^2_{1-\alpha}$. If the hypothesis $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_r = 0$, is rejected, it means that there is at least one non-zero eigenvalue. In fact, it is the first to eigenvalue λ_1, since the eigenvalues are arranged in descending order, so the first of them is the greatest. Next, we proceed to test the other eigenvalues, testing, actually, the hypothesis $H_0 : \lambda_2 = \lambda_3 = \ldots = \lambda_r = 0$, applying the testing statistics:

$$V = \left(n_i - 1 - \frac{p+K}{2} \right) \sum_{i=2}^{r} \ln(1+\lambda_i), \tag{13}$$

which has $\chi^2((p-1)(K-2))$ distribution. The test procedure runs until the hypothesis is rejected, i.e. until no non-zero eigenvalue remains. The hypothesis $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_r = 0$ can be also tested by Rao's statistic *F* whose formula we do not present here. The test is used to determine the number of discriminatory functions that significantly separate the groups.

For a statistical unit the vector of discriminant scores is computed as well as Mahalanobis distance (Euclidean distance, as the canonical functions are not correlated) of this vector from the group centroid of each group by the formula:

$$d_{ik}^2 = \sum_{r=1}^{P} (y_{ir} - \overline{y}_{kr})^2, \;\; k = 1, 2, ..., K, \tag{14}$$

where y_{ir} is the r^{th} discriminant score of the i^{th} observation and $\overline{y}_{kr}$ is the vector of the mean scores for the centroid. The statistical unit is then categorized in the group for which this distance is the smallest.

2 RESULTS AND DISCUSSION

Before using the discriminant analysis method, it is useful to verify the assumptions that the analyzed data must meet so that this method can be used. The assumptions of using discriminant analysis are:

- multivariate normality (it is assumed that data represent a sample from multivariate normal distribution),
- homogeneity of variances/covariances. (it is assumed that the variance/covariance matrices of variables are homogeneous across groups),
- multicollinearity (there must be no correlation between independent variables).

The assumptions of the multivariate normality were verified by the test of Henze-Zirklerovym IRR (Henze, B. and Zirkler, N., 1990) and the test of Royston (Royston, J. P., 1983) in the program R through the package IRR (Korkmaz, et al., 2014; R Core Team, 2017).

Based on the test of Henze-Zirkler's IRR the assumption of multivariate normality only in the group of the tin bronze was rejected and for the other two groups it was not rejected (tin bronze: $HZ = 0.450$ $p = 0.420$; leaded bronze: $HZ = 0.680$, $p = 0.110$; tin bronze: $HZ = 1.730$, $p < 0.001$).

Based on Royston's test the assumption of multidimensional normality in the tin bronze group was also rejected but in the other two groups, it was not be rejected (tin lead bronze: $H = 1.570$, $p = 0.181$; leaded bronze: $H = 4.640$, $p = 0.062$; tin bronze: $H = 9.040$, $p = 0.005$).

When checking one-dimensional normality, only one variable was problematic (Pb) and only in one group. Based on this fact, we consider the assumption of the multivariate normality as fulfilled.

Based on the result of Box's M test ($M = 39.080$, $F = 5.770$, $p < 0.001$), we reject the hypothesis about the equality of the variance-covariance matrix among the groups. The test is very sensitive for multivariate normal distribution. However, the logarithms of the determinants of covariance matrices for each group do not show a significant difference (tin lead bronze: log |D| = 4.670; tin bronze: log |D| = 3.790; leaded bronze: log |D| = 6.490), which indicates the non-breach of the variance-covariance matrices equality assumption.

Based on the values in the pooled within-groups matrix, we noted that in the data (if we consider all 5 original variables), there is multicollinearity caused by a pair of Cu and Sn variables ($r = –0.910$). The remaining correlations of the pairs of the variables are in the absolute value less than 0.400. However, if we consider that the stepwise procedure was the Cu variable eliminated in the DA and the DA was already realized with only a pair of variables Pb and Sn, which are uncorrelated ($r = 0.030$), we can conclude that the multicollinearity in the data is not present.

In our case we applied canonical discriminant analysis to analyze 54 pieces of bronze ferrules, observing 5 variables in each ferrule, i.e. the detected content of the percentage part of five chemical elements – copper (Cu), tin (Sn), lead (Pb), zinc (Zn), and silver (Ag), contained in the ferrules. Canonical discriminant analysis was carried out in the software STATISTICA. A stepwise method was used. Having set the input data, we received the following results (Table 2).

Table 2 Results of stepwise MANOVA

Element	Wilks' Lambda	Partial Lambda	*F* test	*p*-value
Cu	0.149	0.999	0.002	0.998
Zn	0.147	0.982	0.457	0.636
Ag	0.146	0.973	0.681	0.511
Pb	0.712	0.210	94.070	0.000
Sn	0.175	0.857	4.204	0.020

Source: Own construction

In Table 2, the results are of the stepwise MANOVA in Table 2. The aim MANOVA was to find out which variables are unnecessary in the presence of other variables when separating groups. Based on the results shown in Table 2, we can see that the Cu, Zn and Ag variables do not contribute to the separation of the groups and therefore need to be excluded in the next analysis.

In further analysis of the bronze ferrules only those variables (Sn and Pb) were calculated by the discriminant analysis whose eigenvalues most contribute to the minimal value of the discriminative criterion, i.e. those variables which according to the results of the previous analysis discriminate the alloys in a significant way.

Table 3 Chi-square test of gradual roots

	Eigenvalue	Canonical correlation *R*	Wilk's Lambda	Chi-square	df	*p*-value
0	4.928	0.912	0.150	95.948	4	0.000
1	0.128	0.337	0.887	6.074	1	0.014

Source: Own construction

The eigenvalues (Table 3) are computed as the ratio of between-group and within-group sums of squares. A high eigenvalue (4.928) corresponds to a strong discriminant function. Since the value of the canonical correlation coefficient is high (r = 0.912), the first discriminant function discriminates the groups well. Wilks lambda is the ratio of the within-group squares and the total sum of squares. Wilks lambda equals 1 if the group means for the first canonical variable whose the equivalence verify by this test. These group means are mutually equal; lambda is small if the group means are different.

The significance of the difference is expressed by the *p*-value. Since in this case p = 0.000, the group means are significantly different.

Next, the standardized coefficients of canonical discriminant functions for variables Pb and Sn were computed (Table 4).

Table 4 Coefficient of canonical discriminant function

Variable	Root 1	Root 2
Sn	–0.207	–0.979
Pb	–0.972	0.240
Eigenvalue	4.928	0.128
Constant	2.346	1.271
Sn	0.175	0.020

Source: Own construction

Therefore, the obtained canonical discriminant functions are as follows:

for $\lambda = 4.928$

$Y_1 = -0.207\ \text{Sn} - 0.971\ \text{Pb} + 2.346,$

for $\lambda = 0.127$

$Y_2 = -0.979\ \text{Sn} - 0.240\ \text{Pb} + 1.271.$

On the bases canonical discriminant functions above we can see, that the Pb variable most contributes to the separation of groups in the direction of the first canonical discriminant function, whereas in the case of the second canonical discriminant function, it is the variable Sn. By these functions we can compute scores for every statistical unit, i.e. for every bronze ferrule.

Finally, applying the model of canonical DA we obtained the means of the discriminant scores of the objects in groups for the first and second canonical discriminant functions (Table 5). These numbers express the coordinates of the centroids in two-dimensional space of the canonical discriminant functions. Canonical functions represent the transformation of two-dimensional vectors determining particular variables into plane.

Table 5 Functions at Group Centroids

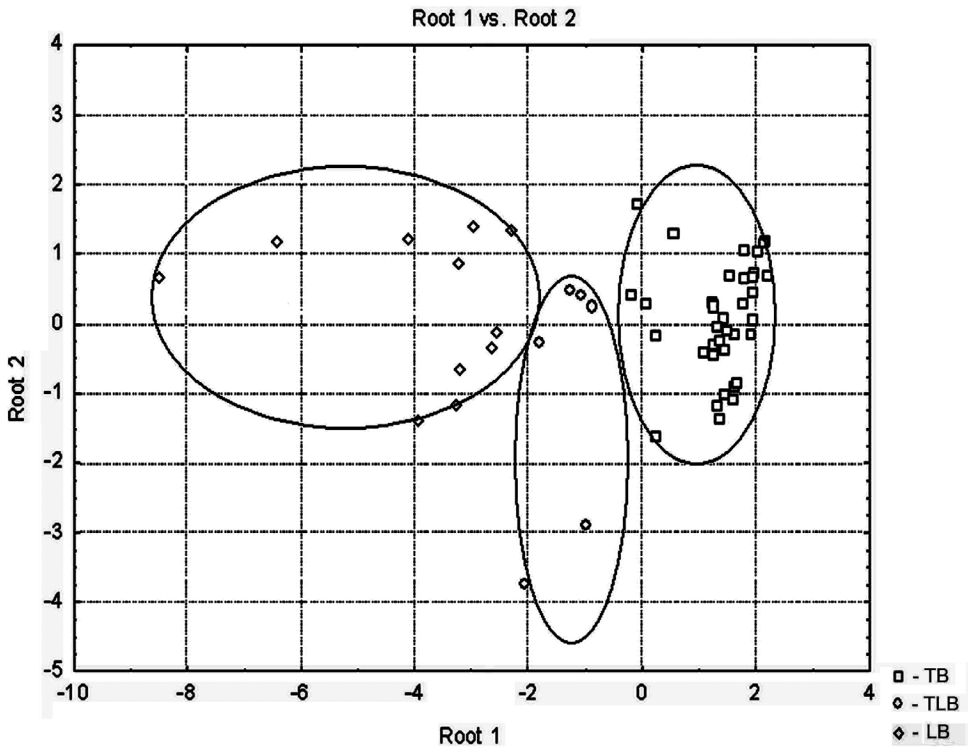
Groups	Root 1	Root 2
Tin bronze	1.386	0.075
Tin lead bronze	–1.354	–0.958
Leaded bronze	–3.923	0.270

Source: Own construction

The situation is displayed in the following scatterplot (Figure 2). Based on the scatterplot it can be decided if both discriminant functions contribute to the group discrimination, or whether one

of the functions is sufficient. As shown in the scatterplot, it suffices to apply one of the discriminant functions to discriminate the groups, and it discriminates in a very good way the groups, i.e. the alloys which the bronze ferrules were made of.

Figure 2 Scatter plot of the canonical scores (TB = tin bronze; TLB = tin lead bronze; LB = leaded bronze)



Source: Own construction

One of the well-known Czech archaeologists who were concerned with the statistical methods in archaeological research which are used to acquire, discover and explore various archeological structures, was professor Neustupný, E. (1993). Professor Neustupný in his study (Neustupný, E., 2005) describing the properties of archaeological artifacts, stated as one of the properties the localization of these artifacts in the space, and then he compared the results of the analysis with the reality. Our main goal was to find a classification criterion for classifying bronze ferrules into three groups only based on their chemical composition, not the spatial localization of the bronze ferrule findings. For illustration, we have plotted the occurrence of all the analyzed bronze ferrules in the map of the burial site Obid, Slovakia (Figure 3).

If we simultaneously consider the real occurrence of bronze ferrules (Figure 3) and the results obtained by the discriminant analysis (Figure 2), it can be observed that the spatial localization of the bronze ferrules is not closely related to the chemical composition of the ferrules. It can be explained by the fact that the bronze ferrules were found only in 11 graves out of the total 199 graves uncovered by the archaeological excavation. Furthermore, the bronze ferrules were placed unevenly in the site, in other words in some of the graves many more bronze ferrules were found than in other graves. The occurrence of a larger number of bronze ferrules

Figure 3 The occurrence of bronze ferrules in the map of the burial site Štúrovo-Obid (TB = tin bronze; TLB = tin lead bronze; LB = leaded bronze)

Source: Own construction

in the graves was usually related to the level of social layer the buried individual had belonged to. Nevertheless, certain group localization of the occurrence of individual types of bronze ferrules can be recognized.

If we simultaneously consider the real occurrence of bronze ferrules (Figure 3) and the results obtained by the discriminant analysis (Figure 2), it can be observed that the spatial localization of the bronze ferrules is not closely related to the chemical composition of the ferrules. It can be explained by the fact that the bronze ferrules were found only in 11 graves out of the total 199 graves uncovered by the archaeological excavation. Furthermore, the bronze ferrules were placed unevenly in the site, in other words in some of the graves many more bronze ferrules were found than in other graves. The occurrence of a larger number of bronze ferrules in the graves was usually related to the level of social layer the buried individual had belonged to. Nevertheless, certain group localization of the occurrence of individual types of bronze ferrules can be recognized.

CONCLUSION

Having employed the canonical discriminant analysis in processing the results of X-ray fluorescence spectrometry performed on bronze archaeological artefacts found at the burial site in Obid, Slovakia out

of seven observed chemical elements two elements, namely tin (Sn) and lead (Pb), with high discriminative ability were identified, i.e. these two alloy components in a significant way discriminate three types of bronze ferrules. The analysis confirmed that the proportional content of the two elements in bronze ferrules was characteristic for a particular type of bronze alloy. The discriminant functions obtained by means of the discriminant analysis enable us to calculate the score for every bronze ferrule, including ferrules which have not been previously categorized, and, thus, also to categorize them into groups according to the bronze alloy types. Based on the presented results we argue that in order to categorize bronze archaeological artefacts into one of the three groups according to the bronze alloy type it is sufficient to determine the proportional content of as few as two chemical elements, tin and lead.

The results of the presented archaeometric study confirmed, inter alia, the important role of the non-destructive analytical methods in archaeological interpretation of artefacts, their composition, origin of the raw material, and the production technology. In addition, the findings support the fact that the excavated burial site shows signs of social stratification via the contents of the graves. The objects found in the graves had been made of precious as well as common alloys, which indicate the economic level as well as the sophistication of the production technologies in the period society.

References

BAYLEY, J. *Analysis and examination of roman brooches from Tiddington, Warwks.* Ancient Monuments Laboratory Report 85/89, London: Historic Buildings and Monuments Commision for England, 1989, 11 p. 01-973-3320.

BECK, F., MENU, M., BERTHOUD, T., HURTEL, L-P. Me´tallurgiedes bronzes. In: HOIRS, J. eds. *Recherches Gallo-Romaines I.* Paris: Laboratoire de Recherches de Muse´es de France, 1985, pp. 70–139.

CONDAMIN, J. AND BOUCHER, S. Recherches techniques sur des bronzes de Gaule Romaine. *Gallia*, 1973, 31, pp. 157–183.

CRADDOCK, P. T. The composition of the copper alloys used by the Greek, Etruscan and Roman Civilizations 3. The origins and early use of brass. *Journal of Archaeological Science*, 1978, 5, pp. 1–16.

CRADDOCK, P., COWELL, M., HOOK,D., HUGHES, M., LA NIECE,S., MEEKS, N. Change and stasis: the technology of Dark Age metalwork from the Carpathian Basin. *British Museum Technical research bulletin*, London, 2010, Vol. 4, pp. 55–65.

FRÁNA, J. AND MAŠTALKA, A. Röntgenfluoreszenzanalyse von frühmittelalterlichen Bronzen aus Böhmen und Mähren. In: DAIM, F. *Awarenforschungen II*, Wien, 1992, pp. 779–801.

GOODALE, N., BAILEY, D. G., JONES, G. T., PRESCOTT, C., SCHOLZ, E., STAGLIANO, N., LEWI, CH. pXRF: a study of inter-instrument performance. *Journal of Archaeological Science*, 2012, 39, pp. 875–883.

HEBÁK, P., HUSTOPECKÝ, J., et al. *Vícerozměrné statistické metody* (Multivariate Statistical Methods). 2nd Ed. Prague: INFORMATORIUM, 2007, 236 p. ISBN 978-80-7333-056-9.

HENZE, N. AND ZIRKLER, B. A class of invariant consistent tests for multivariate normality. Communications in Statistics. *Theory and Methods*, 1990, 19(10), pp. 3595–3617.

HUNT, A. M. W. AND SPEAKMAN, R. J. Portable XRF analysis of archaeological sediments and ceramics. *Journal of Archaeological Science*, 2015, 53, pp. 626–638.

KORKMAZ, S., GOKSULUK, D., ZARARSIZ, G. MVN: An R Package for Assessing Multivariate Normality. *The R Journal*, 2014, 6(2), pp. 151–162.

KÖLTŐ, L. Avar kori bronztárgyak röntgenemissziós analízise (X-ray emission analysis of bronze objects from the Avar Age). *SMK*, 5/1, 1982, pp. 5–68.

LIRITZIS, I. AND ZACHARIAS. N. Portable XRF of archaeological artifacts: current research, potentials and limitations. *X-ray fluorescence spectrometry (XRF) in geoarchaeology.* New York: Springer, 2011, pp. 109–142.

NEUSTUPNY, E. *Archaeological method.* Cambridge University Press, 1993.

NEUSTUPNÝ, E. Syntéza struktur formalizovanými metodami–vektorová syntéza. *Příspěvky k archeologii*, 2005, 2, pp. 127–152.

PROFANTOVÁ, N. Awarische Funde in der Tschechischen Republik. Forschungsstand und neue Erkenntnisse, *Acta archaeologica Carpathica*, 2010, 45, pp. 203–270.

R CORE TEAM. *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Vienna, Austria, 2017. <https://www.R-project.org>.

ROYSTON, J. P. Some techniques for assessing multivariate normality based on the Shapiro-Wilk W. *Applied Statistics*, 1983, 32(2), pp. 121–133.

RIEDERER, J. AND BRIESE, E. Metallanalysen römischer Gebrauchsgegenstände. *Jahrbuch des Römisch-Germanischen Zentral Museums Mainz*, 1974, 19, pp. 83–88.

SPEAKMAN, R. J. AND STEVEN, M. S. Silo science and portable XRF in archaeology: a response to Frahm. *Journal of Archaeological Science*, 2013, 40, pp. 1435–1443.

TOBIAS, B. Néhány érdekes tárgy a Zillingtal-Unterer Kapellenberg D 41. Sírból. Csatfibulák és ecsetek. *Arch. Ért.*, Budapest, 2008, 132, pp. 325–341.

ZÁBOJNÍK, J. *Výskum slovansko-avarského pohrebiska v Štúrove-Obide* (Research of the Slavonic-Avars burial ground in Štúrovo-Obid). AVANS 1981, publ. 1982, pp. 305–308.

ZÁBOJNÍK, J. *Výskum pohrebiska a sídliska z doby avarskej ríše v Štúrove-Obide* (The research of the burial ground and settlement from the period of the Avarian Empire in Štúrovo-Obid). AVANS 1982, publ. 1983, pp. 261–264.

ZÁBOJNÍK, J. *Tretia sezóna výskumu pohrebiska a sídliska z obdobia avarskej ríše v Štúrove-Obide* (The third season of burial research and settlement from the period of the Avarian Empire in Štúrovo-Obid). AVANS 1983, publ. 1984, pp. 225–227.

ZÁBOJNÍK, J. *Výskum pohrebiska a sídliska z obdobia avarského kaganátu v Štúrove-Obide* (The research of the burial ground and settlement from the period of the Avarian Caganate Štúrovo-Obid). AVANS 1984, publ. 1985, pp. 257–259.

Correlates of Multidimensional Indicator of Quality of Life – Fractional Outcome Model Approach

Hanna Dudek[1] | *Warsaw University of Life Sciences – SGGW, Warsaw, Poland*
Wiesław Szczesny[2] | *Warsaw University of Life Sciences – SGGW, Warsaw, Poland*

Abstract

Quality of life indicators need to be measured through a multidimensional framework. In this study, the data from the survey 'Social Diagnosis' is used. The survey encompasses a set of 16 items relating to the evaluation of satisfaction with particular aspects of life. The item's categories are converted into a [0, 1] interval by using a membership function and then they are aggregated into a composite indicator. Fractional output models are applied to assess the impact of various socio-economic and demographic factors on values of this indicator. Such models are useful tools in cases when the response variable ranges between 0 and 1. It is found that satisfaction with life is U-shaped in age. Furthermore, it increases with education and association membership and decreases with disability, urbanisation, and being widowed or divorced. The results of the estimation indicate that the demographic composition of the household, region of residence and source of income all have a statistically significant impact on the quality of life in Poland.

Keywords

Social diagnosis, quality of life, membership function, fractional outcome models

JEL code

I31, C25

INTRODUCTION

Quality of life is a phrase encountered ever more frequently. It is used in so many contexts and for most different purposes that it is difficult to pin down a universally agreed meaning (Phillips, 2006). In the full sense of the term, Quality of Life (QoL) can be approached from an interdisciplinary perspective – the manner of its use depends on the discipline, and many are involved: sociology, economics, political science, social psychology, medicine, philosophy, marketing, environmental sciences and others (Glatzer, 2004). It has even been claimed that there may be as many definitions of QoL as there are people (Hoe et al., 2011). The recent trend has been to address methodologies that take into account individuals' opinions

[1] Department of Econometrics and Statistics, Warsaw University of Life Sciences – SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland. Corresponding author: e-mail: hanna_dudek@sggw.pl, phone: (+48)225937220.
[2] Department of Informatics, Warsaw University of Life Sciences – SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland. E-mail: wieslaw_szczesny@sggw.pl.

using broadly-designed tools based on questions about the subjective quality of life. Such an approach has an advantage – it prevents the risk of a person's QoL being judged by others, hence avoiding 'diminishing empowering people' in evaluating their own well-being (Rojo-Perez et al., 2015). While in the literature there is a lack of consensus on the meaning of 'quality of life', its multidimensional nature is universally accepted (Betti, 2017; Stiglitz et al., 2009; Eurostat, 2017). When measuring QoL, various domains should be analyzed, including subjective well-being. Indicators of satisfaction with various aspects of personal life are regarded as an important part of monitoring social situation. They enable the comparison of people's feelings against the objective data on living conditions, and thus are an indispensable and crucial element in the multidimensional measurement and analysis of the quality of life (Dudek and Szczesny, 2016).

This study examines the subjective perception of QoL using data from the 2015 survey 'Social Diagnosis – the objective and subjective quality of life in Poland'. It uses methodology first used in a multidimensional poverty analysis, and originally proposed by Cerioli and Zani (1990) and developed by Cheli and Lemmi (1995) and Betti and Verma (1999). It also employs methods of fuzzy set theory (Zadeh, 1965), according to which data on subjective assessments of QoL are converted by a membership function into a [0, 1] interval. Fuzzy set theory has become of particular interest to poverty researchers, since conventional crisp-set applications separating the poor and non-poor are increasingly believed not to adequately capture complex social phenomena like poverty (Neff, 2013).

In order to obtain a synthetic indicator encompassing many areas and aspects of life, weights reflecting the relative importance of satisfaction items are used. Such a framework was first applied in multidimensional poverty analysis (Betti and Verma, 2008; Panek, 2010), but recently also in various other socio-economic areas including job satisfaction (De Battisti et al., 2015) and quality of life (Betti et al., 2016; Betti, 2017). The interesting results obtained by Betti encouraged us to apply his approach to analyze the subjective QoL in Poland.

The present study often refers to Betti's work, where multidimensional fuzzy indicator methodology was first proposed and used to measure QoL (Betti et al., 2016; Betti, 2017). As in those articles, we calculate average values of the QoL indicators for the entire Polish population. As shown in (Betti et al., 2016), estimates based on sub-samples (population groups, regions and the like) can statistically differ from each other. In order to identify such differences, Betti et al. (2016) calculated standard errors for fuzzy indicators of QoL using Jackknife repeated replication. We propose another approach: applying fractional outcome models to explain the indicators. Thus, the main contribution of our research is the use of fractional response models and beta regression models in the fuzzy multidimensional analysis of QoL.

The paper is structured as follows: Following the present introduction, section 1 focuses on the data and methodology. Sub-section 1.1 briefly describes the 'Social Diagnosis' survey, sub-section 1.2 introduces the concept of the fuzzy set approach in a multidimensional measurement of quality of life and sub-section 1.3 gives insights on fractional outcome models. Section 2 presents and discusses the results of the analysis and section 3 provides our conclusions.

1 DATA AND METHODOLOGY

1.1. Data

The empirical analysis in this study is based on a 'Social Diagnosis – the objective and subjective quality of life in Poland' (SD) survey conducted in 2015. The SD is a cyclic survey that collects microdata on Poles' living conditions and quality of life as they report it themselves. The database is available free of charge at the website: <*www.diagnoza.com*>.

The 'Social Diagnosis' research project is undertaken by the members of the 'Council for Social Monitoring'. SD report authors and experts invited to participate by the 'Council' comprise economists, demographers, psychologists, sociologists, insurance specialists and statisticians. Headed by professor Janusz Czapiński, a social psychologist, and professor Tomasz Panek, a statistician, the project focuses

on uncovering fundamental facts, behaviours, attitudes and experiences; not just an ordinary descriptive opinion poll, it is a scientific project.

The research was conducted in March and April 2015 by professional interviewers from the Central Statistical Office. The organisation of the questionnaire survey is supervised by the Polish Statistical Association's Office for Statistical Analyses and Research. Two separate questionnaires were used in the SD research.[3] The first provides the information about the household composition and living conditions completed by the interviewer during a meeting with the best-informed household representative. The second questionnaire was completed by all available household members aged 16 and above and contributes the information about their quality of life (Czapiński and Panek, 2015). The 2015 survey involved 11 740 households and 24 324 household members over 16 years of age. Since our study deals with a subjective assessment of QoL, we used the data derived from the second questionnaire, which was completed by 22 208 persons, which is the study's sample size.

The DS survey uses a two-stage stratified sampling method for selecting households[4]. Census areas were the primary sampling units, sampled with probabilities proportional to the number of dwellings they covered. Urban strata were divided into large towns with more than 100 000 residents, medium-sized towns of 20 000–100 000 and small towns with fewer than 20 000. In the five largest cities, the strata covered individual districts. In the second stage, three dwellings were sampled per census area in large towns, four per area in medium-sized ones, five per area in the smallest towns and six dwellings for rural areas (Czapiński and Panek, 2015). To preserve the representative character on the national study scale and in the identified classification cross-sections, weights for individuals were taken into account in the DS database.

The DS survey questionnaires contain numerous questions about respondent satisfaction with regard to particular areas and aspects of life. The scale of domain satisfaction covers 16 different items exhausting nearly the entire scope of the average person's interests and activities. Czapiński (2015) broke these items down into the following five dimensions:

- social aspects (satisfaction with relationships with closest family members, friends, spouses and children),
- material aspects (satisfaction with the family's financial situation and housing conditions),
- environmental aspects (satisfaction with the situation in the country, place of residence, and level of safety in place of residence),
- health-related aspects (satisfaction with one's health condition, sex-life and way of spending free time),
- self-assessment (satisfaction with one's own achievements, prospects for the future, educational level, work).

Respondents were asked to assess all 16 areas and indicate the extent of their satisfaction with each. There is a range of possible replies: 1) very satisfied 2) satisfied 3) rather satisfied 4) rather not satisfied 5) not satisfied 6) very dissatisfied 7) not applicable.[5] In our study, we assign a value of 3.5 to those individuals who indicated answer '7' and to those who did not give any answers, thus attributing them a neutral position. For 12 items such answers did not exceed 2% of all data (at most 2% of all respondents gave answer 7 or did not give any answers). However, there were individuals who were unmarried, had no children, no sex-life, or who did not work. They had no choice but to answer '7' because they could not assess a spouse, children, sex-life or work. These individuals accounted for 37%, 26%, 27% and 48% of respondents, respectively. So, for 4 of the 16 items, there is a very significant amount of missing information. Thus, we analyse the variant of the data with reduced list of 12 items.

---

[3] Questionnaires and instructions for interviewers can be found at the website: <*www.diagnoza.com*> (Czapiński and Panek, 2015).

[4] Details on sampling design can be found on the website: <*www.diagnoza.com*> (Czapiński and Panek, 2015).

[5] See corresponding questionnaire item in the Appendix.

1.2 The multidimensional indicator of quality of life

In the study we analyze the multidimensional indicator of QoL. Our approach requires the following steps:
1) identify the relevant items and group them into dimensions,
2) convert the items' categories into item scores belonging to a [0, 1] interval,
3) assign weights to the aggregate items' scores in the QoL indicators,
4) calculate the QoL indicators.

As mentioned in the description of the SD research, one of the questionnaires includes 16 questions about satisfaction with particular areas and aspects of life. The answers (replies) to these questions created the items we analyzed in our study. According to SD research head Czapiński (2015), they are grouped into five dimensions: social, material, environmental, health-related and self-assessment. Thus, concerning the first step, omitting 4 items with a significant amount of missing information, we analyze 12 items grouped into 5 dimensions.[6]

In the second step, we construct a membership function for each item. Several methods have been proposed in the literature (Cerioli and Zani, 1990; Cheli and Lemmi, 1995; Betti and Verma, 2008) for how to construct this function. We opt to use the empirical distribution function of each item. Such an approach takes into account a given field's relative position in society. We use the formula fulfilling this requirement (Cheli and Lemmi, 1995):[7]

$$d_{k,j,i} = \frac{1-F(c_{k,j,i})}{1-F(1)}, \tag{1}$$

where: $c_{k,j,i}$ is the category of the j-th item in k-th dimension for the i-th individual, $1 \leq c_{k,j,i} \leq 6$,
F is the corresponding cumulative distribution functions.

The item's categories are ordered from the highest value of QoL to the lowest. Formula (1) converts them into a [0, 1] interval. The item score d can be interpreted as the degree of membership in the fuzzy set of satisfied people. In particular, the value 0 refers to the answer 'very dissatisfied' ($c = 6$) and the value 1 to 'very satisfied' ($c = 1$).

In the third step, weights of items were assigned within each of the five dimensions separately. Weights have to be considered as measures of relative importance of the items in the QoL indicators, relative to the other items in the dimension (Guio, 2009). They are essentially value judgements, and several approaches can be followed for defining them (Desai and Shah, 1988; Cerioli and Zani, 1990; Cheli and Lemmi, 1995; Filippone et al., 2001; Lazim and Osman, 2009).[8] In this study, we use the method proposed by Betti and Verma (1999) for two reasons: it assigns less importance to poorly differentiated items and it takes into account data redundancy. To do both, Betti and Verma (1999) defined weights as the product of two components:

$$W_{k,j} = W^a_{k,j} \cdot W^b_{k,j} \tag{2}$$

with the first factor being the coefficient of variation $V_{k,j}$ for j-th item score d in the k-th dimension, i.e.:

---

6 To identify dimensions, one can use statistical methods, for example factor analysis (Betti et al., 2016; Betti, 2017), but in this study we use a classification applied in the 'Social Diagnosis' Report, according to which there are five dimensions encompassing analyzed items.

7 The analogous formula was used in (Betti, 2017; Betti et al., 2016); the only difference lies in considering opposite ordering of categories c – in Betti's research they are ordered from the lowest value of QoL to the highest.

8 In research (Dudek and Szczesny, 2015) applying SD data methods proposed in papers (Desai and Shah, 1988; Cerioli and Zani, 1990; Betti and Verma, 1999) it was determined that the choice of weights does not significantly affect the distribution of synthetic indicators.

$$W^a_{k,j} = V_{k,j}, \tag{3}$$

and the second factor takes into account correlations among item scores:

$$W^b_{k,j} = \left(\frac{1}{1+\sum_{j=1}^{m_k} r_{k,jj'} | r_{k,jj'} < r^*_k}\right)\left(\frac{1}{\sum_{j=1}^{m_k} r_{k,jj'} | r_{k,jj'} \geq r^*_k}\right), \tag{4}$$

where: $r_{k,jj}$, is the correlation coefficient between the two different scores $d_{k,j}$ and $d_{k,j'}$,
r^*_k is a predetermined cut-off correlation level in the k-th dimension,
m_k is the total number of items in the k-th dimension.

Thresholds r^*_k are determined by the point of the largest gap between the ordered set of correlation values encountered (Betti and Verma, 2008).

Using Formulas (3)–(4) results in weight $W_{k,j}$ being directly proportional to the variability of the $d_{k,j}$ and inversely proportional to its correlation with items in the k-th dimension. The low value of the factor $W^a_{(k,j)}$ means that item score $d_{k,j}$ discriminates individuals poorly, while the low value of the factor $W^b_{(k,j)}$ means that $d_{k,j}$ is highly correlated with other item scores in k-th dimension, thus reducing the effect of redundancy (Betti, 2017). Weights are normalized to unity by setting:

$$w_{k,j} = \frac{w_{k,j}}{\sum_{j=1}^{m_k} w_{k,j}}. \tag{5}$$

In the fourth step we calculate the QoL indicator. First, the sub-indicators in each dimension are calculated. For an i-th individual, aggregation over a set of item scores in a k-th dimension ($k = 1, 2, \ldots, K$) is given by formula (Betti et al., 2016; Betti, 2017):

$$S_{k,i} = \sum_{j=1}^{m_k} w_{k,j}\, d_{k,j,i}, \tag{6}$$

where: $d_{k,j,i}$ – the value of j-th item score in the k-th dimension for the i-th individual,
$w_{k,j}$ – normalised weight for j-th item score in the k-th dimension,
m_k – the total number of items in the k-th dimension.

Next, an overall QoL indicator for the i-th individual is calculated as the mean of sub-indicators $S_{k,i}$:

$$S_i = \frac{1}{K}\sum_{k=1}^{K} S_{k,j}, \tag{7}$$

where K is the number of dimensions.

In the next step of the analysis, to gain a deeper insight into the subject matter, we try to explain the values of the indicator S by socio-economic and demographic factors.

1.3 Fractional outcome models

The aim of our research is to estimate a model with the dependent variable S ranged between 0 to 1, inclusive. To handle these data properly, one should take the bounded nature of the response into account. A comprehensive survey of the models and estimation methods suitable to deal with fractional response variables can be found in (Carrasco et al., 2014; Ramalho et al., 2011). The use of linear regression model can generate predictions outside the unit interval. Moreover, it is conceptually flawed to assume normal

distribution for a response variable in the [0, 1] range. As Papke and Wooldridge (1996) pointed out, the drawbacks of a linear model for fractional data are analogous to the drawbacks to a linear probability model for binary data. One way to handle this for response variables' values belonging to a closed unit interval is to apply a fractional response model (FRM). Papke and Wooldridge introduced such a model in a paper in 1996 (Papke and Wooldridge, 1996).

Fractional regression is a model of the mean of the dependent variable y conditional on covariates $\boldsymbol{x}$, which we denote by $E(y|\boldsymbol{x}) = \mu_x$ Because y is in the [0, 1] interval, to ensure that μ_x also belongs to it [0, 1], in an FRM it is assumed that:

$$\mu(\boldsymbol{x}_i) = G(\boldsymbol{x}_i^{'}\boldsymbol{\beta}), \tag{8}$$

where: $\mu(\boldsymbol{x}_i) = E(y_i|\boldsymbol{x}_i)$
$G(\bullet)$ is a known function with $0 < G(z) < 1$ for $z \in R$,
$\boldsymbol{x}_i$ is a vector of explanatory variables representing the characteristics of individual i,
$\boldsymbol{\beta}$ is a vector of parameters to be estimated.

Typically, non-linear functional forms used for G are chosen to be a cumulative distribution function (cdf). The two most popular examples used in FRM are the logistic function $(z) = \Lambda(z) = \frac{\exp(z)}{1+\exp(z)}$ and $G(z) = \Phi(z)$, where Φ is the standard normal cumulative distribution function. Note that G is the inverse function for the so-called link function that specifies the link between the random and systematic components. It indicates how the expected value of the response variable relates to the linear predictor of explanatory variables. For a discussion on link functions in fractional outcome models, see (Smithson and Verkuilen, 2006; Ramalho et al., 2011).

The nonlinear estimation of an FRM's parameters is performed via maximization of the log-likelihood. The Bernoulli log-likelihood function for the FRM is of the form:

$$lnL = \sum_{i=1}^{n} v_i\, y_i\, lnG(\boldsymbol{x}_i^{'}\boldsymbol{\beta}) + v_i(1 - y_i)\ln\left(1 - G(\boldsymbol{x}_i^{'}\boldsymbol{\beta})\right), \tag{9}$$

where: y_i is the dependent variable for the i-th individual,
x_i are the covariates for individual i, and
v_i denotes sample weight of the i-th individual,
n is the sample size.

To obtain robust estimation of an FRM, the quasi-maximum likelihood (QML) is used. It is important that the QML estimator does not require full distributional assumption of the dependent variable for consistency. The only information that it needs is the conditional mean to be correctly specified for consistent parameter estimates. The QML estimator of $\boldsymbol{\beta}$ is consistent and asymptotically normal, regardless the distribution of the dependent variable, conditional on the predictors (Papke and Wooldridge, 1996). To test the correct link specification of the conditional mean function, Ramsey's RESET test, more common in econometrics literature, can be applied.

The partial effects in an FRM of a given variable, say X_j, are given by:

$$\frac{\partial E(y_i|\boldsymbol{x}_i)}{\partial x_{ji}} = \beta_j\, g(\boldsymbol{x}_i^{'}\boldsymbol{\beta}), \tag{10}$$

where: $g(\boldsymbol{x}_i^{'}\boldsymbol{\beta}) = \dfrac{\partial G(\boldsymbol{x}_i^{'}\boldsymbol{\beta})}{\partial(\boldsymbol{x}_i^{'}\boldsymbol{\beta})}$,

x_{ji} is a value of j-th explanatory variable for i-th individual.

Hence, the significance and the direction of the marginal effects may be analyzed simply by examining the significance and sign of β_j (Ramalho and Vidigal da Silva, 2013).

FRMs have been applied in a variety of disciplines, including the social sciences, health sciences and economics. To see how FRMs have been used, see (Cardoso et al., 2010; Czarnitzki and Kraft, 2004; Flores et al., 2015) to name a few.

A beta regression models (BRMs) may be a valid alternative to FRMs. Though the beta distribution has been known in statistics for about a century, the research that has been done on BRM is relatively recent. BRMs have gained traction thanks to their flexibility for modelling dependent variables ranging to the open unit interval. For papers introducing these models, see (Paolino, 2001; and Ferrari and Cribari-Neto, 2004). BRMs are applied across variety fields, including finance, medicine, psychology and economics, for examples, see (Grzybowska and Karwański, 2015; Karwański et al., 2015; Rogers et al., 2012; Smithson and Verkuilen, 2006; Zanin, 2017).

BRMs are based on the assumption that the dependent variable y is beta-distributed and that its mean is related to a set of explanatory variables through a linear predictor with unknown coefficients and a link function. They also include a precision parameter which may be constant or depend on a set of regressors through a scale-link function as well. The density of a beta-distributed dependent variable y conditional on covariates (explanatory variables) $\mathbf{x}$ can be written as (Ferrari and Cribari-Neto, 2004):

$$f(y, \mu_x, \psi) = \frac{\Gamma(\psi)}{\Gamma(\mu_x\psi)\Gamma((1-\mu_x)\psi)}\, y^{\mu_x\psi-1}\,(1-y)^{(1-\mu_x)\psi-1}, \tag{11}$$

where: $\mu_x = E(y|\boldsymbol{x})$ is the mean of the dependent variable y conditional on covariates $\mathbf{x}$,
ψ scales the conditional variance according to:

$$\mathrm{Var}(y|\boldsymbol{x}) = \frac{\mu_x(1-\mu_x)}{1+\psi}. \tag{12}$$

The parameter ψ is known as the precision parameter[9] since, for fixed μ_x, the larger the ψ, the smaller the conditional variance of y. Note also that conditional variance of y is a function of μ_x which renders the regression model based on this parameterization naturally heteroskedastic (Cribari-Neto and Zeileis, 2010).

A BRM is a model of $\mu_x = \mathrm{E}(y|\boldsymbol{x})$. It is appropriate when y takes values from the (0, 1) interval to ensure that μ_x is also in (0, 1), link function for the conditional mean is used. As for FRMs, it is assumed that the mean μ_x is given by Formula (8), thus the partial effects in the BRM are given by (10).

According to (11), the log-likelihood function is of the form:

$$lnL = \sum_{i=1}^{n} \nu_i\,(ln\Gamma(\psi) - ln\Gamma(\mu_x\psi) - ln\Gamma((1-\mu_x)\psi) + (\mu_x\psi - 1)ln_i + ((1-\mu_x)\psi - 1)ln(1-y_i)), \tag{14}$$

where: y_i is the dependent variable for the i-th individual,
μ_x is given by Formula (8),
ψ is the precision parameter,
n is the sample size,

[9] Precision parameter may be constant or depend on set of regressors through a scale-link function (Smithson and Verkuilen, 2006).

v_i denotes the sample weight of the i-th individual.

Parameter estimation is performed by maximum likelihood (ML), simply replacing μ_x with (8).

In our study we try to explain the values of fractional variable *S*, being the QoL indicator, by explanatory variables using a FRM and a BRM. All computations are performed using STATA 14. In order to ensure a representative character on the national scale and in the identified classification cross-sections, we use a sample weight for each individual.

To compare a goodness of fit of the models to the data, we calculated simple measures by taking the observed (y) value minus its corresponding predicted conditional mean ($\hat{y}$). A lot of measures based on such differences can be obtained. The goodness of fit of models in our research was evaluated using the root mean square error (RMSE) and the mean absolute error (MAE):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n}}, \tag{15}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{16}$$

These are common statistics used to assess models. Large values indicate a poor fit.

2 RESULTS AND DISCUSSION

As described in Section 1, we considered five dimensions encompassing 12 items. All items were converted by membership function (1) into item scores. To calculate weights for them, we applied the procedure *mdepriv*[10] – a Stata command written by Pi Alperin and Van Kerm (2014). These weights were used to calculate the QoL indicators given by Formulas (2)–(5). Table 1 reports descriptive statistics for the overall summary QoL indices *S* and the indices S_1, S_2, S_3, S_4, S_5 corresponding to the five dimensions.

Table 1 Descriptive statistics for overall and dimension-specific QoL indices

Descriptive statistic	Overall	Social	Material	Environmental	Health-related	Self-assessment
	(S)	(S_1)	(S_2)	(S_3)	(S_4)	(S_5)
Mean	0.3917	0.4002	0.3993	0.3732	0.3978	0.3914
Standard deviation	0.1754	0.2575	0.2325	0.1975	0.2290	0.2203
Median	0.3740	0.4139	0.4268	0.3463	0.3852	0.3483
Maximum	1	1	1	1	1	1
Minimum	0	0	0	0	0	0
Skewness	0.6202	0.7283	0.5568	0.5841	0.5280	0.5049
Kurtosis	3.3934	2.8453	2.8686	3.2607	2.7736	2.6840

Source: Authors' calculations

As shown in Table 1, Poles, on average, were best satisfied in relationships with other people (social aspects) and least satisfied with environmental aspects. Mean values of all QoL indices stand at about 0.4 with a standard deviation of about 0.2. A minimum of 0 and a maximum of 1 for the indices *S* and S_1, S_2, S_3, S_4, S_5 means that there existed individuals who were very dissatisfied with each aspect of life and others who were very satisfied. All QoL indices exhibit slightly positive asymmetry, indicating distributions with

[10] We found weak or moderately strong positive correlations among all pairs of item scores in a given dimension.

an asymmetric tail extending toward more positive values. Skewness values close to 0 and kurtosis values close to 3 indicate that distributions of the QoL indices do not differ much from the normal distribution.

The next part of the study explores statistical significance and the impact of various socio-demographic factors on the QoL indicator (*S*). As described in Section 1, we applied the FRMs and BRMs using the logit and the probit link function. Because beta-regression is designed to model values on the interval (0,1) we coded values 0 as 0.0001 and values 1 as 0.9999. There were 3 observations with a value of 0 and 60 observations with a value of 1. We considered a number of socio-economic and demographic variables that can shed light on QoL. Akaike (AIC) and Bayesian (BIC) information criteria were used to compare alternative models with various sets of explanatory variables. See the Appendix for a description of these variables.

Table 2 reports the estimation results for the FRMs and the BRMs with logit and probit variants. We found that for both BRMs, AIC and BIC information criteria clearly indicate the choice of explanatory variables presented in Table 2, while results for the FRMs are not so explicit – the AIC criterion prefers the same set of variables used in the BRMs, but the BIC criterion prefers the set of variables without the variable describing the class of respondents' place residence. To compare the results obtained with the various models, we used the same explanatory variables in each of them.

Table 2 Estimates of fractional regression and beta regression models

Variable	FRM with logit link function		FRM with probit link function		BRM with logit link function		BRM with probit link function	
	b	*S(b)*	*b*	*S(b)*	*b*	*S(b)*	*b*	*S(b)*
Age	–0.0293***	0.0020	–0.0181***	0.0013	–0.0301***	0.0023	–0.0187***	0.0014
Age2	0.0003***	2E-5	0.0002***	1E-05	0.0003***	2E-05	0.0002***	1E-05
Disability	–0.1833***	0.0186	–0.1126***	0.0113	–0.1778***	0.0207	–0.1094***	0.0126
Association membership	0.1069***	0.0190	0.0665***	0.0118	0.1081***	0.0218	0.0673***	0.0136
Civil state								
Married	Ref.	–	Ref.	–	Ref.	–	Ref.	–
Unmarried	–0.0293	0.0211	–0.0180	0.0131	–0.0261	0.0251	–0.0160	0.0156
Widowed	–0.1383***	0.0250	–0.0839***	0.0153	–0.1283***	0.0258	–0.0779***	0.0158
Divorced/separated	–0.1118***	0.0344	–0.0693***	0.0211	–0.1081***	0.0355	–0.0669***	0.0218
Education								
1 (primary)	Ref.	–	Ref.	–	Ref.	–	Ref.	–
2 (basic vocational)	0.1935***	0.0208	0.1184***	0.0127	0.1886***	0.0209	0.1155***	0.0128
3 (secondary)	0.2813***	0.0213	0.1730***	0.0115	0.2929***	0.0224	0.1803***	0.0137
4 (higher)	0.4409***	0.0235	0.2723***	0.0144	0.4499***	0.0251	0.2781***	0.0154
Class of place of residence								
Town bigger then 20 000[12]	Ref.	–	Ref.	–	Ref.	–	Ref.	–
Very small town	0.0918***	0.0204	0.0568***	0.0126	0.0942***	0.0227	0.0583***	0.0135
Village	0.0486***	0.0152	0.0300***	0.0094	0.0632***	0.0181	0.0390***	0.0112
Regions								
Central	Ref.	–	Ref.	–	Ref.	–	Ref.	–
South	0.0776***	0.0196	0.0485***	0.0121	0.0685***	0.0204	0.0429***	0.0126
East	–0.0721***	0.0185	–0.0444***	0.0114	–0.0742***	0.0193	–0.0456***	0.0119

[12] Differences between very big towns, big towns, medium-sized towns and small towns were not statistically significant even at the 0.1 level, therefore we used aggregation to towns bigger than 20 000.

Table 2 Estimates of fractional regression and beta regression models (continuation)

Variable	FRM with logit link function		FRM with probit link function		BRM with logit link function		BRM with probit link function	
	b	*S(b)*	*b*	*S(b)*	*b*	*S(b)*	*b*	*S(b)*
Northwest	0.0775***	0.0217	0.0484***	0.0134	0.0952***	0.0270	0.0593***	0.0167
Southwest	–0.0022	0.0232	–0.0012	0.0144	–0.0062	0.0246	–0.0038	0.0152
North	0.1254***	0.0215	0.0778***	0.0133	0.1635***	0.0257	0.1015***	0.0160
Household type								
MC without children	Ref.	–	Ref.	–	Ref.	–	Ref.	–
MC with 1 child	–0.0394*	0.0214	–0.0246*	0.0132	–0.0307	0.0255	–0.0193	0.0158
MC with 2 children	–0.0091	0.0227	0.0059	0.0141	–0.0165	0.0259	–0.0105	0.0161
MC with 3+ children	–0.0915***	0.0268	–0.0568***	0.0166	–0.1014***	0.0287	–0.0630***	0.0177
Single-parent	–0.1966***	0.0274	–0.1214***	0.0169	–0.2086***	0.0279	–0.1289***	0.0172
Multi-family	0.0059	0.0250	0.0037	0.0155	0.0199	0.0291	0.0123	0.0180
One-person	–0.0583**	0.0267	–0.0365**	0.0165	–0.0802***	0.0278	–0.0500***	0.0172
Non-family	–0.1390*	0.0764	–0.0973*	0.0471	–0.1672**	0.0750	–0.1047**	0.0462
The socio-economic group								
Employees	Ref.	–	Ref.	–	Ref.	–	Ref.	–
Entrepreneurs	–0.0152	0.0307	–0.0090	0.0191	–0.0332	0.0343	–0.0202	0.0213
Farmers	–0.0136	0.0228	–0.0084	0.0141	–0.0337	0.0240	–0.0208	0.0149
Retirees	–0.0272*	0.0166	–0.0167*	0.0103	–0.0315*	0.0179	–0.0194*	0.0110
Pensioners	–0.0742**	0.0301	–0.0458**	0.0184	–0.0684**	0.0352	–0.0424**	0.0216
Living on unearned sources	–0.1647***	0.0430	–0.1027***	0.0263	–0.1663***	0.0510	–0.1038***	0.0311
Constant	0.0343	0.0618	0.0210	0.0383	0.0860	0.0709	0.0538	0.0440
Scale parameter	–	–	–	–	1.8476***	0.0254	1.8476***	0.0254

Note: *b* are estimates, *S*(*b*) – their standard errors. All standard errors are robust (with heteroscedasticity-robust asymptotic variance). * means statistical significance at 0.10, ** – statistical significance at 0.05, *** – statistical significance at 0.01.
Source: Authors' calculations

It is evident that most of the explanatory variables are statistically significant at the 0.01 level. In addition, almost all coefficients in all models have the same sign and statistical significance. This means that the impact of the socio-economic and demographic variables on quality of life can be interpreted in the same way for the FRMs and the BRMS. All of the interpretations presented here were made under the assumption of *ceteris paribus*.

We have determined that age had a negative sign while its squared term had a positive sign, implying a U-shaped effect. In other words, people tend to be more satisfied with life when they are younger and older than when they are middle-aged. A number of other researchers have reached the same conclusion (Blanchflower and Oswald, 2008; Sanfey and Teksoz, 2007; Pierewan and Tampubolon, 2015). Our investigation indicates that Poles were the least satisfied with their life at around age 54, a higher age than the turning point for most developed countries, which is typically in the forties (Blanchflower and Oswald, 2008).

As in other studies, we found that being a member of a political party or union has a positive effect on QoL, while being disabled has a negative one (Wang and VanderWeele, 2011; Christoph, 2010).

A widowed individual is likely to be less satisfied than one who is married. The same can be said of those who are divorced or separated. This confirms the findings of other studies (Sanfey and Teksoz, 2007; Pierewan and Tampubolon, 2015).

Education may be one of the most important and consistent determinants of QoL. As a human capital indicator, this covariate predicts the well-being. A number of studies have also investigated the relation between education and QoL (Betti et al., 2016; Sanfey and Teksoz, 2007). In general, the impact of education on satisfaction with one's life is ambiguous across the studies we analysed: there is no clear correlation. Malešević Perović (2010) found a positive correlation, while Clark and Oswald (1994) uncovered a negative one. Still, others have observed a mixed correlation: Betti et al. (2016) found that people with a middle level of education were the most satisfied. Finally, Sanfey and Teksoz (2007) stated that there is no correlation between happiness and education in transition countries. In our study, QoL tended to rise alongside the level of education.

Also in line with other studies (Gerdtham and Johannesson, 2001; Requena, 2016), our results show that living in the countryside or in towns with less than 20,000 inhabitants improved the perception of QoL. Requena (2016) observed that in wealthier countries, rural living standards are high enough to create a higher level of subjective well-being; while in less developed countries the rural environment cannot compete with urban resources for creating subjective well-being. Also in agreement with other research, we found territorial differences in the QoL (Cracolici et al., 2014; Malešević Perović, 2010). Comparing the Central Region, where Poland's capital city Warsaw is located, the South, the Northwest and the North exhibit significantly better QoL, while the East and the Southwest were perceived as significantly worse and not significantly worse, respectively.

With regard to type of household, we stated that the composition of the household affected the perception of QoL. Married couples with three children, single-parent families, non-family households perceived their situation as significantly worse than married couples. In this respect, we did not find a significant difference between married couples and the remaining types (i.e. married couples with one child and with two children and multi-family households). The impact of the composition of the household on subjective well-being has been confirmed by many studies. For example, Cracolici et al. (2014) found that couples with no children were better off than others, while Betti et al. (2016) found that one-person households were in a worse situation than others.

Our results show the impact of a socio-economic group identified on the basis of the household's main source of income. Others reported similar findings on the influence of socio-economic group membership on a subjective perception of QoL (Cracolici et al., 2014; Wang and VanderWeele, 2011). Setting employees as the reference group, we found retirees, pensioners and those living on unearned sources other than retirement pay and pension to be in a significantly worse situation, while the self-employed and farmers exhibited not significantly worse position. The members of households living on unearned sources other than retirement pay and pensions were often the unemployed and poor. Such households generally assess various aspects of life with more pessimism than others. Sen (1997) mentioned a variety of reasons that unemployment may impact the QoL, including a lack of purpose in life, a lower social status and sense of self-esteem and a reduced sense of freedom and financial control.

Unlike the studies carried out for data from various European countries (Corazzini et al., 2012; Pierewan and Tampubolon, 2015), we found that in our study, gender does not reveal different patterns in explaining QoL.

Because this study is the first to explain QoL through the application of fractional outcome models, we considered various types of such models. As previously stated, the results concerning the estimates of significance and the impact of socio-economic and demographic variables obtained by the models considered in our study are very similar. In the next step we compared the models' goodness of fit. The predictive accuracy of the models is assessed using two performance measures: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Models with lower RMSE and MAE more accurately estimate the QoL indicator.

Table 3 Values of goodness of fit measures

Goodness of fit measure	FRM with logit link function	FRM with probit link function	BRM with logit link function	BRM with probit link function
RMSE	0.1748	0.1748	0.1756	0.1756
MAE	0.1396	0.1396	0.1406	0.1406

Source: Authors' calculations

The results reported in Table 3 show that the FRMs exhibit both RMSE and MAE only slightly better than the respective errors of all the BRMs. It should be also stressed that the Ramsey's RESET test reveals no misspecification of the conditional mean function in all estimated models. Thus, it cannot be determined to what extent one model is superior to another.

CONCLUDING REMARKS

This study has examined a new methodological framework for assessing the subjective perception of life by using methods of fuzzy set theory proposed by Betti (Betti et al., 2016; Betti, 2017). The main contribution of this analysis is its application of fractional outcome models to explain the quality of life through various socio-economic and demographic factors.

The data employed for the analysis came from the 'Social Diagnosis' survey conducted in 2015, a good deal of which was devoted to aspects of personal life. According to Betti's approach, the ordered data on subjective assessments were converted by a membership function into a [0,1] interval and then the synthetic QoL indicator encompassing all the aspects of life under consideration was computed. Because all of the QoL indicator values lay in the unit interval, we proposed to explain them using fractional outcome models. We applied a fractional regression model proposed by Papke and Wooldridge (1996) and a beta regression model developed by Ferrari and Cribari-Neto (2004). We included various socio-economic variables and demographic factors as explanatory variables: age, gender, education, civil status, disability, association membership, place of residence, household type and main source of income. We found that the QoL was U-shaped in age, minimizing around the age of 54. Furthermore, the perception of QoL increases with education, association membership, and decreases with disability, urbanisation, and being widowed or divorced. Results of our estimation indicate that the demographic composition of the household, region of residence and source of income all had a statistically significant impact. Our findings are largely in line with other studies.

It should be stressed that our study omits sociological nuances of the definition of 'quality of life' concept. Our goal is to demonstrate the potential for using modern methods to identify factors affecting the multidimensional indicator of QoL. The application of fractional outcome models has many advantages. Such models allow the assessment of whether given socio-economic and demographic factor is associated with response variable bounded by 0 and 1 while controlling the outcomes overlapping associations with other explanatory variables. Also, their ability to capture non-linearities is an important advantage.

We hope that our study with using fractional outcome models approach can provide some insight into the subjective perception of the quality of life. We plan various extensions of our study. Future research could apply panel data models for controlling unobserved heterogeneity of individuals and monitoring changes of QoL over time.

References

BETTI, G. AND VERMA,V. Measuring the degree of poverty in a dynamic and comparative context: A multidimensional approach using fuzzy set theory. *Proceedings of the ICCS-VI.* Lahore, Pakistan, 1999, 11, pp. 289–301.

BETTI, G. AND VERMA, V. Fuzzy measures of the incidence of relative poverty and deprivation: a multi-dimensional perspective. *Statistical Methods and Applications*, 2008, 17(2), pp. 225–250.
BETTI, G., SOLDI, R., TALEV I. Fuzzy multidimensional indicators of quality of life: The empirical case of Macedonia. *Social Indicators Research*, 2016, 127(1), pp. 39–53.
BETTI, G. Fuzzy measures of quality of life in Germany: a multidimensional and comparative approach. *Quality and Quantity*, 2017, 51(1), pp. 23–34.
BLANCHFLOWER, D. G. AND OSWALD, A. J. Is well-being U-shaped over the life cycle? *Social Science & Medicine*, 2008, 66, pp. 1733–1749.
CARDOSO, A. R., FONTAINHA, E., MONFARDINI, C. Children's and parents' time use: empirical evidence on investment in human capital in France, Germany and Italy. *Review of Economics of the Household*, 2010, 8, pp. 479–504.
CARRASCO, J. M. F., FERRARI, S. L. P., ARELLANO-VALLE, R. B. Errors-in-variables beta regression models. *Journal of Applied Statistics*, 2014, 41(7), pp. 1530–1547.
CERIOLI, A. AND ZANI, S. A fuzzy approach to the measurement of poverty. In: DAGUM, C., ZENGA M., eds. *Income and Wealth Distribution, Inequality and Poverty.* Berlin: Springer, 1990.
CHELI, B. AND LEMMI, A. A totally fuzzy and relative approach to the multidimensional analysis of poverty. *Economic Notes*, 1995, 24, pp. 115–134.
CHRISTOPH, B. The relation between life satisfaction and the material situation: A re-evaluation using alternative measures. *Social Indicators Research*, 2010, 98(3), pp. 475–499.
CLARK, A. E. AND OSWALD, A. J. Unhappiness and Unemployment. *Economic Journal*, 1994, 104(424), pp. 648–659.
CORAZZINI, L., ESPOSITO, L., MAJORANO, F. Reign in hell or serve in heaven? A cross-country journey into the relative vs absolute perceptions of wellbeing. *Journal of Economic Behavior & Organization*, 2012, 81(3), pp. 715–730.
CRACOLICI, M. F., GIAMBONA, F., CUFFARO, M. Family structure and subjective economic well-being: some new evidence. *Social Indicators Research*, 2014, 118, pp. 433–456.
CRIBARI-NETO, F. AND ZEILEIS, A. Beta Regression in R. *Journal of Statistical Software*, 2010, 34(2), pp. 1–24.
CZAPIŃSKI, J. Indywidualna jakość i styl życia. *Contemporary Economics*, 2015, 9/4, pp. 200–331.
CZAPIŃSKI, J., PANEK, T., et al. Diagnoza Społeczna 2015. *Warunki i Jakość Życia Polaków – Raport.* Warsaw, 2015, available at the website: <*www.diagnoza.com*>.
CZARNITZKI, D. AND KRAFT, K. Firm leadership and innovative performance: evidence from seven EU countries. *Small Business Economics*, 2004, 22, pp. 325–332.
DE BATTISTI, F., MARASINI, D., NICOLINI, G. A measure of job satisfaction by means of fuzzy set theory. *Statistica Applicata – Italian Journal of Applied Statistics*, 2015, 23(3), pp. 361–374.
DESAI, M. AND SHAH, A. An econometric approach to the measurement of poverty. *Oxford Economic Papers*, 40(3), 1988, pp. 505–522.
DUDEK, H. AND SZCZESNY, W. Subjective perception of quality of life – multidimensional analysis based on fuzzy sets approach. *Wrocław University of Economics Research Papers, Quality of Life, Human and Ecosystem Well-being*, 2016, 435, pp. 55–68.
DUDEK, H. AND SZCZESNY, W. Zastosowanie funkcji przynależności w analizie subiektywnego postrzegania jakości życia. *Ekonometria*, 2015, 50(40), pp. 62–78.
EUROSTAT. *Quality of life indicators – measuring quality of life* [online]. Eurostat, 2017. <*http://ec.europa.eu/eurostat/statistics-explained/index.php/Quality_of_life_indicators_-_measuring_quality_of_life*>.
FERRARI, S. L. P. AND CRIBARI-NETO, F. Beta Regression for Modeling Rates and Proportions. *Journal of Applied Statistics*, 2004, 31(7), pp. 799–815.
FILIPPONE, A., CHELI, B., D'AGOSTINO, A. Addressing the interpretation and the aggregation problems in totally fuzzy and relative poverty measures. *ISER Working Paper Series number 2001–22*, University of Essex, 2001.
FLORES, G., INGENHAAG, M., MAURER, J. An anatomy of old-age disability: Time use, affect and experienced utility. *Journal of Health Economics*, 2015, 44, pp. 150–160.
FUWA, N., ITO, S., KUBO, K., KUROSAKI, T., SAWADA, Y. Gender discrimination, intrahousehold resource allocation, and importance of spouses' fathers: evidence on household expenditure from rural India. *Developing Economies*, 2006, 44(4), pp. 398–439.
GERDTHAM, U.-G. AND JOHANNESSON, M. The relationship between happiness, health, and socioeconomic factors: results based on Swedish microdata. *Journal of Socio-Economics*, 2001, 30(6), pp. 553–557.
GLATZER, W. Challenges for Quality of Life. In: GLATZER, W., VON BELOW, S., STOFFREGEN, M., eds. *Challenges for Quality of Life in the Contemporary World.* Dordrecht, Boston, London: Kluwer, 2004.
GRZYBOWSKA, U. AND KARWAŃSKI, M. Application of Mixed Models and Families of Classifiers to Estimation of Financial Risk Parameters. *Quantitative Methods in Economics*, 2015, 16(1), pp. 108–115.
GUIO, A.-C. What can be learned from deprivation indicators in Europe? *Eurostat methodologies and working paper*, Luxembourg: Eurostat, 2009.

HOE, J., ORRELL, M., LIVINGSTON, G. Quality of life measures in old age. In: ABOU-SALEH, M. T., KATONA, C., KUMAR, A., eds. *Principles and practice of geriatric psychiatry.* London: Wiley, 2011.
KARWAŃSKI, M., GOSTKOWSKI, M., JAŁOWIECKI, P. Loss given default modeling. *Journal of Risk Model Validation*, 2015, 9(3), pp. 23–40.
LAZIM, M. A. AND OSMAN, M. T. A. A new Malaysian quality of life index based on fuzzy sets and hierarchical needs. *Social Indicators Research*, 2009, 94(3), pp. 499–508.
MALEŠEVIĆ PEROVIĆ, L. M. Life satisfaction in Croatia. *Croatian Economic Survey*, 2010, 12(1), pp. 45–81.
NEFF, D. Fuzzy set theoretic applications in poverty research. *Policy and Society*, 2013, 32, pp. 319–331.
PANEK, T. Multidimensional approach to poverty measurement: Fuzzy measures of the incidence and the depth of poverty. *Statistics in Transition*, 2010, 11(2), pp. 361–379.
PAOLINO, P. Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, 2001, 9(4), pp. 325–346.
PAPKE, L. E., WOOLDRIDGE, J. M. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics*, 1996, 11, pp. 619–632.
PHILLIPS, D. *Quality of Life: Concept, Quality, Practice.* London: Routledge, 2006.
PI ALPERIN, M. N., VAN KERM, P. MDEPRIV: Stata module to compute synthetic indicators of multiple deprivation. *Statistical Software Components S457806*, Boston College Department of Economics, 2014.
PIEREWAN, A. C. AND TAMPUBOLON, G. Happiness and health in Europe: A multivariate multilevel model. *Applied Research in Quality of Life*, 2015, 10(2), pp. 237–252.
RAMALHO, J. J. S. AND VIDIGAL DA SILVA, J. Functional form issues in the regression analysis of financial leverage ratios. *Empirical Economics*, 2013, 44(2), pp. 799–831.
RAMALHO, E. A, RAMALHO, J. J. S., MURTEIRA, J. M. R. Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys*, 2011, 25(1), pp. 19–68.
REQUENA, F. Rural–urban living and level of economic development as factors in subjective well-being. *Social Indicators Research*, 2016, 128(2), pp. 693–708.
ROGERS, J. A, POLHAMUS, D., GILLESPIE, W. R. et al. Combining patient-level and summary-level data for Alzheimer's disease modeling and simulation: a beta regression meta-analysis. *Journal of Pharmacokinetics and Pharmacodynamics*, 2012, 39(5), pp. 479–498.
ROJO-PÉREZ, F., FERNÁNDEZ-MAYORALAS, G., RODRÍGUEZ-RODRÍGUEZ, V. Global Perspective on Quality in Later Life. In: GLATZER, W., CAMFIELD, L., MØLLER, V., ROJAS M., eds. *Global Handbook of Quality of Life. International Handbooks of Quality-of-Life.* Dordrecht: Springer, 2015.
SANFEY, P. AND TEKSOZ, U. Does transition make you happy? *Economics of Transition*, 2007, 15(4), pp. 707–731.
SEN, A. Inequality, unemployment and contemporary Europe. *International Labour Review*, 1997, 136(2), pp. 155–72.
SMITHSON, M. AND VERKUILEN, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, 2006, 11(1), pp. 54–71.
STIGLITZ, J. E., SEN, A., FITOUSSI, J. *Report by the commission on the measurement of economic performance and social progress.* 2009, retrieved from: <*www.stiglitz-sen-fitoussi.fr*>.
WANG, P. AND VANDERWEELE, T. J. Empirical research on factors related to subjective wellbeing of Chinese urban residents. *Social Indicators Research*, 2011, 101(3), pp. 447–459.
ZADEH, L. A. Fuzzy sets. *Information and Control*, 1965, 8, pp. 338–353.
ZANIN, L. The effects of various motives to save money on the propensity of Italian households to allocate an unexpected inheritance towards consumption. *Quality and Quantity*, 2017, 51(4), pp. 1755–1775.

APPENDIX

Table A1 Items in the individual questionnaire concerning respondent satisfaction with regard to particular areas and aspects of life[13]

Please, assess the specific areas of your life and state to what extent you are satisfied with them. Please, give your answers by crossing the box next to the appropriate digit for the given area of life. The specific digits mean:

1 – VERY SATISFIED
2 – SATISFIED
3 – RATHER SATISFIED
4 – RATHER NOT SATISFIED
5 – NOT SATISFIED
6 –VERY NOT SATISFIED
7 – not applicable

To what extent are you satisfied with:

1.	your relations with your close family members	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
2.	the financial situation of your family	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
3.	your relations with friends (a group of friends)	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
4.	your health condition	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
5.	your life achievements	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
6.	the situation in the country	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
7.	your housing conditions	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
8.	the town/city you live in	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
9.	your future prospects	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
10.	your sex life	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
11.	your education	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
12.	the manner in which you spend your free time	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
13.	your work	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
14.	children	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
15.	marriage	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □
16.	safety in your town/city of residence	1 □ 2 □ 3 □ 4 □ 5 □ 6 □ 7 □

Source: Own construction based on (Czapiński and Panek, 2015)

Table A2 List and description of explanatory variables

Variable	Description
Age	The individual's age
Age2	The individual's age squared
Female	1 if the individual is female
Civil state	Four groups of formal civil states are considered:
married	1 if married
unmarried	1 if unmarried
widowed	1 if widowed
divorced/separated	1 if divorced or separated
Education	The educational level achieved by the individual is classified as:
1 (primary)	primary or lower
2 (basic vocational)	basic vocational or lower-secondary

[13] All items of questionnaire can be found on the website: <*www.diagnoza.com*> (Czapiński and Panek, 2015).

Table A2 List and description of explanatory variables (continuation)

Variable	Description
3 (secondary)	secondary
4 (higher)	higher or post-secondary
Disability	1 if the individual is disabled
Association membership	1 if the individual is a member of any organization, party or clubs
Class of place of residence	The class of place of residence is divided into urban and rural areas, with urban areas further subdivided by resident size units:
Very big town	Towns over 500 000 residents
Big town	Towns with 200 000–500 000
Medium-sized town	Towns with 100 000–200 000 residents
Small town	Towns with 20 000–100 000 residents
Very small town	Towns up to 20 000 residents
Village	Rural areas
Regions	Regions are the first level NUTS regions of the European Union. They include corresponding second-level sub-regions:
Central	Łódź, Mazovia
South	Lesser Poland, Silesia
East	Lublin, Subcarpathian, Świętokrzyskie, Podlaskie
Northwest	Greater Poland, West Pomerania, Lubusz
Southwest	Lower Silesia, Opola Voivodeship
North	Kuyavian-Pomeranian, Varmia-Masuria, Pomerania
Household type	Household type was established on the basis of the number of families and biological family type
MC without children	married couples with no children
MC with 1 child	married couples with one child
MC with 2 children	married couples with two children
MC with 3+ children	married couples with three or more children
Single-parent	single-parent families
Multi-family	multi-family households
One-person	non-family one-person households
Non-family	non-family multi-person households
The socio-economic group	The socio-economic group is identified on the basis of the household's main source of income. The following groups of households are taken into account:
Employees	households where the sole or main (dominant) source of income is from gainful employment in the public or private sector and from performing home-based work or on the basis of agency agreements
Self-employed	households whose exclusive or main (prevailing) source of income is self-employment (other than from private farming)
Farmers	households where the sole or main (dominant) source of income is from a farm with agricultural land exceeding 1 ha (including users of plots up to 1 ha of agricultural land and owners of domestic animals but no agricultural land if the livestock is the sole or main source of income)
Retirees	households where the sole or main (dominant) source of income is a retirement pension
Pensioners	households where the sole or main (dominant) source of income is a form of disability welfare support
Living on unearned sources	households where the sole or main (dominant) source of income are sources other than paid work (except for retirement pension, disability benefit or other type of pension)

Source: Own construction

The Evaluation of a Concomitant Variable Behaviour in a Mixture of Regression Models

Kristýna Vaňkátová[1] | *Palacký University Olomouc, Czech Republic*
Eva Fišerová | *Palacký University Olomouc, Czech Republic*

Abstract

Finite mixture of regression models are a popular technique for modelling the unobserved heterogeneity that occurs in the population. This method acquires parameters estimates by modelling a mixture conditional distribution of the response given explanatory variables. Since this optimization problem appears to be too computationally demanding, the expectation-maximization (EM) algorithm, an iterative algorithm for computing maximum likelihood estimates from incomplete data, is used in practice. In order to specify different components with higher accuracy and to improve regression parameter estimates and predictions the use of concomitant variables has been proposed. Based on a simulation study, performance and obvious advantages of concomitant variables are presented. A practical choice of appropriate concomitant variable and the effect of predictors' domains on the estimation are discussed as well.[2]

Keywords

Mixture of regression models, linear regression, EM algorithm, concomitant variable

JEL code

C11, C38, C51, C52

INTRODUCTION

The basic requirement for the proper use of a standard linear regression model is a homogeneity in the studied population. If this assumption is violated and a standard regression model is inapplicable due to several heterogeneous groups in data, an alternative approach to modelling by means of mixture of regression models can be utilized (DeSarbo and Cron, 1988; McLachlan and Peel, 2000). While a standard regression mainly aims to estimate regression parameters, a mixture of regression models is also used as a tool for data clustering and therefore works as a clusterwise regression.

Mixtures of linear regression models, originally called switching regressions, are a special case of mixture density models (also known as a mixture of distributions) that were initially studied by means of a moment-generating function (Pearson, 1894). Recently, however, a likelihood point of view has been preferred for mixture models with a fixed number of components. A standard technique to obtain the maximum likelihood estimates is the expectation-maximization (EM) algorithm (Dempster et al., 1977).

[1] 17. listopadu 12, 771 46 Olomouc, Czech Republic. E-mail: kristyna.vankatova@upol.cz, phone: (+420)733348062.
[2] This article is based on contribution at the conference *Robust 2016*.

In addition to the method of moments and the maximum-likelihood approach, a variety of other methods have been proposed for estimating parameters in mixture densities. These methods include graphic procedures; an estimate determined by a least squares criterion in the spirit of the minimum-distance method; a procedure based on a linear operator reducing the variances of the component densities; the confusion matrix method and related methods; a stochastic approximation algorithm; and a minimum chi-square estimation. A short description of these and related methods can be found in Redner and Walker (1984) along with necessary references.

Modelling of unobserved heterogeneity using a maximum likelihood methodology is presented for instance in Bengalia et al. (2009), De Veaux (1989), DeSarbo and Cron (1988), and Faria and Soromenho (2010). An extensive review of finite mixture models can be found in McLachlan and Peel (2000). The methodology of mixtures of regression models can be applied in various research fields, such as climatology, biology, economics, medicine and genetics; see e.g. Grün et al. (2012), Vaňkátová and Fišerová (2016), and Hamel et al. (2016).

Grün and Leisch (2008) proposed the concomitant variable models for the component weights that allow to allocate the data into the mixture components through other variables called concomitant. This extension can provide both more precise parameter estimates and better components identification. Since the concomitant variable is still a new concept in mixture modelling, the aim of this paper is to evaluate its role. Accordingly, a simulation study was conducted and results concerning the impact of the concomitant variable on the model quality are presented. Both precision of regression parameters and clusterwise properties of the model are addressed in cases of categorical and continuous concomitant variables. A practical choice of appropriate concomitant variable and the effect of predictors' domains on the estimation are discussed as well.

This paper is structured as follows. In Section 1, some fundamentals of mixtures of linear regression models with and without concomitant variables are presented. The theory behind parameters estimation is summarized in Section 2. Section 3 is dedicated to a simulation study investigating the performance of mixture models with and without concomitant variables. At the end, the conclusions of the study are drawn and additional comments are given.

1 REGRESSION MODELS

1.1 Mixtures of regression models

A mixture distribution (Pearson, 1894) is the probability distribution of a random variable obtained from a set of other random variables in such a way that, firstly, a random variable from the set is drawn according to given probabilities that sum to one; and that, secondly, the value of the selected variable is realized. Formally, the probability density function f can be represented by a convex combination of probability density functions f_i:

$$f(y) = \sum_{i=1}^{c} \pi_i f_i(y), \tag{1}$$

where f_i are called component densities and $\pi_1, \ldots, \pi_c$ are positive mixing proportions that sum to one. A Gaussian mixture distribution assumes that all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters (McLachlan and Peel, 2000).

Introduced by Goldfeld and Quandt (1976) as switching regressions, the mixture of regression models is formed analogously to the mixture distribution (1). Let Y_j denote the response variable, and let $\boldsymbol{x}_j$ denote the vector of predictors for the jth subject. Assuming the random errors are normally distributed and subpopulations are present within an overall population, the response variable Y_j is given as a finite sum (mixture) of conditional univariate normal densities $\boldsymbol{\varphi}$ with the expectation $\boldsymbol{x}_j^T \boldsymbol{\beta}_i$, and the variance

$\sigma_i^2, i = 1, \dots, c$. Following the mixture models structure, the conditional density of $Y_j \mid \boldsymbol{x}_j$ is defined by (Bengalia et al., 2009):

$$f\left(y_j \middle| \boldsymbol{x}_j, \boldsymbol{\Psi}\right) = \sum_{i=1}^{c} \pi_i \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right) = \sum_{i=1}^{c} \pi_i (2\pi\sigma_i^2)^{-\frac{1}{2}} \mathrm{e}^{\frac{-\left(y_j - \boldsymbol{x}_j^T \boldsymbol{\beta}_i\right)^2}{2\sigma_i^2}}. \tag{2}$$

The symbol $\boldsymbol{\Psi}$ denotes the vector of all unknown parameters for a mixture of regression models with c components:

$$\boldsymbol{\Psi} = \left(\pi_1, \dots, \pi_c, (\boldsymbol{\beta}_1^T, \sigma_1^2), \dots, (\boldsymbol{\beta}_c^T, \sigma_c^2)\right)^T,$$

where $\boldsymbol{\beta}_i$ denotes the q-dimensional vector of unknown regression parameters for the ith component and σ_i^2 is the unknown error variance for the ith component. The mixing proportions $\pi_1, \dots, \pi_c$ satisfy the conditions $\pi_i > 0$ and $\sum_{i=1}^{c} \pi_i = 1$. A more transparent way of expressing a mixture of regression models is:

$$Y_j = \begin{cases} \boldsymbol{x}_j^T \boldsymbol{\beta}_1 + \varepsilon_{1j} \text{ with probability } \pi_1, \\ \boldsymbol{x}_j^T \boldsymbol{\beta}_2 + \varepsilon_{2j} \text{ with probability } \pi_2, \\ \vdots \\ \boldsymbol{x}_j^T \boldsymbol{\beta}_c + \varepsilon_{cj} \text{ with probability } \pi_c, \end{cases} \tag{3}$$

where ε_{ij} are independent random errors with a normal distribution $\mathrm{N}(0, \sigma_i^2)$, $i = 1, \dots, c$, $j = 1, \dots, n$.

1.2 Mixtures of regression models with concomitant variables

The mixture of regression models consists of c components where each component follows a specific parametric distribution. Each individual component has been assigned a weight indicating the prior probability for an observation to come from this component. Hence, the mixture distribution is given by the weighted sum over c components with weights corresponding to the prior probabilities. If the weights depend on further variables, the latter are referred to as concomitant variables. The mixture of regression models with concomitant variables was introduced and is described in detail by Grün and Leisch (2008).

The mixture of regression models with s concomitant variables is in the form of:

$$f\left(y_j \middle| \boldsymbol{x}_j, \boldsymbol{\omega}_j\right) = \sum_{i=1}^{c} \pi_i\left(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right), \tag{4}$$

where $\boldsymbol{\omega}_j$ denotes the s-dimensional vector of concomitant variables for the jth observation. The symbol $\boldsymbol{\alpha}_i$ denotes the vector of parameters of the concomitant variable model for the ith component. The dimension of $\boldsymbol{\alpha}_i$ relates to the chosen concomitant model and the dimension of concomitant variables. Dimensions of these vectors remain the same over all observations.

The set of unknown parameters for a mixture of regression models with concomitant variables with c components is:

$$\boldsymbol{\Psi} = \left((\boldsymbol{\alpha}_1^T, \boldsymbol{\beta}_1^T, \sigma_1^2), \dots, (\boldsymbol{\alpha}_c^T, \boldsymbol{\beta}_c^T, \sigma_c^2)\right)^T.$$

The component weights π_i need to satisfy conditions:

$$\sum_{i=1}^{c} \pi_i\left(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i\right) = 1, \qquad \text{for } j = 1, \dots, n, \tag{5}$$

$$\pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) > 0, \qquad \text{for } i = 1, \dots, c, \text{and } j = 1, \dots, n.$$

Although the function of a concomitant variable model may have an arbitrary form, it has to fulfill the conditions (5). In this paper, a multinomial logit model for the π_i is considered, as seen below:

$$\pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i) = \frac{e^{\omega_j^T \alpha_i}}{\sum_{h=1}^{c} e^{\omega_j^T \alpha_h}} \quad \text{for} \quad i = 1, \dots, c, \; j = 1, \dots, n, \tag{6}$$

with $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \dots, \boldsymbol{\alpha}_c^T)^T$ and $\boldsymbol{\alpha}_1 \equiv \mathbf{0}$. This settings means that the first component is a baseline. The vector $\boldsymbol{\alpha}_i$ is s-dimensional, $i = 1, \dots, c$, provided the model contains s concomitant variables (Grün and Leisch, 2008).

A classical linear regression model can be applied to heterogeneous population problem only in a case when the component membership of every observation is deterministically known or described by the observable random variable. As the result of the first option (deterministically determined membership), c independent regression models are analyzed separately. Concerning the latter scenario, the cluster identification information in a form of a categorical random variable is included in the model as dummy variables (indicators of categories) together with interactions between predictors. However, both suggested approaches are inapplicable in the situation discussed in this paper since the cluster membership of observations is considered to be latent.

The categorical concomitant variable can be potentially used in a classical linear regression model as a random variable carrying the information about a cluster membership but the effect of such a variable on the estimated model is highly exaggerated. Also, a number of other problems arise in this case. For example, there is a problem with a number of categories versus a number of components. In addition, it is not ideal that the assignment of an observation to the cluster is no longer weighted but fixed as 1 or 0.

Mixtures of regression models are frequently used specifically for theirs clustering properties. Unlike classical clustering methods, mixture regression models are able to deal with clustering of the data following a certain function, therefore we refer to the clusterwise regression method.

2 PARAMETERS ESTIMATION

In order to obtain parameters estimates for a standard mixture of regression models with a fixed number of components c, the log-likelihood function is maximized:

$$\log L(\boldsymbol{\Psi}, \boldsymbol{x}_1, \dots, \boldsymbol{x}_n, y_1, \dots, y_n) = \sum_{j=1}^{n} \log\left(\sum_{i=1}^{c} \pi_i \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right)\right). \tag{7}$$

Within the framework of mixture models the observations are viewed as an incomplete data. The data consists of triples $\left(\boldsymbol{x}_j^T, y_j, \boldsymbol{z}_j^T\right)^T$, where $\boldsymbol{z}_j$ is an unobserved vector indicating from which mixture component the observation $\left(\boldsymbol{x}_j^T, y_j\right)^T$ is drawn. More precisely, z_{ij} is equal to one if the observation $\left(\boldsymbol{x}_j^T, y_j\right)^T$ comes from the ith component; otherwise z_{ij} is zero. These values z_{ij} are unobservable and therefore treated as missing, and the data are augmented by estimates of the component memberships, i.e. the estimated posterior probabilities τ_{ij} (McLachlan and Peel, 2000). Using the Bayes rule, any jth observation can be assigned to the ith cluster with a probability given by:

$$\tau_{ij} = \frac{\pi_i \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right)}{\sum_{h=1}^{c} \pi_h \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_h, \sigma_h^2\right)}. \tag{8}$$

Since mixing proportions sum to unity, the log-likelihood function can be optimized using the Lagrange multipliers method with $\sum_{i=1}^{c} \pi_i = 1$ constraint. In order to obtain stationary equations, we compute the first order partial derivatives of the augmented log-likelihood function and equate them to zero. In the next step, it is a matter of few simple modifications to acquire a new system of equations obviously corresponding to stationary equations of another optimization problem formulated as (DeSarbo and Cron, 1988):

$$\log L_c(\boldsymbol{\Psi}) = \sum_{i=1}^{c}\sum_{j=1}^{n} \tau_{ij} \log \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right). \tag{9}$$

The function (9) is called the expected complete log-likelihood due to the fact it works with the estimated posterior probabilities τ_{ij} instead of unobservable values z_{ij}. This particular structure gainfully lends itself to the development of the EM algorithm (Dempster et al., 1977), an iterative procedure which alternates between an Expectation step and a Maximization step. The EM algorithm takes advantage of the expected likelihood that is in general easier to maximize than the original one.

The EM algorithm is widely exploited in practice. The estimators are viewed as some form of a local maximum likelihood estimator (Behboodian, 1970). However, it is not guaranteed that the EM algorithm provides a global maximum. A complication may occur in the case of normal mixtures with component specific variances, where the log-likelihood is unbounded and attains $+\infty$ for certain values of the parameter space. For this specific case, the EM algorithm adds to its advantage and provides, according to many practitioners, rather reasonable solutions unlike algorithmic approaches of global character such as a gradient function based techniques. Although the EM algorithm is often used, there is surprisingly little theoretical knowledge available for this estimator. In fact, it might be unclear to which extent asymptotic properties of the EM algorithm estimators, such as consistency, asymptotic efficiency and asymptotic normality, hold (Nityasuddhi and Böhning, 2003).

In the E-step, posterior probabilities τ_{ij} are estimated. Consequently, the expected complete log-likelihood is maximized in the M-step and the vector of unknown parameters $\boldsymbol{\Psi}$ is updated. The $(k+1)$th iteration of the EM algorithm can be summarized as follows:

E-step: Given the observed data $\boldsymbol{y}$ and current parameter estimates $\widehat{\boldsymbol{\Psi}}^{(k)}$ in the kth iteration, replace the missing data z_{ij} by the estimated posterior probabilities:

$$\hat{\tau}_{ij}^{(k)} = \frac{\hat{\pi}_i^{(k)} \varphi\left(y_j \middle| \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_i^{(k)}, \hat{\sigma}_i^{2(k)}\right)}{\sum_{h=1}^{c} \hat{\pi}_h^{(k)} \varphi\left(y_j \middle| \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_h^{(k)}, \hat{\sigma}_h^{2(k)}\right)}. \tag{10}$$

M-step: Given the estimates $\hat{\tau}_{ij}^{(k)}$ for the posterior probabilities τ_{ij} (which are functions of $\widehat{\boldsymbol{\Psi}}^{(k)}$), obtain new estimates $\widehat{\boldsymbol{\Psi}}^{(k+1)}$ of the parameters by maximizing the expected complete log-likelihood:

$$Q\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}^{(k)}\right) = \sum_{i=1}^{c}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \log \varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right). \tag{11}$$

This maximization problem is equivalent to solving the weighted least squares problem, where the vector $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ of observations and the design matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ are each weighted by $\sqrt{\hat{\tau}_{ij}^{(k)}}$. That means that we get:

$$\widehat{\boldsymbol{\beta}}_i^{(k+1)} = \left(\boldsymbol{X}^T \boldsymbol{W}_i^{(k)} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{W}_i^{(k)} \boldsymbol{y} \tag{12}$$

for estimates of regression parameters, assuming the $n \times n$ matrix $\boldsymbol{W}_i^{(k)} = Diag\left\{\sqrt{\hat{\tau}_{i1}^{(k)}}, \ldots, \sqrt{\hat{\tau}_{in}^{(k)}}\right\}$ is a diagonal matrix of weights, and:

$$\hat{\sigma}_i^{2(k+1)} = \frac{\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_i^{(k+1)}\right)\boldsymbol{W}_i^{(k)}(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_i^{(k+1)})}{n} \tag{13}$$

for the error variance estimate.

Thus, the entire set of $\widehat{\boldsymbol{\beta}}_i^{(k+1)}$ is derived by performing c separate weighted least squares analyses. In the same spirit, $\hat{\sigma}_i^{2(k+1)}$ is estimated and, lastly, the estimates of the mixing proportions π_i are updated using:

$$\hat{\pi}_i^{(k+1)} = \frac{\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)}}{n} \quad \text{for } i = 1, \ldots, c. \tag{14}$$

The principle of parameters estimation is very similar for the mixture of regression models with concomitant variables. The expected complete log-likelihood function can be derived analogously to the previous case. Thus, the EM algorithm for the mixture models with concomitant variables follows the following two steps (Grün and Leisch, 2008):

E-step: Given the observed data $\boldsymbol{y}$ and current parameter estimates $\widehat{\boldsymbol{\Psi}}^{(k)}$ in the kth iteration, replace the missing data z_{ij} by the estimated posterior probabilities τ_{ij}:

$$\hat{\tau}_{ij}^{(k)} = \frac{\pi_i(\boldsymbol{\omega}_j, \widehat{\boldsymbol{\alpha}}_i^{(k)})\varphi\left(y_j \middle| \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_i^{(k)}, \hat{\sigma}_i^{2(k)}\right)}{\sum_{h=1}^{c} \pi_i(\boldsymbol{\omega}_j, \widehat{\boldsymbol{\alpha}}_h^{(k)})\varphi\left(y_j \middle| \boldsymbol{x}_j^T \widehat{\boldsymbol{\beta}}_h^{(k)}, \hat{\sigma}_h^{2(k)}\right)}. \tag{15}$$

M-step: Given the estimates $\hat{\tau}_{ij}^{(k)}$ for the posterior probabilities τ_{ij} (which are functions of $\widehat{\boldsymbol{\Psi}}^{(k)}$), obtain new estimates $\widehat{\boldsymbol{\Psi}}^{(k+1)}$ of the parameters $\boldsymbol{\Psi}$ by maximizing:

$$Q\left(\boldsymbol{\Psi}, \widehat{\boldsymbol{\Psi}}^{(k)}\right) = Q_1\left(\boldsymbol{\beta}_i, \sigma_i^2, i = 1, \ldots, c; \widehat{\boldsymbol{\Psi}}^{(k)}\right) + Q_2\left(\boldsymbol{\alpha}, \widehat{\boldsymbol{\Psi}}^{(k)}\right), \tag{16}$$

where:

$$Q_1\left(\boldsymbol{\beta}_i, \sigma_i^2, i = 1, \ldots, c; \widehat{\boldsymbol{\Psi}}^{(k)}\right) = \sum_{i=1}^{c}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \log\left(\varphi\left(y_j \middle| \boldsymbol{x}_j^T \boldsymbol{\beta}_i, \sigma_i^2\right)\right) \tag{17}$$

and:

$$Q_2\left(\boldsymbol{\alpha}, \widehat{\boldsymbol{\Psi}}^{(k)}\right) = \sum_{i=1}^{c}\sum_{j=1}^{n} \hat{\tau}_{ij}^{(k)} \log\left(\pi_i(\boldsymbol{\omega}_j, \boldsymbol{\alpha}_i)\right). \tag{18}$$

Formulas Q_1 and Q_2 can be maximized separately. The formula Q_1 is maximized using the weighted ML estimation of linear models with weights $\sqrt{\hat{\tau}_{ij}^{(k)}}$. The maximization of Q_1 gives new estimates $\widehat{\boldsymbol{\beta}}_i^{(k+1)}, \hat{\sigma}_i^{2(k+1)}$, $i = 1, \ldots, c$. The term Q_2 is maximized by means of the weighted ML estimation of multinomial logit models and provides new estimates $\widehat{\boldsymbol{\alpha}}_i^{(k+1)}$.

Initial values of regression parameters may be based on a random division of observations into c components, i.e. on initial $\hat{\tau}_{ij}^{(0)}$ probabilities, where for each observation y_j only one of these c probabilities

equals to 1 and the other ones are set to zero. The EM algorithm is stopped when the (relative) change of the log-likelihood is smaller than a chosen tolerance.

The number of components can be chosen by comparing information criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC) of various models, each with a different number of components.

3 SIMULATION STUDY

A simulation study is conducted to assess the performance of both a standard mixture of regression models and a mixture of regression models containing concomitant variables. The standard regression model could only be applied in the case of statistically significant concomitant variables that could be used as additional explanatory variables. However, such a model lacks the clustering properties that are essential in the following analysis; therefore, only mixtures of regression models are considered.

The study is mainly focused on the impact of the concomitant variable on the model quality (the accuracy of estimation and clustering), practical choice of appropriate concomitant variable and the effect of predictors' domains on the estimation. Accordingly, data are simulated under a two and three component mixture of linear regressions and concomitant variables are considered either categorical or continuous. The statistical software R (R Core Team, 2016) containing several extension packages for the estimation of a mixture of regression models is used. The results are built on flexmix package, introduced in Leisch (2004).

3.1 Design of the study

Each observation $(\boldsymbol{x}_j^T, y_j)^T$, $j = 1, \ldots, n$, is generated by the following scheme. Firstly, the component membership is determined. Assuming the observation comes from the ith component with the probability π_i, it is possible to randomly select the component membership by means of the outcome of the multinomial distribution with mixing proportions as multinomial probabilities. With established membership, the value of the predictor x for the assigned ith component is randomly generated from a given distribution (a uniform distribution on the interval $[x_L, x_U]$ or a normal distribution with parameters μ_x and σ_x^2). Next, a normal random error ε_{ij} with the mean 0 and variance σ_i^2 is generated. Finally, the observed value y_j is computed using the regression model form $\boldsymbol{x}_j^T \boldsymbol{\beta}_i + \varepsilon_{ij}$, where the true values of regression parameters $\boldsymbol{\beta}_i$ are considered. Two typical positions of the true regression lines are considered, in which the lines are either parallel or concurrent. The effect of these alternative positions is also studied.

In order to examine the performance of both mixture models (with and without a concomitant variable), the following statistical characteristics of estimators of $\boldsymbol{\Psi}$ are calculated:

- The mean square error of the regression parameter estimates over all replications:

$$MSEPAR(\hat{\psi}_p) = \frac{1}{M}\sum_{m=1}^{M}\left(\psi_p - \hat{\psi}_p^{(m)}\right)^2, \tag{19}$$

where ψ_p is the pth parameter of the vector $\boldsymbol{\Psi}$. While ψ_p is a true parameter, $\hat{\psi}_p^{(m)}$ is the final estimate of a given parameter in the mth replication, $m = 1, \ldots, M$. We desire to examine MSEPAR for all mixture model parameters, i.e. for regression coefficients $\boldsymbol{\beta}_i$, error variances σ_i^2, and mixing proportions $\pi_i, i = 1, \ldots, c$. For mixing proportions, however, the true values of component weights are not constant and vary over replications, denoted as $\pi_i^{(m)}$ (with increasing sample size, these values converge to the true mixing proportions). Hereby, the mean square error of mixing proportions is computed according to:

$$MSEPAR(\hat{\pi}_i) = \frac{1}{M}\sum_{m=1}^{M}\left(\pi_i^{(m)} - \hat{\pi}_i^{(m)}\right)^2. \tag{20}$$

- The mean variance of estimated regression parameters:

$$VAR(\hat{\psi}_p) = \frac{1}{M}\sum_{m=1}^{M} \mathrm{var}\left(\hat{\psi}_p^{(m)}\right). \tag{21}$$

  Here, $\mathrm{var}\left(\hat{\psi}_p^{(m)}\right)$ represents the estimate of a variance of the pth parameter estimator in the mth replication. The variance-covariance matrix of the regression parameters estimators is estimated by the inverted negative Hesse matrix of the full likelihood of the model (Grün and Leisch, 2008).
- The misclassification error:

$$\mathrm{Err}_{\mathrm{M}} = 1 - \frac{1}{nM}\sum_{m=1}^{M}\sum_{j=1}^{n} I\left(\hat{\boldsymbol{z}}_j = \boldsymbol{z}_j\right), \tag{22}$$

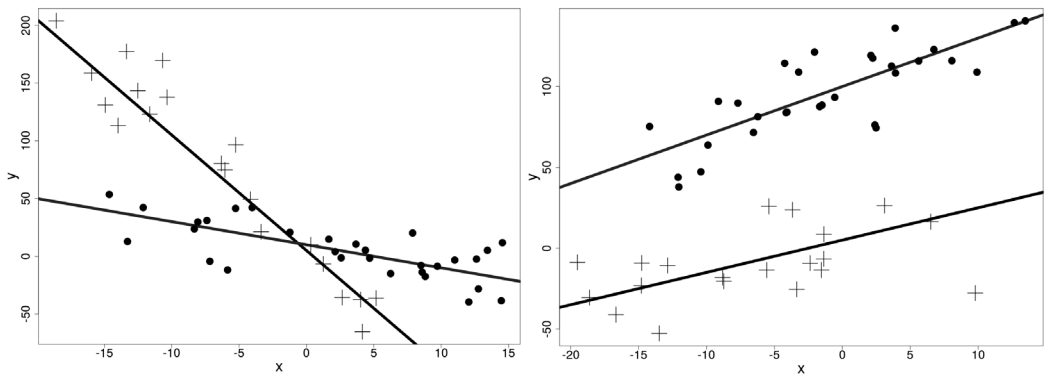
  where $\boldsymbol{z}_j$ is the true component membership of each observation and $\hat{\boldsymbol{z}}_j$ is its estimate. The misclassification error states a mean ratio of incorrectly assigned observations over all replications.

For the simplification, mixtures of regression lines are only considered in the following simulations. This simplification is not restrictive. The similar results are also valid in more complex regression models such as models with a polynomial trend.

3.2 Two component mixtures of linear regression models

For two component models, samples of three different sizes ($\mathrm{n} = 50, 100, 300$) are considered. Values of the predictor x are drawn from a uniform distribution on the interval $[-20,15]$ for both components. True parameter values (regression lines' coefficients and variances) are shown in Table 1 along with true mixing proportions. Scatter plots for samples of size 50 together with true regression lines are demonstrated in Figure 1. The number of replications is set to $M = 200$ considering how slow the algorithm is in practice.

Figure 1 Scatter plots for two configurations of mixtures of two regression lines of a sample of size 50 together with true regression lines



Source: Own construction

The concomitant variable is chosen as a categorical variable with four levels. Each of these levels labels the corresponding component with approximately 90% accuracy; values 1 and 2 label the first component, while values 3 and 4 label the second one. Since the concomitant variable is a univariate categorical variable, we can create three dummy variables that reflect the original variable in terms of a linear regression model.

Table 1 True parameter values for a two component mixture of regression lines

Position	β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	π_1	π_2
Parallel	100	3	15	5	2	20	6/10	4/10
Concurrent	10	–2	15	5	–10	20	6/10	4/10

Source: Own construction

Let us consider for example the level one as the reference category. Then, every level of the concomitant variable can be replaced with the 4-dimensional vector $\boldsymbol{\omega_j} = \left(1, \delta_{j2}, \delta_{j3}, \delta_{j4}\right)^T$, where δ_{jl} is an indicator of the level l for the jth observation, i.e. $\delta_{jl} = 1$ if the jth observation is labelled to the lth level, otherwise $\delta_{jl} = 0$. The resulting logit model is of the form:

$$\text{logit}\left[\pi_i\left(\boldsymbol{\omega_j}, \boldsymbol{\alpha}_i\right)\right] = \alpha_{i0} + \alpha_{i2}\delta_{j2} + \alpha_{i3}\delta_{j3} + \alpha_{i4}\delta_{j4}, \quad i = 1,2, \quad j = 1, \ldots, n, \tag{23}$$

meaning that the 90% accuracy of classification by a concomitant variable corresponds to a vector of parameters $\boldsymbol{\alpha}_1 = (\alpha_{10}, \alpha_{12}, \alpha_{13}, \alpha_{14})^T = (0,0,0,0)^T$ and $\boldsymbol{\alpha}_2 = (\alpha_{20}, \alpha_{22}, \alpha_{23}, \alpha_{24})^T = (-2.2,0,4.4,4.4)^T$. The vector $\boldsymbol{\alpha}_1$ is set to zero as the theory in Section 2 determines. To demonstrate the basic scheme of a concomitant model, we aim to show the selected probabilities π_i given by the multinomial logit model (23); for the clarity, the level l is also indicated:

$$\pi_1\left(l = 1, \boldsymbol{\omega}_j = (1,0,0,0)^T, \boldsymbol{\alpha}_1\right) = \frac{e^{\alpha_{10}}}{e^{\alpha_{10}} + e^{\alpha_{20}}} = 0.9\,, \tag{24}$$

$$\pi_1\left(l = 3, \boldsymbol{\omega}_j = (1,0,1,0)^T, \boldsymbol{\alpha}_1\right) = \frac{e^{\alpha_{10}+\alpha_{13}}}{e^{\alpha_{10}+\alpha_{13}} + e^{\alpha_{20}+\alpha_{23}}} = 0.1\,, \tag{25}$$

$$\pi_2\left(l = 1, \boldsymbol{\omega}_j = (1,0,0,0)^T, \boldsymbol{\alpha}_2\right) = \frac{e^{\alpha_{20}}}{e^{\alpha_{10}} + e^{\alpha_{20}}} = 0.9\,, \tag{26}$$

$$\pi_2\left(l = 3, \boldsymbol{\omega}_j = (1,0,1,0)^T, \boldsymbol{\alpha}_2\right) = \frac{e^{\alpha_{20}+\alpha_{23}}}{e^{\alpha_{10}+\alpha_{13}} + e^{\alpha_{20}+\alpha_{23}}} = 0.1\,. \tag{27}$$

Software R provides a detailed summary for the concomitant model, so that both parameters estimates and their significance test statistics are displayed.

The effect of a concomitant variable on the estimation in mixture models is visible on the resulting statistical characteristics of estimators, such as the mean square errors (MSEPAR), the mean variances (VAR) and the misclassification errors (Err_M), see Tables 2 and 3. It is rather obvious the concomitant variable helps to optimize parameters estimates in both regression lines configurations (parallel and concurrent). Its benefit is apparent mainly for a small sample size. In case of a parallel model of a sample of size 50, the MSEPAR is about 1.7-fold to 3.2-fold smaller for a concomitant model than for a standard mixture. With an increasing sample size, the MSEPAR from both models are comparable. The accuracy of estimators is slightly higher in a model with a concomitant variable. The same tendency is also valid for the accuracy of mixing proportions (Table 3). For two component mixtures, the MSEPAR of both mixing proportions is the same. The MSEPAR of $\hat{\pi}_1$ is minor in both mixture models, and the difference is most significant for the parallel position of regression lines. It is less than 0.1%, with the exception for a sample of size 50, when the mixture model with a concomitant variable is used. For the standard mixture, the MSEPAR ($\hat{\pi}_1$) is 2-fold greater for both parallel and concurrent position, except sample size

Table 2 The mean square error (*MSEPAR*) and the mean variance (*VAR*) of the regression parameters, and standard error estimates for a two component mixture of regression models

Parallel							
MSEPAR		β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2
n = 50	standard	24.1535	0.4206	10.3827	74.1701	1.1376	32.1636
	concomitant	8.7147	0.1405	4.9659	44.6211	0.3585	14.7015
n = 100	standard	7.1550	0.1165	3.5546	13.8645	0.3098	8.7427
	concomitant	4.0437	0.0606	1.8087	14.9526	0.1335	6.7200
n = 300	standard	1.7086	0.0244	0.7518	5.0777	0.0452	1.9845
	concomitant	1.3909	0.0166	0.6388	4.3677	0.0403	1.6901
VAR							
n = 50	standard	8.6111	0.1200	0.0207	33.4614	0.3381	0.0310
	concomitant	7.4635	0.1003	0.0181	25.9662	0.2717	0.0273
n = 100	standard	4.1556	0.0566	0.0098	13.4476	0.1390	0.0157
	concomitant	3.8123	0.0517	0.0091	13.6837	0.1399	0.0144
n = 300	standard	1.3314	0.0180	0.0033	4.6263	0.0464	0.0052
	concomitant	1.2862	0.0173	0.0030	4.5854	0.0454	0.0048
Concurrent							
MSEPAR		β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2
n = 50	standard	12.6919	0.1459	5.9812	56.5914	0.5026	18.6374
	concomitant	9.1072	0.1245	5.2192	45.6830	0.3635	14.0798
n = 100	standard	4.6781	0.0528	2.8718	28.0681	0.2231	9.3213
	concomitant	4.5898	0.0609	2.4462	16.5644	0.1752	5.8688
n = 300	standard	1.4862	0.0210	0.8740	7.1206	0.0641	2.5656
	concomitant	1.2614	0.0155	0.7271	5.9025	0.0569	2.4797
VAR							
n = 50	standard	8.9620	0.1091	0.0258	39.0396	0.3477	0.0371
	concomitant	8.4352	0.1076	0.0206	31.1324	0.3036	0.0303
n = 100	standard	4.5738	0.0542	0.0125	20.3584	0.1774	0.0189
	concomitant	3.9854	0.0506	0.0103	16.3407	0.1526	0.0159
n = 300	standard	1.5218	0.0181	0.0041	6.6889	0.0582	0.0062
	concomitant	1.3803	0.0177	0.0034	5.2361	0.0489	0.0054

Source: Own construction

of 50. For the smallest sample size in this study, the difference in MSEPAR of $\hat{\pi}_1$ is more significant considering parallel configuration of regression lines. In this case the model with a concomitant variable achieves more than 9-fold better results in π_i estimation.

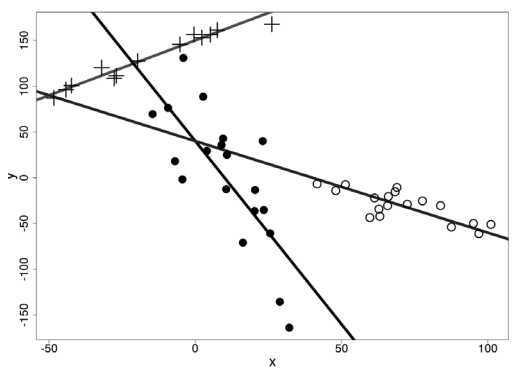
In addition, the misclassification error is considerably smaller when additional information on membership of observations is taken into account (Table 3). The misclassification error also depends on the configuration of regression lines in the mixture. For the mixture of concurrent lines, the misclassification error is more than 10% when the standard mixture model is used, while it decreases by half using the concomitant model. The classification is better for parallel regression lines. Although the misclassification error is still worse in the standard mixture (2% in contrast to 0.6% for a sample of size 50), with an increasing sample size the differences become negligible (0.6% and 0.3% for a sample of size 300).

Table 3 The misclassification errors ($\mathrm{Err}_{\mathrm{M}}$) and the mean square errors for the estimate of the first mixing proportion ($\mathrm{MSEPAR}(\hat{\pi}_1)$) for a two component mixture of regression lines

	n = 50		n = 100		n = 300	
Position	standard	concomitant	standard	concomitant	standard	concomitant
$\mathrm{Err}_{\mathrm{M}}$						
Parallel	0.0209	0.0061	0.0080	0.0038	0.0057	0.0034
Concurrent	0.1112	0.0502	0.1083	0.0437	0.1040	0.0416
$\mathrm{MSEPAR}(\hat{\pi}_1)$						
Parallel	0.0019	0.0002	0.0002	< 0.0001	< 0.0001	< 0.0001
Concurrent	0.0028	0.0017	0.0014	0.0007	0.0004	0.0002

Source: Own construction

Figure 2 The scatter plot of a three component mixture of a sample of size 50. True regression lines visualized

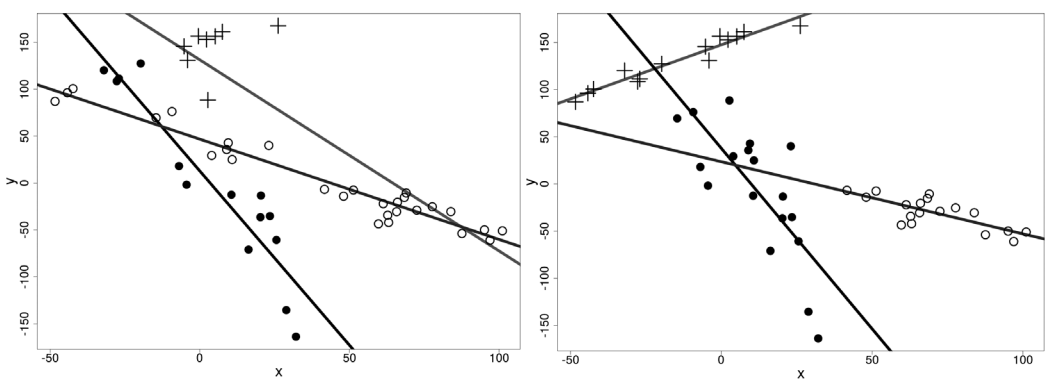


Source: Own construction

It should be noted that similarly to other clustering problems it is not guaranteed in general that the misclassification error converges to zero as n tends to infinity. It rather converges to some fixed value depending on the variance of parameter estimators and the distance or the angle between regression lines.

It is worth mentioning that even a random choice of a concomitant variable (a concomitant variable is generated as a completely random variable with zero correlation to the observation's component membership) does not affect this kind of mixture models in a negative way. It merely causes that the results given by a model including the concomitant variable are comparable to the results of a standard mixture model. This is due to the fact that a multinomial logit model describing

Figure 3 Fitted regression lines via a standard mixture of regression lines (left) and a mixture model with a continuous concomitant variable (right)



Source: Own construction

the effect of a concomitant variable on mixing proportions is only secondary in a process of clustering. A possible contribution of concomitant variables can be assessed by statistical significance of parameters in a multinomial logit model (6).

The mixture model tends to maintain its behaviour no matter how many categories the concomitant variable has. Favourable characteristics of mixture models containing concomitant variables are preserved even when the concomitant variable is continuous. The superiority of the mixture of regression models using concomitant variables does not deteriorate with a rising number of components.

3.3 Three component mixtures of linear regression models

In this section, a three component mixture and a continuous concomitant variable represented by the normally distributed predictor itself are investigated. The aim is to show how problematic the usage of mixture models is when components are defined on different parts of the predictor space, which is in our case the x-axis. In other words, if the values of the predictor x are generated from a uniform distribution, the interval $[x_L, x_U]$ is not the same for all components. Assuming normally distributed predictor, the mean μ_x is different for each component. Even relatively small nuances significantly affect estimates in a negative way, as it can be seen in the following example. In this type of a configuration of mixtures of regression functions, the estimation can be improved by using the predictor as a concomitant variable.

Table 4 True parameter values for a three component mixture of regression lines and probability distributions of a predictor

reg. parameters	β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	β_{30}	β_{31}	σ_3^2
	150	1.2	10	40	−4	40	40	−1	10
mixing proportions π_i	3/10			5/10			2/10		
$x \sim \mathrm{N}(\mu_x, \sigma_x^2)$	$\mathrm{N}(-20, 20^2)$			$\mathrm{N}(10, 20^2)$			$\mathrm{N}(70, 20^2)$		

Source: Own construction

The design of the mixture model containing three regression lines is presented in Table 4. The predictor x is considered as a concomitant variable ω and the logit of the mixing proportion π_i is assumed to be a linear function of a concomitant variable, as seen below:

$$\mathrm{logit}\left[\pi_i(\omega_j, \boldsymbol{\alpha}_i)\right] = \alpha_{i0} + \alpha_{i1}\omega_j, \quad i = 1,2,3, \qquad j = 1, \ldots, n. \tag{28}$$

Let us recall that the vector $\boldsymbol{\alpha}_1$ is set to zero. Apparently, the mixing proportions can be expressed as:

$$\pi_i(\omega_j, \boldsymbol{\alpha}_i) = \frac{\mathrm{e}^{\alpha_{i0}+\alpha_{i1}\omega_j}}{\mathrm{e}^{\alpha_{10}+\alpha_{11}\omega_j} + \mathrm{e}^{\alpha_{20}+\alpha_{21}\omega_j} + \mathrm{e}^{\alpha_{30}+\alpha_{31}\omega_j}}, \quad i = 1,2,3, \quad j = 1, \ldots, n. \tag{29}$$

An example of such a mixture for a sample of size 50 is visualized in Figure 2. Apart from visualization of a data coming from a given three component mixture, true regression lines for individual clusters are demonstrated. As it was indicated above, this particular mixture of regression models causes severe inaccuracy in estimates. This problematic phenomenon is noticeable in Figure 3, where one fitted regression line from a standard mixture model is completely inaccurate due to incorrect classification, while a model with a concomitant variable fits all lines correctly.

In this type of configuration, a standard mixture model in many cases does not even estimate the right number of components, let alone remotely accurate regression parameters and component memberships of observations. 2 000 simulations were performed and the ratio of these highly imprecise estimates was 79% for a sample size of 50 and even 93% for a sample size of 300 (Table 5), which indicates that this

is a systematic effect. Conversely, the proportion of inaccurate estimates obtained from a model with a concomitant variable is significantly smaller, accounting for only 27% for a sample size of 50 and decreasing to 6% for a sample size of 300. The estimates so dissimilar to the true parameter values that individual components cannot be recognized or efficiently identified were considered highly inaccurate. In practice, acceptance intervals for regression parameters $\boldsymbol{\beta}_i$ from a given component may be used. These intervals are as wide as possible to allow identification of a component and its distinction from the remaining components in the model. If no component or more components correspond to some acceptance interval, the whole mixture model is marked as inaccurate (see Figure 3, left).

Table 5 The ratio of entirely inaccurate estimates of parameters in a three component mixture of regression lines with a different space of the predictor from 2 000 simulations. The misclassification errors are evaluated from 200 correctly fitted models as well as the mean square errors for mixing proportion estimates

	n = 50		n = 100		n = 300	
	standard	concomitant	standard	concomitant	standard	concomitant
Inaccurate param. ratio	0.7850	0.2700	0.8200	0.1400	0.9250	0.0600
$\mathrm{Err}_{\mathrm{M}}$	0.2316	0.0703	0.1925	0.0456	0.1742	0.0392
$\mathrm{MSEPAR}(\hat{\pi}_1)$	0.0066	0.0018	0.0062	0.0005	0.0047	0.0002
$\mathrm{MSEPAR}(\hat{\pi}_2)$	0.0098	0.0046	0.0054	0.0010	0.0036	0.0002
$\mathrm{MSEPAR}(\hat{\pi}_3)$	0.0254	0.0023	0.0187	0.0003	0.0154	<0.0001

Source: Own construction

Table 6 The mean square errors (MSEPAR) and the mean variances (VAR) of the regression parameters, and standard error estimates for a three component mixture of regression lines with a different space of the predictor. Characteristics are calculated from 200 correctly fitted models

MSEPAR		β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	β_{30}	β_{31}	σ_3^2
n = 50	standard	35.7349	0.0388	42.2389	268.0571	0.3215	90.3459	277.8042	0.0527	33.0111
	concomitant	21.4600	0.1367	32.7982	154.9637	0.4086	74.5576	365.5693	0.0667	14.7752
n = 100	standard	7.6366	0.0096	6.4402	89.8041	0.1335	34.0492	194.2027	0.0367	12.4032
	concomitant	9.1976	0.0515	12.2703	55.8374	0.1272	24.2869	126.9932	0.0241	5.4217
n = 300	standard	2.3826	0.0029	2.2685	28.5231	0.0348	10.2868	91.1527	0.0178	5.2866
	concomitant	1.4447	0.0144	0.9264	21.0425	0.0384	6.3211	32.4438	0.0062	0.9661
VAR		β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	β_{30}	β_{31}	σ_3^2
n = 50	standard	16.3324	0.0314	0.0742	116.0794	0.1842	0.0329	40.9276	0.0097	0.0512
	concomitant	11.8080	0.0927	0.0425	103.0376	0.2121	0.0262	164.2889	0.0319	0.0497
n = 100	standard	6.0410	0.0086	0.0364	53.9657	0.0841	0.0151	26.5701	0.0056	0.0315
	concomitant	4.7048	0.0353	0.0235	49.8695	0.0975	0.0117	81.7340	0.0155	0.0260
n = 300	standard	2.0882	0.0030	0.0115	18.1297	0.0299	0.0047	13.1638	0.0027	0.0125
	concomitant	1.3823	0.0108	0.0080	16.6439	0.0314	0.0040	27.1625	0.0050	0.0087

Source: Own construction

In order to evaluate the quality of estimates, all entirely inaccurate estimated models were identified and discarded. Quality characteristics for both types of models evaluated from 200 correctly estimated models are reported in Table 6. In contrast to the previous simulation study (Table 2), the superiority of a model with a concomitant variable is not so apparent and the accuracy of estimators from both models is comparable. The misclassification error is still much worse in a standard mixture and the same

goes for the MSEPAR of mixing proportion estimates (Table 5). However, it should be kept in mind that the quality characteristics were calculated from the correctly estimated models and that a standard mixture tends to give entirely inaccurate results. Therefore, for this type of regression function configuration, a mixture model with a concomitant variable should be only used for the estimation.

CONCLUSION

The paper is focused on a concomitant variable introduced by Grün and Leisch (2008) and its role in the mixture of regression models. Two representative simulation studies were performed in order to assess the quality of regression estimates and clustering properties of both a standard mixture of regression models and a mixture of regression models with concomitant variables. Obviously, the possibilities of mixture models setting are various and this paper is focused only on two of them. However, the models presented here were chosen as a representative sample, assuming at the same time that each model works with different number of components, diverse distributions of predictor, various regression lines configuration and, most importantly, distinct characters of the concomitant variable.

The results of both studies indicate that the concomitant variables present a beneficial extension of mixture models. In case of a categorical concomitant variable, the results are straightforward and provide evidence in favour of a mixture model including a concomitant variable, since for this model, both the mean square error and the mean variance of estimates are, with very few exceptions, smaller. These characteristics are not so unambiguous for a three component mixture and a covariate as a concomitant variable. However, these indicators are only valid for a small portion of estimates that are close enough to the true values of parameters. In practice, the ratio of highly inaccurate estimates is more informative and is significantly reduced as a concomitant variable is added into the model.

Clustering properties are assessed through the mean misclassification error of each model. Again, a concomitant variable enhances estimated component membership in both cases, especially for a small sample size. In general, concomitant variables themselves prove to be useful in the mixture of regression models. Particularly, the concomitant variable in a form of the predictor itself seems to be a common choice for reasonable regression parameters estimates.

As models in the mixture get more complicated, estimates can become less precise and reliable. Nevertheless, the conclusions of the simulation study remain similar as the concomitant variables still enhance the performance of the mixture of regression models for both categorical and continuous concomitant variables.

ACKNOWLEDGMENT

The authors are grateful to the referees for their helpful comments that significantly improved this article. The authors also gratefully acknowledge the support received from the grant IGA_PrF_2016_025 and IGA_PrF_2017_019 of the Internal Grant Agency of Palacký University Olomouc.

References

BEHBOODIAN, J. On a mixture of normal distributions. *Biometrika*, 1970, 57, pp. 215–217.

BENGALIA, T., CHAUVEAU, D., HUNTER, D. R., YOUNG, D. S. Mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, 2009, 32(6), pp. 1–29.

DE VEAUX, R. D. Mixtures of Linear Regressions. *Computational Statistics & Data Analysis*, 1989, 8(3), pp. 227–245.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. Maximum Likelihood from Incomplete Data Via the EM-Algorithm. *Journal of the Royal Statistical Society*, Series B, 1977, 39, pp. 1–38.

DESARBO, W. S. AND CRON, W. L. A Maximum Likelihood Methodology for Clusterwise Linear Regression. *Journal of Classification*, 1988, 5(2), pp. 249–282.

FARIA, S. AND SOROMENHO, G. Fitting Mixtures of Linear Regressions. *Journal of Statistical Computation and Simulation*, 2010, 80(2), pp. 201–225.

GOLDFELD, S. M. AND QUANDT, R. E. Techniques for Estimating Switching Regressions. In: *Studies in Nonlinear Estimation*, Cambridge, Massachussets, Ballinger, 1976, pp. 3–35.

GRÜN, B. AND LEISCH. F. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, 2008, 28(4), pp. 1–35.

GRÜN, B., SCHARL, T., LEISCH, F. Modelling Time Course Gene Expression Data with Finite Mixtures of Linear Additive Models. *Bioinformatics*, 2012, 28(2), pp. 222–228.

HAMEL, S., YOCCOZ, N. G., GAILLARD, J. M. Assessing Variation in Life-history Tactics within a Population Using Mixture Regression Models: A Practical Guide for Evolutionary Ecologists. *Biological Reviews*, Cambridge Philosophical Society, 2017, 92(2), pp. 754–775.

LEISCH, F. FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *Journal of Statistical Software*, 2004, 11(8), pp. 1–18.

MCLACHLAN, G. AND PEEL, D. *Finite Mixture Models.* New York: John Wiley & Sons, 2000.

NITYASUDDHI, D. AND BÖHNING, D. Asymptotic Properties of the EM Algorithm Estimate for Normal Mixture Models with Component Specific Variances. *Computational Statistics & Data Analysis*, 2003, 41, pp. 591–601.

PEARSON, K. Contributions to the Mathematical Theory of Evolution. *The Royal Society*, 1894, 185, pp. 71–110.

QUANDT, R. E. AND RAMSEY, J. B. Estimating Mixtures of Normal Distributions and Switching Regressions. *Journal of the American Statistical Association*, 1978, 73, pp. 730–752.

R CORE TEAM. R: A *Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing, 2016.

REDNER, R. A. AND WALKER, H. F. Mixture Densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, 1984, 26, pp. 195–239.

VAŇKÁTOVÁ, K. AND FIŠEROVÁ, E. Analysis of Income of EU Residents Using Finite Mixtures of Regression Models. In: *34th International Conference Mathematical Methods in Economics MME 2016 – Conference proceedings.* Liberec: Technical University of Liberec, 2016, 1, pp. 875–880.

Comparison of Severity Estimators' Efficiency Based on Different Data Aggregation Levels

Pavel Zimmermann[1] | *University of Economics, Prague, Czech Republic*

Abstract

Estimates of the ultimate claim value occur in many actuarial models. Detailed data about each claim are available for estimation: each claim is at first booked at an initial value and processed over a random number of years, during which it is adjusted until closure. The ultimate value can be estimated based on observations of the ultimate value directly, which in this context, means using aggregated data. A more detailed, distribution-free estimator based on estimates of the initial claim value, the closure probability, and development factors is constructed in this article. It is proved that this estimator is asymptotically unbiased and an approximate analytical formula is derived for its variance. The efficiency of this estimator is compared to the efficiency of the simple arithmetic average of the ultimate claim value. Results are illustrated on an example and complemented with a simulation. The example results in significantly lower variability of the detailed estimator.

Keywords

Severity estimators, efficiency, data aggregation

JEL code

C13, G22

INTRODUCTION

In many actuarial tasks such as reserving or pricing, an estimate of the claim value is necessary. Usually, the focus is on the ultimate claim value that is the value at which the claim is closed. Prior to claim closure, the claim passes through the settlement process. Non-life insurers often collect detailed data about a variety of variables from the settlement process. In (Arjas, 1989), a mathematical description and a list of important variables is presented. Insurers, however, prefer traditional approaches and quite often aggregate their data prior to modelling. Three basic levels of aggregation can be distinguished: 1) Models based on aggregates from multiple claims. For example triangle schemes. 2) Models based on data from individual claims at its 'ultimate' state. 3) Models based on data collected throughout the settlement process, i.e. data containing whole claim 'trajectories' from its registration until its closure. Such detailed data are nowadays commonly available, however, rarely used in full detail. On the one hand, aggregation is usually connected with loss of information that can be used for efficient estimates. On the other hand,

[1] Faculty of Informatics and Statistics, W. Churchill Square 4, 130 67 Prague 3, Czech Republic. E-mail: zimmerp@vse.cz.

if models are based on more granular data, more parameters are usually involved and, hence, higher estimation error may appear. In this article we derive and compare properties of two ultimate claim value (claim severity) estimators based on level 2 and level 3 aggregation of the above mentioned typology. The term 'ultimate claim value' is preferred here to the term 'claim severity' to distinguish the value at claim closure from its value during the settlement process.

In general insurance, models based on triangles, i.e. level 1 aggregation, are presently most popular and have been studied by many authors. See, for example, (England and Verral, 2002) for an extensive list. Estimates based on less aggregated data (level 2 or even 3) are studied by far fewer authors. The research is often focused on stochastic processes underlying the claim occurrence and its development. The theoretical background of individual claim level models was originally set in (Norberg, 1993) and extended in (Norberg, 1999). The author considered a full time-continuous model of the settlement process using a non-homogeneous marked Poisson process. Another model based on the marked processes using simulation techniques was published in (Larsen, 2007). A potential bootstrap algorithm to asses the sampling error is also outlined. A simulation model based on individual claims was also developed in (Antonio and Plat, 2014). In (Herbst, 1999), the author applies survival analysis to derive an analytical formula for the estimate of incurred but not yet reported claims. Estimates based on fitting the multivariate skewed normal distribution were developed in (Pigeon, Antonio, Denuit, 2013) and (Pigeon, Antonio, Denuit, 2014). The topic of individual claim modeling was, from a practical point of view, also analyzed in several consultancy articles such as (Taylor, McGuire, Sullivan, 2008) or (Murphy and McLennan, 2006) in the context of large claims. A similar model was also assumed in (Drieskens et al., 2012). Simulation studies such as (Pigeon, Antonio, Denuit, 2014) or (Antonio and Plat, 2014) proved, on real examples, that higher efficiency of prediction of liabilities can be achieved using an individual claims approach.

Estimators of the ultimate claim value based on level 2 aggregated data appear in many actuarial models. They appear in a variety of simple frequency severity models, in collective risk models, and in more complex schemes such as (Herbst, 1999) or (Huang et al., 2015). Many of the individual claim models mentioned above, such as (Pigeon, Antonio, Denuit, 2013) or (Pigeon, Antonio, Denuit, 2014), are based on level 3 detailed data. To the author's knowledge, a comparison of efficiency of severity estimators based on these two levels of data aggregation has not been tackled previously. We assume that claims follow similar process specification as in (Murphy and McLennan, 2006) and (Drieskens et al., 2012). Each claim consists of a random initial registered value which is further adjusted by random number of random development factors that are independent but not identically distributed. See Formula (2). The first estimator considered is the simple arithmetic average of the ultimate claim value of all observed claims. This means it is calculated based only on data aggregated at level 2 of the above mentioned typology. This estimator does not consider the knowledge of the data from the whole settlement process, just the ultimate values. The second estimator (referred to as detailed) is based on the more granular level 3 data. It is constructed as the estimate of the initial value of the claim at reporting multiplied by a weighted average of estimates of development factors, from initial to a particular development year, where the weights are estimated probabilities of the claim being settled in a particular development year. See Formula (41). The estimate is constructed as an empirical counterpart of the variable defined in Formula (2). There are no specific requirements on the distribution so the estimator can be considered distribution-free.

The detailed estimator requires much more variables to be estimated (probabilities of claim settlement in each development year, development factors for each development year and the initial claim value). On the other hand, it uses more data than the simple average. The main task is to quantify to which extent such granularity contributes to the efficiency of the estimate of the ultimate claim value. The questions answered in this article are: Is it worth to construct more detailed estimate? What is the gain in efficiency?

An approximate formula of the variance of the detailed estimator is derived and compared to the variance of the simple average. Although it was not proved that the variance of the second estimate

is always lower than in the case of the simple average, the presented realistic application suggests that the simple average is much less efficient under practical conditions. If the true process follows our assumptions, the detailed estimator shows, in the example case, approximately 55% lower variance.

The article is structured as follows: In the next section, the components of the ultimate claim value are introduced and the moments of the variables are derived. In Section 3, estimators of these components are constructed and their properties are derived. The main results are in Section 4 where the estimator of the ultimate claim value is constructed, its expected value and approximate formula for its variance are derived. It is further compared to the variance of simple arithmetic average. The formulas derived are applied on a realistic example in Section 5.

1 ULTIMATE CLAIM VALUE

We first define the ultimate claim value and some associated variables. Some relations and properties of these variables are stated. At the end, the first two moments of the ultimate claim value are derived.

1.1 Basic Notation and Assumptions

The following notation is used:

1. Maximum development year is denoted ω.
2. The initial value is denoted X_0. Its expected value and variance are denoted $\mathrm{E}(X_0)=\mu_0$ and $\mathrm{Var}(X_0) = \sigma_0^2$.
3. Vector $\mathbf{I} = (I_1,I_2,\ldots I_\omega)'$ is a vector indicating in which development year was the claim closed. For a claim closed in k-th development year $I_j = 1$ for $j = k$ and $I_j = 0$ otherwise. For simplicity no re-openings are assumed and therefore $\sum_{j=1}^{\omega} I_j = 1$. Expected value of I_j is denoted $\mathrm{E}(I_j)=p_j$.
4. In every development year, the claim value is updated by a random development factor. Vector of these incremental development factors is denoted $\mathbf{D} = (D_1,D_2, \ldots D_\omega)'$.
5. Vectors of cumulative development factors is denoted $\mathbf{F} = (F_1,F_2, \ldots ,F_\omega)'$ where F_j is defined as $F_j = \prod_{u=1}^{j} D_u$. The adjective 'cumulative' will often be omitted. The expected value andvariance of F_j are denoted $\mathrm{E}(F_j) = \mu_j$ and $\mathrm{Var}(F_j) = \sigma_j^2$.
6. Development factor from a period j to a period k is denoted:

$$ {}_jF_k = \prod_{u=j+1}^{k} D_u \,. \tag{1}$$

7. The ultimate claim value (the severity of the claim) is denoted X. It is defined as:

$$ X = X_0 \sum_{j=1}^{\omega} I_j \, F_j \,. \tag{2}$$

Variables X_0, $\mathbf{F}$ and $\mathbf{I}$ will be referred to as the components of the ultimate claim value. Further notation for corresponding estimators is presented in Section 3.

The following is assumed:

A 1. Maximum development year ω is known and deterministic.

A 2. Development factors D_j are mutually independent.

A 3. Vector of development factors $\mathbf{D}$ is independent on the vector of indicators $\mathbf{I}$.

A 4. Initial value X_0 is independent on $\mathbf{I}$ and $\mathbf{D}$.

A 5. The moments μ_0,σ_0^2,μ_j, and σ_j^2 are all finite.

All these assumptions are simplification of reality. Assumption 1 means that 'reasonably' high maximum number of development years have to be chosen in order to cover almost all reasonably observable

cases on one hand and to have reasonable number of observations for the latest development years, on the other hand. Assumption 2 is also a simplifying assumption. Similar assumption is often assumed in aggregate models. This assumption allows derivation of analytic formulas for the variance of the estimators. Independence for given portfolio has to be tested prior to application of the estimators.

1.2 Properties of *F* and *I*

Assumption A3 means that technically we assume that claims develop even after the closure. Such development however does not influence the ultimate claim value. Due to Assumption A4 X_0 is also independent on **F**. Given the independence of D_j stated in Assumption A2, mean of j-th development factor is:

$$ {}_jF_k = \prod_{u=j+1}^{k} D_u \,. \tag{3}$$

The expected value of ${}_jF_k$ is then:

$$ \mathrm{E}\left({}_jF_k\right) = \prod_{u=j+1}^{k} \mathrm{E}\left(D_u\right) = \frac{\mu_k}{\mu_j}. \tag{4}$$

Independence of the incremental development factors D_j means that factors F_j and ${}_jF_k$ are also independent. For any $j < k$ the relation $F_k = F_j \cdot {}_jF_k$ holds. We can write for the covariance of F_j and F_k, $j < k$

$$\begin{aligned} \mathrm{Cov}(F_j, F_k) &= \mathrm{Cov}\left(F_j, F_j \cdot {}_jF_k\right) = \mathrm{E}\left(F_j F_j\right)\mathrm{E}\left({}_jF_k\right) - \mathrm{E}\left(F_j\right)\mathrm{E}\left(F_j\right)\mathrm{E}\left({}_jF_k\right) \\ &= \mathrm{E}\left({}_jF_k\right)\mathrm{Var}(F_j) = \frac{\mu_k}{\mu_j}\sigma_j^2. \end{aligned} \tag{5}$$

Vector I always contains only one element equal to one and $\omega - 1$ elements equal to zero. Vector I has multinomial distribution with parameters $v = 1$ and $p = (p_1, p_2, \ldots, p_\omega)$ and therefore $\mathrm{E}(I_j) = p_j$,

$$ \mathrm{Var}(I_j) = p_j(1 - p_j). \tag{6}$$

and

$$ \mathrm{Cov}(I_j, I_k) = -p_j p_k. \tag{7}$$

A possible modification of the assumed model considering growth curves was published in (Pešta and Okhrin, 2014). The development factors are replaced by some parametric growth curves with number of parameters usually lower than ω. Lowering thenumber of parameters would lead to a more precise prediction, but it also requires backtesting of the growth curve's fit.

1.3 Properties of the Ultimate Claim Value

In this subsection, moments of the ultimate claim value are derived based on moments of the components. Firstly, it is necessary to derive the variance of the sum of products $I_j F_j$ which is further denoted ψ.

Lemma 1. Under Assumptions A1, A2, A3, and A5 variance of the sum of products $I_j F_j$ equals:

$$\psi = \mathrm{Var}\left(\sum_{j=1}^{\omega} I_j F_j\right) = \sum_{j=1}^{\omega} p_j \sigma_j^2 + p_j(1-p_j)\mu_j^2 - 2\sum_{j<k} \mu_k p_j p_k \mu_j. \tag{8}$$

Proof. The proof of this lemma is in the Appendix in subsection A2.

Theorem 1. Under Assumptions A1–A5 the expected value of the ultimate claim value is:

$$\mathrm{E}(X) = \mathrm{E}(X_0)\sum_{j=1}^{\omega} \mathrm{E}\,(I_j F_j) = \mu_0 \sum_{j=1}^{\omega} p_j \mu_j, \tag{9}$$

and the variance is:

$$\mathrm{Var}(X) = \sigma_0^2\psi + \sigma_0^2\left(\sum_{j=1}^{\omega} p_j \mu_j\right)^2 + \mu_0^2\psi. \tag{10}$$

Proof. The variance of the variable X may be written using Formula (A6) from the Appendix as:

$$\mathrm{Var}(X) = \mathrm{Var}(X_0)\mathrm{Var}\left(\sum_{j=1}^{\omega} I_j F_j\right) + \mathrm{Var}(X_0)\mathrm{E}^2\left(\sum_{j=1}^{\omega} I_j F_j\right) + \mathrm{Var}\left(\sum_{j=1}^{\omega} I_j F_j\right)\mathrm{E}^2(X_0). \tag{11}$$

After some algebraic operations this formula can be simplified to Formula (10).

2 ESTIMATORS OF THE COMPONENTS

As a first step to derive properties of the detailed estimator of the ultimate claim value, moments and covariances of estimators of the components and its multiples are derived.

2.1 Random Sample and its Notation

All estimators are denoted with a 'hat' sign. Observations are denoted by adding additional index u to the variable. Random sample of a fixed size of n claims is assumed. Random vector of numbers of claims closed in each development year $j = 1,2, \ldots ,\omega$ is denoted $\mathbf{N} = (N_1,N_2, \ldots , N_\omega)^{'}$. Sum of the elements $\sum N_j = n$. It is automatically assumed that conditioning by a random event (e.g., in case of conditional expectation) means conditioning by an indicator of this random event.

Random vector $\boldsymbol{N}$ may be thought of as the sum of the n independent observations of the vector I and therefore has also multinomial distribution, this time with parameters $v = n$ and again $\mathbf{p} = (p_1,p_2, \ldots ,p_\omega)^{'}$. The following relations hold:

$$\mathrm{E}(N_j) = np_j, \tag{12}$$

$$\mathrm{Var}(N_j) = np_j(1-p_j), \tag{13}$$

and

$$\mathrm{Cov}(N_j, N_k) = -np_jp_k = n\mathrm{Cov}(I_j, I_k). \tag{14}$$

Development factors F_j are only observed for claims which are closed in j or later. This means we have a random number of observations denoted $\bar{N}_j$ defined as:

$$\bar{N}_j = N_j + N_{j+1} + \cdots + N_\omega. \tag{15}$$

By definition the following implications hold for any $k > j$:

$$\bar{N}_j = 0 \Rightarrow \bar{N}_k = 0, \tag{16}$$

$$\bar{N}_k > 0 \Rightarrow \bar{N}_j > 0, \tag{17}$$

$$\bar{N}_j = 0 \Rightarrow N_j = 0. \tag{18}$$

In theory we have to consider a case for which all n claims are closed prior to the development year j and hence no observation of F_j is available for j, i.e. $\bar{N}_j = 0$. The probability of such event is denoted $\pi_j^{(0)}$ and equals:

$$\pi_j^{(0)} = \left(\sum_{k=1}^{j-1} p_k\right)^n. \tag{19}$$

In practical cases $\pi_j^{(0)}$ will be very close to 0 and also for all j if the sum in (19) is less then 1,

$$\lim_{n\to\infty} \pi_j^{(0)} = 0. \tag{20}$$

As $\bar{N}_j$ might be thought of as an outcome of n independent trials with the probability of being closed in development year j or later, equal to $\bar{p}_j = \sum_{k=j}^{\omega} p_k$, we may state that $\bar{N}_j$ has binomial distribution. First negative moment of $\bar{N}_j$ truncated at $\bar{N}_j = 0$, denoted as $\bar{n}_j^{-1}$ is defined as:

$$\bar{n}_j^{-1} = \mathrm{E}\left(\bar{N}_j^{-1} | \bar{N}_j > 0\right) = \frac{1}{1-\pi_j^{(0)}} \sum_{u=1}^{n} \frac{1}{u} \binom{n}{u} \bar{p}_j^u \left(1-\bar{p}_j\right)^{n-u}. \tag{21}$$

2.2 Estimator of Probability of Claim Closure

For multinomial distribution, the maximum likelihood estimate of p_j is the average of the observed indicators, i.e.

$$\hat{p}_j = \frac{\sum_{u=1}^{n} I_{j,u}}{n} = \frac{N_j}{n}. \tag{22}$$

Being the simple average, this estimator is unbiased:

$$\mathrm{E}(\hat{p}_j) = \mathrm{E}(I_j) = p_j, \tag{23}$$

and its variance is:

$$\mathrm{Var}(\hat{p}_j) = \frac{1}{n}\mathrm{Var}(I_j) = \frac{1}{n}p_j(1-p_j). \tag{24}$$

Estimates of the elements of vector $\boldsymbol{p}$ are not independent as, if in some development year more claims are closed, in other development years the number of closed claims will tend to decrease. The covariance of the estimates is, using (14),

$$\mathrm{Cov}\left(\hat{p}_j, \hat{p}_k\right) = \frac{1}{n^2}\mathrm{Cov}(N_j, N_k) = -\frac{1}{n}p_j p_k = \frac{1}{n}\mathrm{Cov}(I_j, I_k). \tag{25}$$

Lemma 2. The expected value of $\hat{p}_j$ conditional on $\bar{N}_j > 0$ equals:

$$\mathrm{E}\left(\hat{p}_j \mid \bar{N}_j > 0\right) = \frac{1}{\left(1-\pi_j^{(0)}\right)}p_j, \tag{26}$$

and covariance conditional on $\bar{N}_j > 0$ and $\bar{N}_k > 0$ equals:

$$\mathrm{Cov}\left(\hat{p}_j, \hat{p}_k \mid \bar{N}_j > 0, \bar{N}_k > 0\right) = \frac{\left(1-n\pi_j^{(0)}\right)}{n\left(1-\pi_k^{(0)}\right)}\mathrm{Cov}\left(\hat{p}_j, \hat{p}_k\right). \tag{27}$$

Proof. Proof of this lemma is presented in the Appendix in subsection 5.2.

2.3 Estimator of the Initial Value and Development Factors

The initial value X_0 can be observed for every loss in the sample. Simple average over all individual losses observed, denoted as $\hat{X}_0$, is considered as a predictor of X_0. Analogous approach may, however, be used for more advanced predictors, if necessary. The moments of this predictor are:

$$\mathrm{E}(\hat{X}_0) = \mathrm{E}(X_0) = \mu_0, \tag{28}$$

i.e. the predictor is unbiased, and

$$\mathrm{Var}(\hat{X}_0) = \frac{Var(X_0)}{n} = \frac{\sigma_0^2}{n}. \tag{29}$$

There are two sources of randomness in the estimate of the development ratio F_j:

1. The number of observations $\bar{N}_j$ defined in (15) available for the estimate of F_j which is the number of losses that were closed in j-th development year and later.
2. The actual observations of development factors $F_{j,u},\ u = 1, \ldots, \bar{N}_j$.

The estimator assumed if $\bar{N}_j = 0$ is the average observed ratio, i.e.

$$\hat{\mu}_j = \frac{\sum_{u=1}^{\bar{N}_j} F_{j,u}}{\bar{N}_j}. \tag{30}$$

The number of observations $\bar{N}_j$ can also be 0 which slightly complicates the inference. We assume for the (theoretical) situation when $\bar{N}_j = 0$ that there is an estimate $\hat{\mu}_j$ available from some external source, for which:

$$\mathrm{E}(\hat{\mu}_j|\bar{N}_j = 0) = \alpha_j < \infty, \tag{31}$$

$$\mathrm{Var}(\hat{\mu}_j|\bar{N}_j = 0) = \beta_j < \infty. \tag{32}$$

As mentioned above $\pi_j^{(0)}$ will in practical tasks be very close to 0 and hence consideration of these external estimates is more formal than practical issue.

Lemma 3. Under Assumptions A3 and A5 the expected value of the estimator $\hat{\mu}_j$ equals:

$$\mathrm{E}(\hat{\mu}_j) = \pi_j^{(0)}\alpha_j + \left(1 - \pi_j^{(0)}\right)\mu_j, \tag{33}$$

and the variance equals:

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_j) = {} & \pi_j^{(0)}\left(1 - \pi_j^{(0)}\right)\left(\mu_j - \alpha_j\right)^2 + \pi_j^{(0)}\beta_j \\ & + \left(1 - \pi_j^{(0)}\right)\bar{n}_j^{-1}\sigma_j^2.\end{aligned} \tag{34}$$

Proof. Proof of this lemma is presented in the Appendix in subsection A2.

In the special case, where the external estimate is unbiased, i.e. $\alpha_j = \mu_j$, the estimate $\hat{\mu}_j$ is also unbiased. As the limit of $\pi_j^{(0)}$ is 0, the estimate is asymptotically unbiased (even if $\alpha_j \neq \mu_j$). Further more, in the special case where the external estimate is unbiased, the first term of (34) equals to 0 and the formula reduces to somewhat intuitive form where the variance of the estimator is weighted average of the variance in the case of the external estimate and the variance of the simple average. Due to limit (20), the estimator is consistent as the influence of the variance of the external estimator vanishes as n is increasing.

Lemma 4. Conditional covariance of the estimators $\hat{\mu}_j, \hat{\mu}_k$, $j < k$, conditioning on $\bar{N}_j > 0$, $\bar{N}_k > 0$ equals under Assumptions A2, A3, and A5:

$$\mathrm{Cov}(\hat{\mu}_j, \hat{\mu}_k|\bar{N}_j, \bar{N}_k > 0) = \bar{n}_j^{-1}\mathrm{Cov}(F_j, F_k). \tag{35}$$

Proof. Proof of this lemma is presented in the Appendix in subsection A2.

Lemma 5. The estimators $\hat{p}_j$ and $\hat{\mu}_k$ $j, k = 1, \ldots, \omega$ are under Assumptions A3 and A5 conditioning on $\bar{N}_j > 0$, $\bar{N}_k > 0$ uncorrelated, i.e.

$$\mathrm{Cov}(\hat{p}_j, \hat{\mu}_k|\bar{N}_j, \bar{N}_k > 0) = 0. \tag{36}$$

Proof. Proof of this lemma is presented in the Appendix in subsection A2.

2.4 Properties of Products of the Estimators

In this section properties of the product of the estimators $\hat{p}_j$ and $\hat{\mu}_j$ are stated. First order approximation of the variance of the the product $\hat{p}_j\hat{\mu}_j$ is further denoted ϕ_j. First order approximation of the conditional covariance $\mathrm{Cov}\left(\hat{p}_j\hat{\mu}_j, \hat{p}_k\hat{\mu}_k|\bar{N}_j, \bar{N}_k > 0\right)$ is denoted as $\xi_{j,k}$.

Lemma 6. Under Assumptions A3 and A5 the expected value of the product $\hat{p}_j \hat{\mu}_j$ equals:

$$\mathrm{E}\left(\hat{p}_j \hat{\mu}_j\right) = p_j \mu_j, \tag{37}$$

and the approximate formula for the variance equals:

$$\phi_j = \mathrm{Var}\left(\hat{p}_j \hat{\mu}_j\right) \approx \frac{1}{n}\mathrm{Var}(I_j)\left(\mathrm{Var}(\hat{\mu}_j) + \mathrm{E}^2(\hat{\mu}_j)\right) + \mathrm{Var}(\hat{\mu}_j)\mathrm{E}^2(I_j), \tag{38}$$

where $\mathrm{E}(\hat{\mu}_j)$ is derived in (33) and $\mathrm{Var}(\hat{\mu}_j)$ is derived in (34). The product $\hat{p}_j \hat{\mu}_j$ is consistent.
Proof. Proof of this lemma is in the Appendix in Subsection 5.3.

For covariance of the product, $\mathrm{Cov}(\hat{p}_j \hat{\mu}_j, \hat{p}_k \hat{\mu}_k)$, $j < k$ is assumed. In order to derive the formula, it is necessary to cover different possible constellations of $\overline{N}_j$ and $\overline{N}_k$ being zero or grater than zero. It is necessary to cover situations $\left(\overline{N}_j > 0, \overline{N}_k > 0\right), \left(\overline{N}_j > 0, \overline{N}_k = 0\right)$ and $\left(\overline{N}_j = 0, \overline{N}_k = 0\right)$. The combination $\left(\overline{N}_j = 0, \overline{N}_k > 0\right)$ cannot appear for $j < k$ due to implication (16). Using the Law of total covariance and approximate formula for the covariance of the product of random variables, the following lemma is proved:

Lemma 7. Under Assumptions A2, A3, and A5 the first order approximation of the conditional covariance $\mathrm{Cov}\left(\hat{p}_j \hat{\mu}_j, \hat{p}_k \hat{\mu}_k \middle| \overline{N}_j, \overline{N}_k > 0\right)$ equals:

$$\mathrm{Cov}\left(\hat{p}_j \hat{\mu}_j, \hat{p}_k \hat{\mu}_k \middle| \overline{N}_j, \overline{N}_k > 0\right) \approx \frac{p_j p_k n_j^{-1}\mathrm{Cov}\left(F_j F_k\right)}{\left(1-\pi_j^{(0)}\right)\left(1-\pi_k^{(0)}\right)} + \frac{\mu_j \mu_k \mathrm{Cov}\left(I_j I_k\right)\left(1-n\pi_j^{(0)}\right)}{n\left(1-\pi_k^{(0)}\right)}, \tag{39}$$

and the unconditional covariance equals approximately:

$$\xi_{j,k} = \mathrm{Cov}\left(\hat{p}_j \hat{\mu}_j, \hat{p}_k \hat{\mu}_k\right) \approx p_j p_k \mu_j \mu_k \left(\frac{n_j^{-1}\sigma_j^2}{\left(1-\pi_j^{(0)}\right)\mu_j^2} - \frac{\left(1-n\pi_j^{(0)}\right)}{n} + \left(1-\pi_k^{(0)}\right)\pi_j^{(0)}\right). \tag{40}$$

Proof. Proof of this lemma is in the Appendix in subsection A4.

On one hand, more precise approximations could be achieved by using stochastic expansions shown in (Hudecová and Pešta, 2013). On the other hand, more restrictive assumptions would be required.

3 ESTIMATOR OF ULTIMATE CLAIM VALUE

In this section we define the ultimate claim value estimator based on data collected over the whole claim settlement trajectory and derive its properties. The detailed estimator proposed is constructed as an empirical counterpart of Formula (2):

$$\hat{X} = \hat{X}_0 \sum_{j=1}^{\omega} \hat{p}_j \hat{\mu}_j. \tag{41}$$

3.1 Properties of the estimator

Mean and first order approximation of the variance are derived based on the properties of the estimators derived in Section 1.

Theorem 2. Estimator $\hat{X}$ is under Assumptions A1–A5 unbiased:

$$\mathrm{E}\left(\hat{X}\right) = \mu_0 \sum_{j=1}^{\omega} p_j \, \mu_j, \tag{42}$$

and the approximate variance is:

$$\mathrm{Var}\left(\hat{X}\right) \approx \frac{\sigma_0^2}{n}\phi + \frac{\sigma_0^2}{n}\left(\sum_{j=1}^{\omega} p_j \, \mu_j\right)^2 + \mu_0^2\phi, \tag{43}$$

where ϕ denotes approximate variance $\mathrm{Var}\left(\sum_{j=1}^{\omega} \hat{p}_j \, \hat{\mu}_j\right)$ and equals:

$$\mathrm{Var}\left(\sum_{j=1}^{\omega} \hat{p}_j \, \hat{\mu}_j\right) \approx \phi = \sum_{j=1}^{\omega} \phi_j + 2\sum_{j<k} \xi_{j,k}. \tag{44}$$

Proof. Using the assumption of independence of *Ij*, F_j and X_0 we can write for the mean:

$$\mathrm{E}\left(\hat{X}\right) = \mathrm{E}\left(\widehat{X_0}\right) \sum_{j=1}^{\omega} \mathrm{E}\left(\hat{p}_j \hat{\mu}_j\right) \tag{45}$$

Using (37) and (28) we get Formula (42).

The approximate variance (44) of the sum of $\hat{p}_j \, \hat{\mu}_j$ is calculated using the first order approximations ϕ_j derived in (38) and $\xi_{j,k}$ derived in (40) for the variances or covariances respectively. $\mathrm{Var}\left(\hat{X}\right)$ is then calculated using Formula (A6) from the Appendix and inserting (28) and (29).

3.2 Asymptotic Relative Efficiency

Now, we compare the asymptotic efficiency of the detailed estimator of the ultimate claim value $\hat{X}$ defined in (41) with the simple average of the ultimate values of the claims observed. The simple average estimator is denoted $\bar{X}$. The variance of $\bar{X}$ is simply the variance of the ultimate claim value derived in (10) divided with the number of observed claims *n*. The asymptotic relative efficiency is then:

$$\lim_{n\to\infty} \frac{Var(\hat{X})}{Var(\bar{X})} = \frac{\sigma_0^2\left(\sum_{j=1}^{\omega} p_j \, \mu_j\right)^2 + \mu_0^2 K}{Var(X)}, \tag{46}$$

where:

$$K = \lim_{n\to\infty} n\,\phi = \sum_{j=1}^{\omega} \mu_j^2 \, p_j(1-p_j) + \frac{\sigma_j^2}{\bar{p}_j} p_j^2 + 2\sum_{j<k} p_j \, p_k \mu_j \mu_k \left(\frac{\sigma_j^2}{\bar{p}_j \mu_j^2} - 1\right) \tag{47}$$

and $\mathrm{Var}(X)$ is defined in (10). The first term of the numerator of (46) is identical to the middle term of $\mathrm{Var}(X)$ hence $\hat{X}$ is more efficient than simple average in case the following relation holds:

$$\frac{\sigma_0^2}{\mu_0^2} + 1 > \frac{K}{\psi}. \tag{48}$$

Left hand side contains only characteristics of the initial loss X_0. We may conclude that higher relative efficiency of $\hat{X}$ may be expected in case of high relative variability of X_0. The right hand side is a fraction of complex sums of characteristics of both F_j and I_j for which some straightforward statements cannot be easily claimed, however, the following practical example suggests that the relative efficiency observed may be well below one (see Table 2).

4 PRACTICAL EXAMPLE

The properties of both estimators are illustrated on an example based on real data from motor third party liability bodily claims. Simulation study is presented to accompany the analytical results. All financial values are in EUR. The following 'true' values are assumed:

- Maximum $\omega = 9$ development years.
- Gamma distribution with $\mathrm{E}(X_0) = 6\,704$ and $\mathrm{Var}(X_0) = 125216729$ is assumed for the initial value X_0.
- Gamma distribution is also assumed for all development factors F_j, $j = 1,2, \ldots, 9$. Moments of the variables are contained in Table 1.
- Sample size (number of claims) $n = 5\,000$.
- Although probabilities of having no observation in j-th development year $\pi_j^{(0)}$ are negligible as maximum equals $\pi_9^{(0)} = 2.1\times10^{-10}$, we set formally also moments of the external estimates. Intentionally, the values are selected to be of a very low quality. Both the mean and the variance (α and β) of the external estimate is set twice the true values. Note that in the case of one observation, the variance of the estimate would be equal to the variance of F_j, i.e. would be $\beta/2$.

Table 1 Parameters used as the true values in the simulation. $\mathbf{E}(F_j)$ and $\mathbf{Var}(F_j)$ are calculated based on Formulas (3) and (A6)

j	$\mathrm{E}(I_j) = p_j$	$\mathrm{E}(D_j)$	$\mathrm{Var}(D_j)$	$\mathrm{E}(F_j) = \mu_j$	$\mathrm{Var}(F_j) = \sigma_j$
1	0.236	1.27	1.73	1.27	1.73
2	0.198	1.08	0.98	1.38	5.29
3	0.138	0.97	0.31	1.34	7.27
4	0.216	0.84	0.13	1.13	6.25
5	0.124	0.79	0.12	0.89	4.78
6	0.050	0.82	0.09	0.73	3.70
7	0.019	0.83	0.08	0.60	2.87
8	0.014	0.76	0.08	0.46	1.93
9	0.004	0.84	0.08	0.38	1.52

Source: Own construction

The parameters shown in Table 1 imply the moments of the ultimate claim value X. The 'true' mean calculated using Formula (9) is $\mathrm{E}(X) = 7833$ and the 'true' variance calculated using Formula (10) is $\mathrm{Var}(X) = 975670440$. Based on these values and assumptions stated in Section 2.1, random portfolios of n claims were generated. For each such portfolio, both estimators of the ultimate claim value $\tilde{X}$ and $\hat{X}$

were calculated. The simple average of the ultimate claim value is calculated directly $\tilde{X}$ from the values observed. For the detailed estimator $\hat{X}$ all the estimators involved such as initial claim size, probabilities of claim closure for each development year, and development factors for each development years are calculated.

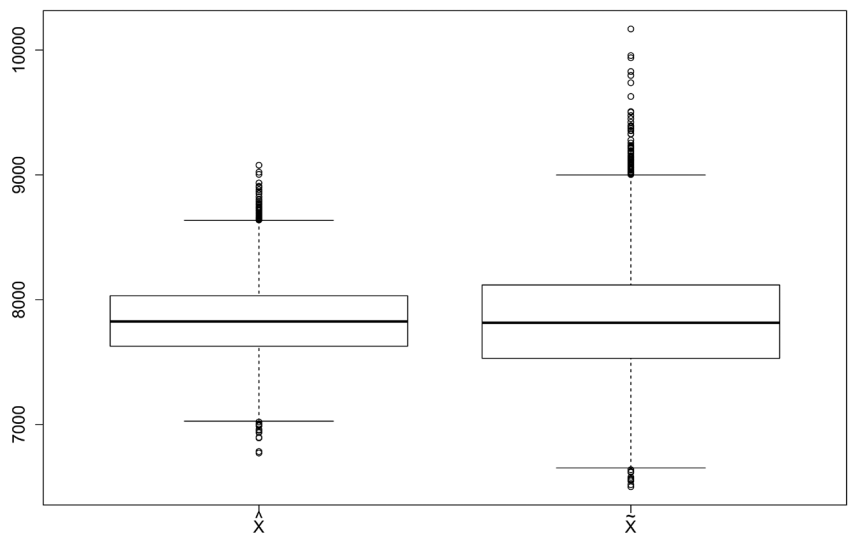
Portfolios were generated randomly 10 000 times and properties of both estimators $\tilde{X}$ and $\hat{X}$ were calculated from the simulations. Both analytic as well as simulated results are for the experiment presented in Table 2. Given the true process follows Assumptions A1–A5, the gain in efficiency using the detailed data to estimate the ultimate claim value is rather high, approximately 55%. The differences in distributions of the estimators are demonstrated in Figure 1. The box plots clearly show smaller variance of the estimator $\hat{X}$. The relative efficiency as a function of the sample size n is plotted in Figure 2.

Table 2 Comparison of the variance of the estimate $\hat{X}$ based on micro data and the simple average $\tilde{X}$

	Analytic result	Simulation
$\psi = \mathrm{Var}(\Sigma\, I_j\, F_j)$	4.729	.
$\phi = \mathrm{Var}(\Sigma\, \hat{p}_j\, \hat{\mu}_j)$	0.001	0.001
$\mathrm{Var}(\hat{X})$ (micro data)	89 293.000	88 638.000
$\mathrm{Var}(\tilde{X})$ (simple avg.)	195 108.000	192 674.000
$\mathrm{Var}(\hat{X})/\mathrm{Var}(\tilde{X})$	**0.458**	0.460
$\lim_{n\to\infty} \mathrm{Var}(\hat{X})/\mathrm{Var}(\tilde{X})$	0.457	.

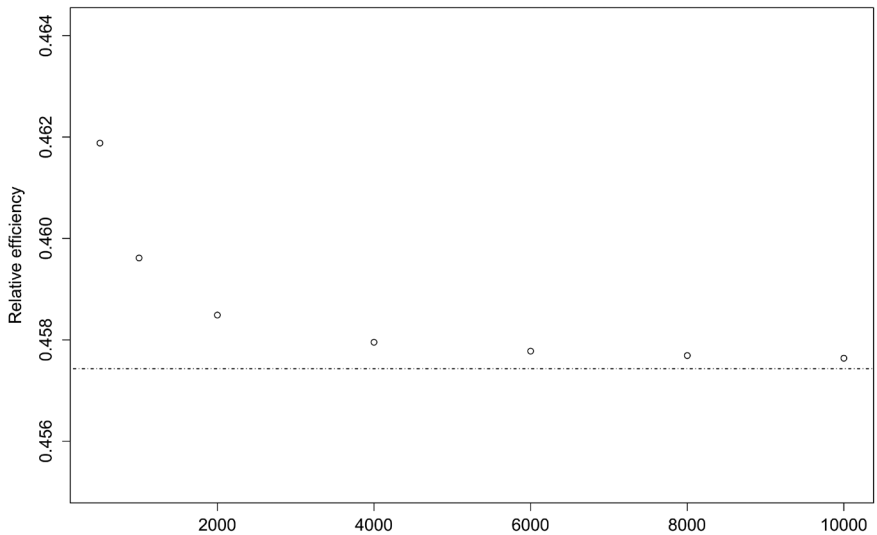
Source: Own construction

Figure 1 Boxplot of the simulated estimates for the simple average $\tilde{X}$ and estimate based on detailed settlement process data $\hat{X}$



Source: Own construction

Figure 2 Relative efficiency as a function of the sample size n and its asymptote



Source: Own construction

CONCLUSION

Models based on aggregated data are rather common in general insurance. Models based on individual data are generally more complex, requiring more calculations and, consequently, more computer time and human capacities. It is not obvious at first sight if this extra effort actually leads to higher efficiency as the number of parameters involved is usually also higher and hence higher estimation error occurs. In this article, we focused on one particular quantity – the ultimate claim value. A simple average, which would usually be the most common estimator employed, was compared to an estimator based on the more detailed data collected during the whole settlement process. It was shown that the estimator proposed is asymptotically unbiased. The approximate formula for its variance was derived and the difference in efficiency from the simple average was evaluated. The estimator is distribution-free. Although it was not proven that the efficiency of the more complex estimator for the assumed process would always be higher, it can be concluded that rather high gains in efficiency may be achieved. In the example presented, the increase of efficiency is almost 55%.

ACKNOWLEDGEMENTS

The author is grateful for comment received by an anonymous referee which helped to improve the quality of the article. The article was written with the support of the Grant Agency of the Czech Republic to the project no. P404/12/0883 and with the institutional support of the long term conceptual development of science and development of the Faculty of Informatics and Statistics of the University of Economics, Prague.

References

ANTONIO, K. AND PLAT R. Micro-Level Stochastic Loss Reserving for General Insurance. *Scandinavian Actuarial Journal*, 2014, 7, pp. 649–69.

ARJAS, E. The Claims Reserving Problem. In: Non-Life Insurance: Some Structural Ideas, *ASTIN Bulletin*, 1989, 19.
BOHRNSTEDT, G. W. AND GOLDBERGER, A. S. On the Exact Covariance of Products of Random Variables. *Journal of the American Statistical Association*, 1969, 64(328), pp. 1439–1442.
DRIESKENS, D., HENRY, M., WALHIN, J. F., WIELANDTS, J. Stochastic Projection for Large Individual Losses. *Scandinavian Actuarial Journal*, 2012(1), pp. 1–39.
ENGLAND, P. D. AND VERRAL, R. J. Stochastic Claims Reserving in General Insurance. *British Actuarial Journal*, 2002, 8(3), pp. 443–518.
HERBST, T. An Application of Randomly Truncated Data Models in Reserving IBNR Claims. *Insurance: Mathematics and Economics*, 1999, 25, pp. 123–131.
HUANG, J., QIU, C., WU, X., ZHOU, X. An Individual Loss Reserving Model with Independent Reporting and Settlement. *Insurance: Mathematics and Economics*, 2015, 64, pp. 232–245.
HUDECOVÁ, Š. AND PEŠTA, M. Modeling Dependencies in Claims Reserving with GEE. *Insurance: Mathematics and Economics*, 2013, 53(3), pp. 786–794.
KENDALL, M. AND STUART, A. *The Advanced Theory of Statistics. Vol. 1: Distribution Theory.* London: Griffin, 1977.
LARSEN, C. R. An Individual Claims Reserving Model. *ASTIN Bulletin International Actuarial Association*, 2007, 37(1), pp. 113–132.
MURPHY, K. AND MCLENNAN, A. A Method For Projecting Individual Large Claims. *Casualty Actuarial Society Forum*, 2006, pp. 205–36.
NORBERG, R. Prediction Of Outstanding Liabilities in Non-Life Insurance. *ASTIN Bulletin International Actuarial Association*, 1993, 23(1), pp. 95–115.
NORBERG, R. Prediction of Outstanding Liabilities II. Model Variations and Extensions. *ASTIN Bulletin International Actuarial Association*, 1999, 29(1), pp. 5–25.
PEŠTA, M. AND OKHRIN O. Conditional Least Squares and Copulae in Claims Reserving for a Single Line of Business. *Insurance: Mathematics and Economics*, 2014, 56, pp. 28–37.
PIGEON, M., ANTONIO, K., DENUIT, M. Individual Loss Reserving with the Multivariate Skew Normal Framework. *ASTIN Bulletin International Actuarial Association*, 2013, 43(3), pp. 399–428.
PIGEON, M., ANTONIO, K., DENUIT, M. Individual Loss Reserving Using Paid–Incurred Data. *Insurance: Mathematics and Economics*, 2014, 58, pp. 121–131.
TAYLOR, G., MCGUIRE, G., SULLIVAN, J. Individual Claim Loss Reserving Conditioned by Case Estimates. *Annals of Actuarial Science*, 2008, 3, pp. 215–56.

APPENDIX

A1 Some Auxiliary Formulas

A1.1 Variance of product of random variables

Lemma 8. The variance of the product of two finite random variables A and B is:

$$Var(AB) = \\ \mathrm{Cov}(A^2, B^2) + (\mathrm{Var}(A) + \mathrm{E}(A))^2(\mathrm{Var}(B) + \mathrm{E}(B))^2 - \\ (\mathrm{Cov}(A, B) - \mathrm{E}(A)\mathrm{E}(B))^2. \tag{A1}$$

Proof. This formula is derived setting:

$$\mathrm{Var}(AB) = \mathrm{E}(A^2B^2) - \mathrm{E}^2(AB) \tag{A2}$$

and inserting:

$$\mathrm{E}(A^2B^2) = \mathrm{Cov}(A^2, B^2) + \mathrm{E}(A^2)\mathrm{E}(B^2), \tag{A3}$$

and

$$\mathrm{E}^2(AB) = (\mathrm{Cov}(A,B) + \mathrm{E}(A)\mathrm{E}(B))^2. \tag{A4}$$

A1.2 Variance of product of independent random variables

The variance of the product of two finite uncorrelated random variables A and B is:

$$\begin{aligned}\mathrm{Var}(AB) = {} & \mathrm{Cov}(A^2, B^2) + \mathrm{Var}(A)\mathrm{Var}(B) + \mathrm{Var}(A)\mathrm{E}^2(B) \\ & + \mathrm{Var}(B)\mathrm{E}^2(A).\end{aligned} \tag{A5}$$

This formula follows directly from inserting $(\mathrm{Cov}(A,B) = 0$ in Formula (A1) in Appendix. If the variables A and B are independent, also $\mathrm{Cov}(A^2, B^2) = 0$, and Formula (A5) reduces to:

$$\mathrm{Var}(AB) = \mathrm{Var}(A)\mathrm{Var}(B) + \mathrm{Var}(A)\mathrm{E}^2(B) + \mathrm{Var}(B)\mathrm{E}^2(A). \tag{A6}$$

A1.3 Approximate covariance of product of random variables

The approximate formula for the covariance of the product of random variables is stated in (Kendall and Stuart, 1977):

$$\begin{aligned}\mathrm{Cov}(AB, UV) \approx {} & \\ & \mathrm{E}(A)\mathrm{E}(U)\mathrm{Cov}(B,V) + \mathrm{E}(A)\mathrm{E}(V)\mathrm{Cov}(B,U) + \\ & \mathrm{E}(B)\mathrm{E}(U)\mathrm{Cov}(A,V) + \mathrm{E}(B)\mathrm{E}(V)\mathrm{Cov}(A,U).\end{aligned} \tag{A7}$$

The exact formula is stated in (Bohrnstedt and Goldberger, 1969).

A2 Moments of the components and estimators

A2.1 Proof of Lemma 1

Variance of the sum can be written as:

$$\mathrm{Var}\left(\sum_{j=1}^{\omega} I_j F_j\right) = \sum_{j=1}^{\omega} \mathrm{V\,ar}(I_j F_j) + 2\sum_{j<k} \mathrm{C\,ov}(I_j F_j, I_k F_k). \tag{A8}$$

The variance of the product $\mathrm{V\,ar}(I_j F_j)$ contained in the first term can be derived using Formula (A6) and inserting (6) for $\mathrm{Var}(I_j)$.

$$\begin{aligned}\mathrm{Var}(I_j F_j) = {} & \mathrm{Var}(I_j)\mathrm{Var}(F_j) + \mathrm{Var}(I_j)\mathrm{E}^2(F_j) + \mathrm{Var}(F_j)\mathrm{E}^2(I_j) \\ = {} & p_j\sigma_j^2 + p_j(1-p_j)\mu_j^2.\end{aligned} \tag{A9}$$

The covariance contained in the second term of (A6) is for $j < k$ using the notation (1):

$$\begin{aligned}cov(I_j F_j, I_k F_k) = {} & \\ & \mathrm{E}(I_j I_k)\mathrm{E}(F_j F_j)\mathrm{E}({}_j F_k) - \mathrm{E}(I_j)\mathrm{E}(F_j)\mathrm{E}(I_k)\mathrm{E}(F_j)\mathrm{E}({}_j F_k) = \\ & \mathrm{E}({}_j F_k)\left[\mathrm{E}(I_j I_k)E(F_j^2) - \mathrm{E}(I_j)\mathrm{E}(I_k)\mathrm{E}^2(F_j)\right].\end{aligned} \tag{A10}$$

As $I_j I_k$ is always equal to 0 due to the fact that if I_j=1, I_k must equal to zero and vice versa, $\mathrm{E}(I_j I_k) = 0$ and we can write using (4):

$$\mathrm{Cov}(I_j F_j, I_k F_k) = -\frac{\mu_k}{\mu_j} p_j p_k \mu_j^2 = -\mu_k p_j p_k \mu_j. \tag{A11}$$

If we now insert (A9) and (A11) into (A8), we get Formula (8).

A2.2 Proof of Lemma 2

Implication (18) implies also:

$$\bar{N}_j = 0 \Rightarrow \hat{p}_j = \frac{N_j}{n} = 0 \tag{A12}$$

and therefore $\mathrm{E}\left(\hat{p}_j | \bar{N}_j = 0\right) = 0$. Using the iterated expectations on $E\left(\hat{p}_j\right)$ gives Formula (26).

For $j < k$ joint probability that $\bar{N}_j > 0 = \bar{N}_k > 0$ equals to probability that $\bar{N}_k > 0$ as the combination $\left(\overline{N}_j = 0, \overline{N}_k > 0\right)$ cannot appear due to implication (16). Using the iterated expectations we can write:

$$\mathrm{E}\left(N_j N_k\right) = \mathrm{E}\left(N_j N_k | \bar{N}_j, \bar{N}_k > 0\right)\left(1 - \pi_k^{(0)}\right). \tag{A13}$$

Inserting this relation and Formula (26) and (14) in the covariance formula, we get:

$$\mathrm{Cov}\left(\hat{p}_j, \hat{p}_j | \bar{N}_j > 0, \bar{N}_k > 0\right) = \frac{Cov\left(N_j, N_k\right) - \pi_j^{(0)} E\left(N_j N_k\right)}{n^2 \left(1 - \pi_j^{(0)}\right)\left(1 - \pi_k^{(0)}\right)}. \tag{A14}$$

Inserting further:

$$\begin{aligned} \mathrm{E}\left(N_j N_k\right) &= \mathrm{Cov}\left(N_j, N_k\right) + \mathrm{E}\left(N_j\right)\mathrm{E}(N_k) \\ &= \mathrm{Cov}\left(N_j, N_k\right) - n^2 \mathrm{Cov}\left(I_j, I_k\right) \end{aligned} \tag{A15}$$

and using (7) yields Formula (25).

A2.3 Proof of Lemma 3

As F_j and I_j are independent we may also state that $\bar{N}_j$ and F_j are in the case $\bar{N}_j > 0$ independent. Furthermore, given the value of $\bar{N}_j$, observations $F_{j,u}$, $u = 1, \ldots, \bar{N}_j$ is series of independent identically distributed variables. Using the iterated expectations the expected value of the estimator equals:

$$\mathrm{E}(\hat{\mu}_j) = \mathrm{E}_{\bar{N}_j > 0} \mathrm{E}(\hat{\mu}_j | \bar{N}_j) = \pi_j^{(0)} \mathrm{E}\left(\hat{\mu}_j | \bar{N}_j = 0\right) + \left(1 - \pi_j^{(0)}\right)\mathrm{E}\left(\hat{\mu}_j | \bar{N}_j > 0\right), \tag{A16}$$

where $\mathrm{E}_{\bar{N}_j > 0}$ denotes expectation over $\bar{N}_j$ conditional on $I(\bar{N}_j > 0)$. Inserting (31) and:

$$\mathrm{E}\left(\hat{\mu}_j | \bar{N}_j > 0\right) = \mu_j \tag{A17}$$

yields Formula (33).

The variance of the estimator $\hat{\mu}_j$ can be derived using the Law of total variance:

$$\mathrm{Var}(\hat{\mu}_j) = \mathrm{Var}_{\bar{N}_j>0}\big(\mathrm{E}(\hat{\mu}_j|\bar{N}_j)\big) + \mathrm{E}_{\bar{N}_j>0}\big(\mathrm{Var}(\hat{\mu}_j|\bar{N}_j)\big). \tag{A18}$$

We can write for first term:

$$\begin{aligned}\mathrm{Var}_{\bar{N}_j>0}\big(\mathrm{E}(\hat{\mu}_j|\bar{N}_j)\big) &= \pi_j^{(0)}\alpha_j^2 + \left(1-\pi_j^{(0)}\right)\mu_j^2 - \left(\pi_j^{(0)}\alpha_j + \left(1-\pi_j^{(0)}\right)\mu_j\right)^2 = \\ &= \pi_j^{(0)}\left(1-\pi_j^{(0)}\right)\left(\alpha_j-\mu_j\right)^2.\end{aligned} \tag{A19}$$

And for the second term:

$$\begin{aligned}&\mathrm{E}_{\bar{N}_j>0}\big(\mathrm{Var}(\hat{\mu}_j|\bar{N}_j)\big) = \\ &\quad = \pi_j^{(0)}\mathrm{Var}\big(\hat{\mu}_j|\bar{N}_j=0\big) + \left(1-\pi_j^{(0)}\right)\mathrm{E}\big(\mathrm{Var}(\hat{\mu}_j|\bar{N}_j>0)\big).\end{aligned} \tag{A20}$$

Inserting (32) and:

$$\begin{aligned}&\mathrm{E}\left(\mathrm{Var}\big(\hat{\mu}_j\big|\bar{N}_j>0\big)\right) = \\ &\quad \mathrm{E}\left(\mathrm{Var}\left(\frac{\sum_{u=1}^{\bar{N}_j}F_{j,u}}{\bar{N}_j} \mid \bar{N}_j>0\right)\right) = \mathrm{E}\left(\frac{1}{\bar{N}_j} \mid \bar{N}_j>0\right)\sigma_j{}^2.\end{aligned} \tag{A21}$$

and using the notation (21) we get Formula (34).

A2.4 Proof of Lemma 4

Let us assume the covariance of the estimators $\mathrm{Cov}(\hat{\mu}_j, \hat{\mu}_k)$ for $j < k$ conditional on some fixed $\bar{N}_j, \bar{N}_k > 0$. The inequality $j < k$ implies $\bar{N}_j \geq \bar{N}_k$ as $\bar{N}_j$ contains all claims contained in $\bar{N}_k$.

$$\begin{aligned}&\mathrm{Cov}(\hat{\mu}_j, \hat{\mu}_k|\bar{N}_j, \bar{N}_k) = \\ &\quad \frac{1}{\bar{N}_j}\frac{1}{\bar{N}_k}\mathrm{E}\left(\sum_{u=1}^{\bar{N}_j}F_{j,u}\sum_{l=1}^{\bar{N}_k}F_{k,l}\right) - \frac{1}{\bar{N}_j}\frac{1}{\bar{N}_k}\mathrm{E}\left(\sum_{u=1}^{\bar{N}_j}F_{j,u}\right)\mathrm{E}\left(\sum_{l=1}^{\bar{N}_k}F_{k,l}\right).\end{aligned} \tag{A22}$$

We may write for the first term:

$$\mathrm{E}\left(\sum_{u=1}^{\bar{N}_j}F_{j,u}\sum_{l=1}^{\bar{N}_k}F_{k,l}\right) = \mathrm{E}\left(\sum_{u=1}^{\bar{N}_j}\sum_{l=1}^{\bar{N}_k}F_{j,u}\,F_{j,l}\;{}_jF_{k,l}\right). \tag{A23}$$

The variables $F_{j,l}$, ${}_jF_{j,l}^k$ are independent. As we assume random sample, we may also state that the variables $F_{j,u}$ and $F_{j,l}$ are independent as long as $u \neq l$. The double sum contains $\min(\bar{N}_j, \bar{N}_k) = \bar{N}_k$ terms for which $u = l$ and $\bar{N}_j\bar{N}_k - \min(\bar{N}_j, \bar{N}_k) = \bar{N}_k$ terms for which $u \neq l$. Therefore we may write:

$$\begin{aligned}
\mathrm{E}\left(\sum_{u=1}^{\bar{N}_j} F_{j,u} \sum_{l=1}^{\bar{N}_k} F_{k,l}\right) &= \sum_{u=1}^{\bar{N}_k} \mathrm{E}\left(F_{j,u}^2\right)\mathrm{E}\left({}_jF_{k,l}\right) + \sum\sum_{u\neq l} \mathrm{E}\left(F_{j,u}\right)\mathrm{E}\left(F_{j,l}\right)\mathrm{E}\left({}_jF_{k,l}\right) = \\
&= \bar{N}_k \mathrm{E}\left(F_j^2\right)\mathrm{E}\left({}_jF_k\right) + \bar{N}_k\bar{N}_j\mathrm{E}\left(F_j\right)^2\mathrm{E}\left({}_jF_k\right) - \bar{N}_k\mathrm{E}\left(F_j\right)^2\mathrm{E}\left({}_jF_k\right) = \\
&= \bar{N}_k \mathrm{E}\left({}_jF_k\right)\mathrm{Var}(F_j) + \bar{N}_k\bar{N}_j\mathrm{E}\left(F_j\right)^2\mathrm{E}\left({}_jF_k\right).
\end{aligned} \tag{A24}$$

The second term of covariance (1) equals:

$$\mathrm{E}\left(\sum_{u=1}^{\bar{N}_j} F_{j,u}\right)\mathrm{E}\left(\sum_{l=1}^{\bar{N}_k} F_{k,l}\right) = \bar{N}_j\bar{N}_k\mathrm{E}\left(F_j\right)^2\mathrm{E}\left({}_jF_k\right). \tag{A25}$$

Therefore inserting (A24) and (A25) into (A22) and using (5), we get:

$$\mathrm{Cov}(\hat{\mu}_j, \hat{\mu}_k|\bar{N}_j, \bar{N}_k) = \frac{1}{\bar{N}_j}\frac{\mu_k}{\mu_j}\sigma_j^2 = \frac{1}{\bar{N}_j}\mathrm{Cov}(F_j, F_k). \tag{A26}$$

Taking the expectation over $\bar{N}_j$ conditional on $\bar{N}_j > 0$ and the notation (21), we get Formula (35).

A2.5 Proof of Lemma 5

For any j, $k = 1, \ldots \omega$, the conditional expected value of the product $\hat{p}_j\,\hat{\mu}_j$ equals to:

$$\begin{aligned}
&E\left(\hat{p}_j\hat{\mu}_k \mid \bar{N}_j > 0, \bar{N}_k > 0\right) = \\
&\mathrm{E}\left(\frac{N_j}{n}\frac{\sum_{u=1}^{\bar{N}_k} F_{k,u}}{\bar{N}_k} \mid \bar{N}_j > 0, \bar{N}_k > 0\right) = \\
&\mathrm{E}_{N_j>0,\bar{N}_k>0}\left(\frac{N_j}{n} \mid \bar{N}_j > 0, \bar{N}_k > 0\right)\mathrm{E}\left(F_k \mid \bar{N}_j > 0, \bar{N}_k > 0\right) = \\
&\mathrm{E}\left(\hat{p}_j|\bar{N}_j > 0, \bar{N}_k > 0\right)\mathrm{E}\left(\hat{\mu}_k|\bar{N}_j > 0, \bar{N}_k > 0\right)
\end{aligned} \tag{A27}$$

which proves (36).

A3 Moments of Product of the Estimators

A3.1 Proof of Lemma 6

The expected value of the product $\hat{p}_j\,\hat{\mu}_j$ can be expressed as:

$$\mathrm{E}\left(\hat{p}_j\hat{\mu}_j\right) = \pi_j^{(0)}\mathrm{E}\left(\hat{p}_j\hat{\mu}_j|\bar{N}_j = 0\right) + \left(1 - \pi_j^{(0)}\right)\mathrm{E}\left(\hat{p}_j\hat{\mu}_j|\bar{N}_j > 0\right). \tag{A28}$$

Implication (18) implies that the first term equals 0.

The expectation in the second term is:

$$\begin{aligned} \mathrm{E}\left(\hat{p}_j\hat{\mu}_j|\bar{N}_j>0\right) &= \mathrm{EE}\left(\frac{N_j}{n}\frac{\sum_{u=1}^{\bar{N}_j}F_{j,u}}{\bar{N}_j} \mid \bar{N}_j>0\right) = \\ &= \mathrm{E}\left(\frac{N_j}{n} \mid \bar{N}_j>0\right)\mathrm{E}\left(F_j\right). \end{aligned} \tag{A29}$$

Using (26) and inserting (A29) into (A28), we get Formula (37).

The estimators $\hat{p}_j$ and $\hat{\mu}_j$ are generally dependent. The variance of the $\hat{p}_j\,\hat{\mu}_j$ can be derived using Formula (A5):

$$\begin{aligned} \mathrm{Var}\left(\hat{p}_j\hat{\mu}_j\right) &= \mathrm{Cov}\left(\hat{p}_j^2,\hat{\mu}_j^2\right) + \mathrm{Var}(\hat{p}_j)\mathrm{Var}(\hat{\mu}_j) + \mathrm{Var}(\hat{p}_j)\mathrm{E}^2(\hat{\mu}_j) \\ &\quad + \mathrm{Var}(\hat{\mu}_j)\mathrm{E}^2(\hat{p}_j). \end{aligned} \tag{A30}$$

The covariance of the squares of the estimators $\mathrm{Cov}\left(\hat{p}_j^2,\hat{\mu}_j^2\right)$ can not be generally derived. Its first order approximation equals to:

$$\mathrm{Cov}\left(\hat{p}_j^2,\hat{\mu}_j^2\right) \approx 4\mathrm{E}(\hat{\mu}_j)\mathrm{E}(\hat{p}_j)\mathrm{Cov}\left(\hat{p}_j,\hat{\mu}_j\right) = 0. \tag{A31}$$

Lemma 5 states that the estimators $\hat{p}_j$ and $\hat{\mu}_j$ are uncorrelated. Formula for the variance of product of independent estimators (A6) is therefore approximately valid for uncorrelated variables.

$$\begin{aligned} \mathrm{Var}\left(\hat{p}_j\hat{\mu}_j\right) &\approx \mathrm{Var}(\hat{p}_j)\mathrm{Var}(\hat{\mu}_j) + \mathrm{Var}(\hat{p}_j)\mathrm{E}^2(\hat{\mu}_j) \\ &\quad + \mathrm{Var}(\hat{\mu}_j)\mathrm{E}^2(\hat{p}_j). \end{aligned} \tag{A32}$$

If we collect $\mathrm{Var}(\hat{p}_j)$ and insert the results (24) and (23) we get the Formula (38). The consistency is implied by the fact that $\hat{\mu}_j$ is a consistent estimator and hence both terms has limit zero.

A4 Covariance of Product of the Estimators

Situations of random sample listed in the first column of Table A1 need to be considered. An approximate formula for the covariance conditioning on the first situation, $\bar{N}_j > 0$ and $\bar{N}_k > 0$, is first derived.

Table A1 Conditional expected values of $\hat{p}_j\,\hat{\mu}_j$ and $\hat{p}_k\,\hat{\mu}_k$ conditioning on different constellations of $\bar{N}_j$ and $\bar{N}_k$

Condition	Probability	$\mathrm{E}(\hat{p}_j\,\hat{\mu}_j)$	$\mathrm{E}(\hat{p}_k\,\hat{\mu}_k)$	$\mathrm{E}(\hat{p}_j\,\hat{\mu}_j\,,\hat{p}_k\,\hat{\mu}_k)$
$\bar{N}_j>0,\ \bar{N}_k>0$	$1-\pi_k^{(0)}$	$p_j\,\mu_j$	$p_k\mu_k$	$p_j\mu_j p_k\mu_k$
$\bar{N}_j>0,\ \bar{N}_j=0$	$\pi_k^{(0)}-\pi_j^{(0)}$	$p_j\,\mu_j$	0	0
$\bar{N}_j=0,\ \bar{N}_k=0$	$\pi_j^{(0)}$	0	0	0

Source: Own construction

A4.1 Proof of Lemma 7

Using Formula (A7) and Lemma 5, we get:

$$\begin{aligned}&\mathrm{Cov}\left(\hat{p}_j\hat{\mu}_j,\hat{p}_k\hat{\mu}_k|\bar{N}_j,\bar{N}_k>0\right)\approx\\&\quad\mathrm{E}\left(\hat{p}_j|\bar{N}_j,\bar{N}_k>0\right)\mathrm{E}\left(\hat{p}_k|\bar{N}_j,\bar{N}_k>0\right)\mathrm{Cov}\left(\hat{\mu}_j,\hat{\mu}_k|\bar{N}_j,\bar{N}_k>0\right)+\\&\quad\mathrm{E}\left(\hat{\mu}_j|\bar{N}_j,\bar{N}_k>0\right)\mathrm{E}\left(\hat{\mu}_k|\bar{N}_j,\bar{N}_k>0\right)\mathrm{Cov}\left(\hat{p}_j,\hat{p}_k|\bar{N}_j,\bar{N}_k>0\right).\end{aligned}\tag{A33}$$

Inserting Formulas (A17), (26), (35) and (25) we get:

$$\begin{aligned}&\mathrm{Cov}\left(\hat{p}_j\hat{\mu}_j,\hat{p}_k\hat{\mu}_k|\bar{N}_j,\bar{N}_k>0\right)\approx\\&\quad\frac{p_jp_kn_j^{-1}Cov\left(F_j,F_k\right)}{\left(1-\pi_j^{(0)}\right)\left(1-\pi_k^{(0)}\right)}+\frac{\mu_j\mu_kCov\left(I_j,I_k\right)\left(1-n\pi_j^{(0)}\right)}{n\left(1-\pi_k^{(0)}\right)},\end{aligned}\tag{A34}$$

which is Formula (39).

Second column of Table 3 contains probabilities of the three situations. The Law of total covariance is applied:

$$\begin{aligned}&\mathrm{Cov}\left(\hat{p}_j\hat{\mu}_j,\hat{p}_k\hat{\mu}_k\right)\\&\quad=\mathrm{E}\left(\mathrm{Cov}(\hat{p}_j\hat{\mu}_j,\hat{p}_k\hat{\mu}_k|\bar{N}_j,\bar{N}_k)\right)\\&\quad+\mathrm{Cov}\left(\mathrm{E}\left(\hat{p}_j\hat{\mu}_j|\bar{N}_j,\bar{N}_k\right),\mathrm{E}\left(\hat{p}_k\hat{\mu}_k|\bar{N}_j,\bar{N}_k\right)\right).\end{aligned}\tag{A35}$$

Due to implication (A12), multiples containing $\hat{p}_k$ conditioning on $\bar{N}_k=0$ all equal 0. Therefore conditional covariance in the first term in both cases where $\bar{N}_k=0$ equals 0 and we may write:

$$\begin{aligned}&\mathrm{E}\left(\mathrm{Cov}(\hat{p}_j\hat{\mu}_j,\hat{p}_k\hat{\mu}_k|\bar{N}_j,\bar{N}_k)\right)\\&\quad=\mathrm{Cov}\left(\hat{p}_j\hat{\mu}_j,\hat{p}_k\hat{\mu}_k|\bar{N}_j>0,\bar{N}_k>0\right)\left(1-\pi_k^{(0)}\right).\end{aligned}\tag{A36}$$

Table A1 can also be used to calculate the second term of Formula (14). Based on this table we may write:

$$\begin{aligned}\mathrm{E}_{\bar{N}_j,\bar{N}_k}\mathrm{E}\left(\hat{p}_j\hat{\mu}_j\right)&=\left(1-\pi_k^{(0)}\right)p_j\mu_j+\left(\pi_k^{(0)}-\pi_j^{(0)}\right)p_j\mu_j\\&=p_j\mu_j\left(1-\pi_j^{(0)}\right),\end{aligned}\tag{A37}$$

$$\mathrm{E}_{\bar{N}_j,\bar{N}_k}\mathrm{E}(\hat{p}_k\hat{\mu}_k)=p_k\mu_k\left(1-\pi_k^{(0)}\right),\tag{A38}$$

$$\mathrm{E}_{\bar{N}_j,\bar{N}_k}\mathrm{E}\left(\hat{p}_j\hat{\mu}_j\hat{p}_k\hat{\mu}_k\right)=p_j\mu_jp_k\mu_k\left(1-\pi_k^{(0)}\right).\tag{A39}$$

The covariance $\mathrm{Cov}\left(\mathrm{E}\left(\hat{p}_j\hat{\mu}_j|\bar{N}_j,\bar{N}_k\right),\mathrm{E}\left(\hat{p}_k\hat{\mu}_k|\bar{N}_j,\bar{N}_k\right)\right)$ is then:

$$\begin{aligned}\mathrm{Cov}\left(\mathrm{E}\left(\hat{p}_j\hat{\mu}_j|\tilde{N}_j,\tilde{N}_k\right),\mathrm{E}\left(\hat{p}_k\hat{\mu}_k|\tilde{N}_j,\tilde{N}_k\right)\right) &= \\ &= \mathrm{E}_{\tilde{N}_j,\tilde{N}_k}\mathrm{E}\left(\hat{p}_j\hat{\mu}_j\hat{p}_k\hat{\mu}_k\right) - \mathrm{E}_{\tilde{N}_j,\tilde{N}_k}\mathrm{E}\left(\hat{p}_j\hat{\mu}_j\right)\mathrm{E}_{\tilde{N}_j,\tilde{N}_k}\mathrm{E}\left(\hat{p}_k\hat{\mu}_k\right) \\ &= p_j\mu_j p_k\mu_k\left(1-\pi_k^{(0)}\right)\pi_j^{(0)}.\end{aligned} \tag{A40}$$

Inserting (5) and (7) in (39) we get Formula (40).

International Conference *Applications of Mathematics and Statistics in Economy* (AMSE 2017)[1]

Stanislava Hronová[2] | *University of Economics, Prague, Czech Republic*

From 30th August to 3rd September 2017 the 20th jubilee international conference *Applications of Mathematics and Statistics in Economy* (20th AMSE) took place in the Jizerské Mountains in Szklarska Poreba, Poland, organized by the department of statistics from the Faculty of Management and Computer Science Wroclaw University of Economics. The Conference was attended by over 60 experts from the Czech Republic, Slovakia and Poland, representing the University of Economics, Prague, Matej Bel University, Banska Bystrica, Wroclaw University of Economics, University of Žilina and Technical University of Zvolen.

This year the AMSE Conference celebrated a jubilee. In 1998, when it was held for the first time, the representatives of the department of statistics from the Faculty of Informatics and Statistics of the University of Economics and the applied informatics department from the Economic Faculty of Matej Bel University agreed on intensification of co-operation between the above work-places. Those, in addition to personal professional contacts of the members of the departments, have built up a tradition to alternate the holding of international conferences with the same or similar aim. In 2000, the Polish colleagues from Wroclaw University of Economics whose statistical work places had very good relations with the departments of statistics of the University of Economics, Prague, were also invited to contribute to organisational issues. The Conference has been gradually upgraded both in terms of programme and participants (bigger share of PhD students) and has become an "essential" part of professional contacts of the members of departments of statistics from the above universities and also the venue of regular friendly gatherings. The aim of this annual international conference is to inform the participants about top modern statistical and mathematical methods which might help to find solution to theoretical and practical issues of economics and economy. The AMSE Conference provides an opportunity to present results of scientific activities of European universities' workplaces. Since its 17th year (from 2014) the AMSE Conference proceedings have been included into the Web of Science database.

Celebration of the 20th jubilee of the conference has added to this meeting of Polish, Slovak and Czech statisticians and mathematicians a certain gala flavour. Polish colleagues have prepared not only an abundant professional programme but they have also reminded previous nineteen conferences by means of photos, programmes and other archival documents. This enabled the participants to recall not only the humble beginnings of the conference and always prevailing warm and open friendly atmosphere but also

[1] More at: <*http://www.amse.ue.wroc.pl/index.html*>.

[2] Faculty of Informatics and Statistics, Department of Economic Statistics, W. Churchill Square 4, 130 67 Prague 3, Czech Republic. E-mail: hronova@vse.cz.

to remember those participants who are not with us any more – Miroslav Abrahám, Anna Sedliecka, Ludwik Adamczyk, Ilja Novák, Josef Kozák, Jiří Trešl, Felix Koschin and Jaroslav Jílek. The loyalty diplomas were awarded to the AMSE Conference founders and to all who took part in all previous twenty conferences (Richard Hindls a Stanislava Hronová from the University of Economics, Prague and to Peter Laco and Rudolf Zimka from Matej Bel University, Banska Bystrica).

Papers presented at this year's conference covered eight subjects (Macroeconomics issues, Financial and Insurance Market, Time Series Analysis Methods, Insurance Market, Demographic and Labour Market Issues, Social Economic Issues, Regional Analysis, Education Issues and Social Welfare and History of Statistics). Proceedings were held in two sections. For the agenda of AMSE 2017 see: <*http://www.amse.ue.wroc.pl/program.html*>. At the conference website you can find the collection of abstracts, the information on the AMSE history and reference to previous years of this international conference.

Papers presented at the conference AMSE 2017 will be published in the book of proceedings that will be sent to Thomson Reuters to be considered for inclusion into the Conference Proceedings Citation Index (CPCI). The proceedings of the past three AMSE conferences (i.e. AMSE 2014, AMSE 2015, AMSE 2016) have been successfully indexed and are available at the Web of Science database.

The tradition of alternate conference holding (Slovakia – Poland – Czech Republic) continues and organizing of the 21st AMSE Conference passes to the department of statistics from the University of Economics, Prague. The conference will take place at the end of August and beginning of September, 2018 in historical town Kutná Hora.

For recap, see the data and venues of twenty previous AMSE conferences:

Year	Date	Place	Organizer
1998	3rd – 4th Sept.	Liptovský Trnovec	Matej Bel University, Banská Bystrica
1999	2nd – 3rd Sept.	Liptovský Trnovec	Matej Bel University, Banská Bystrica
2000	31st Aug. – 1st Sept.	Poprad	Matej Bel University, Banská Bystrica
2001	13th – 14th Sept.	Zadov	University of Economics, Prague
2002	4th – 7th Sept.	Kudova Zdroj	Wrocław University of Economics
2003	4th – 5th Sept.	Banská Bystrica	Matej Bel University, Banská Bystrica
2004	3rd – 4th Sept.	České Budějovice	University of Economics, Prague
2005	1st – 2nd Sept.	Wroclaw	Wrocław University of Economics
2006	31st Aug. – 1st Sept.	Trutnov	University of Economics, Prague
2007	29th Aug. – 1st Sept.	Poprad	Matej Bel University, Banská Bystrica
2008	27th – 29th Aug.	Wisla	Wrocław University of Economics
2009	27th – 28th Aug.	Uherské Hradiště	University of Economics, Prague
2010	25th – 29th Aug.	Demänovská Dolina	Matej Bel University, Banská Bystrica
2011	31st Aug. – 3rd Sept.	Lądek Zdrój	Wrocław University of Economics
2012	30th – 31st Aug.	Liberec	University of Economics, Prague
2013	28th Aug. – 1st Sept.	Gerlachov	Matej Bel University, Banská Bystrica
2014	27th – 31st Aug.	Jerzmanowice	Wrocław University of Economics
2015	2nd – 6th Sept.	Jindřichův Hradec	University of Economics, Prague
2016	31st Aug. – 4th Sept.	Banská Štiavnica	Matej Bel University, Banská Bystrica
2017	30th Aug. – 3rd Sept.	Szklarska Poręba	Wrocław University of Economics

Mathematical Methods in Economics (MME 2017) International Conference[1]

Josef Jablonský[2] | *University of Economics, Prague, Czech Republic*

Mathematical Methods in Economics (MME) conferences have a very long history and tradition. They are one of the most important scientific events organized in the Czech Republic in the field of operational research, econometrics, mathematical economics, and related research areas. In 2017, the 35th international conference *Mathematical Methods in Economics* 2017 was organized in the city of Hradec Králové in September 13–15. Except the local organizer, which was the Faculty of Informatics and Management, University of Hradec Králové, main organizers of MME conferences are the *Czech Society for Operations Research* (CSOR) and the *Czech Econometric Society*.

The total number of participants of this year's conference MME 2017 was more than 150 from the Czech Republic, Iran, Italy, Japan, Poland, and Slovakia. The scientific programme started with a plenary session that was introduced by the chair of the Organizing Committee of the conference and the Vice-dean of the Faculty of Informatics and Management, Professor Petra Poulová. Professor Jana Talašová, President of the *CSOR*, pointed out to a long tradition and importance of MME conferences and their role in supporting development of mathematical modelling from both theoretical and practical point of views. After these introductory talks, two regular invited plenary lectures have been delivered. The first one was given by Professor Shinji Mojida (University of Marketing and Distribution Sciences, Kobe, Japan); its title was *Research and Probabilistic Risk Evaluation of Business System Development Projects Based on Requirements Analysis*. The second plenary talk dealt with *Systemic Risk in Finance and Insurance* and the speaker was Professor Tomáš Cipra (Charles University, Prague, Czech Republic). After the plenary session, the programme of the conference was divided into 4 parallel sessions. The total number of presentations was more than 120. All accepted papers are published in the Proceedings of the MME 2017. They are submitted, as in the previous years, for indexing in the Web of Science.

It has been a long tradition that during MME conferences a competition of PhD students for the best paper takes place. This competition is organized and honoured by the CSOR. All papers submitted were peer-reviewed and the papers with positive referee reports were further evaluated by the Programme Committee. 10 best selected papers have been presented at the conference in two special sessions and finally, the evaluation committee decided about the winners. The winner of the competition was Gabrielle Torri (University of Bergamo, Italy) with the paper *Systemic Risk and Community Structure in the European Banking System*.

Organization of the conference was excellent. All sessions including a conference banquet took place in a new campus of the University of Hradec Králové that offers all necessary up-to-date facilities

[1] More at: <*http://fim2.uhk.cz/mme*>.

[2] Faculty of Informatics and Statistics, Department of Econometrics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: jablon@vse.cz.

for organization of this kind of events. An important part of all conferences is a social programme that always offers many opportunities to discuss various problems in an informal environment. The organizers have prepared 4 options for a half-day tour in the city of Hradec Králové or nearby. After the tour, the conference was officially finished by the conference banquet at which winners of the PhD competition were awarded.

An annual meeting of the CSOR decided that the 36th MME conference will be organized in the town of Jindřichův Hradec by the Faculty of Management, University of Economics, Prague, in September 12–14, 2018.

11th Year of the *International Days of Statistics and Economics* (MSED 2017)[1]

Tomáš Löster[2] | *University of Economics, Prague, Czech Republic*

From 14th to 16th September 2017, a worldwide conference of the International Days of Statistics and Economics (MSED) was held at the University of Economics in Prague. The conference belongs to traditional professional events; this year, the eleventh year of this event was held. The University of Economics, Prague (the Department of Statistics and Probability and the Department of Microeconomics) was the main organizer, as usually; co-organizers were: the Faculty of Economics, the Technical University of Košice, and Ton Duc Thang University. The conference incorporated itself in important statistical and economic conferences, which can be proved by the fact that Online Conference Proceedings have been included in the Conference Proceedings Citation Index (CPCI), which has been since 2011 integrated within the Web of Science, Thomson Reuters. This year, 328 participants registered at the conference; they came from various countries, such as Poland (30), Russian Federation (93), Slovakia (18). Other participants were from Viet Nam, Turkey, Lithuania, France, etc. Conference participants consisted as usually of doctoral students and young scientists of various universities abroad. The aim of the conference was to present scientific papers and discuss current issues in the field of statistics, demography, economics, and human resources, including their mutual interconnection. Regarding statistical topics, the interest was traditionally focused on the cluster analysis, computational statistics, and statistical modelling. This year, a significant invited contribution by Mr. Jiří Rusnok (the Governor of the Czech National Bank) was presented. The lecture hall was full and due to the high erudition of Mr. Rusnok the lecture was accompanied by a rich discussion about current economic topics. To conclude, we wish the conference to be successful in the next year as well, because it is important that through this professional event deeper connections between important disciplines such as statistics and economics are established and the professional community realizes that the mutual cooperation is crucial to the entire system.

[1] More at: <*http://msed.vse.cz*>.

[2] Faculty of Informatics and Statistics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: tomas.loster@vse.cz.

Recent Publications and Events

New publications of the Czech Statistical Office

Demographic Yearbook of the Czech Republic 2016. Prague: CZSO, 2017.
Indicators of Social and Economic Development of the Czech Republic 2000–2nd quarter 2017. Prague: CZSO, 2017.
Životní podmínky v ČR 2016 (Living conditions in the CR 2016). Prague: CZSO, 2017.

Other selected publications

EUROSTAT. *Eurostat-OECD compilation guide on inventories.* Luxembourg: Publication Office of the European Union, 2017.
EUROSTAT. *Eurostat regional yearbook.* Luxembourg: Publication Office of the European Union, 2017.
EUROSTAT. *Key figures on Europe.* Luxembourg: Publication Office of the European Union, 2017.
EUROSTAT. *Monitoring social inclusion in Europe.* Luxembourg: Publication Office of the European Union, 2017.
EUROSTAT. *Smarter, greener, more inclusive? Indicators to support the Europe 2020 Strategy.* Luxembourg: Publication Office of the European Union, 2017.

Conferences

The ***20th ROBUST 2018 Conference*** will take place in **Rybník, Hostouň, Czech Republic, from 21st to 26th January 2018**. More information available at: <*https://robust.nipax.cz*>.
The ***European Conference on Quality in Official Statistics Q2018*** will be held in **Kraków, Poland, during 26–29 June 2018.** More information available at: <*http://www.q2018.pl/call-for-abstracts*>.
The ***21st AMSE 2018 Conference*** will take place in **Kutná Hora, Czech Republic, from 28th August to 2nd September 2018**. More information available at: <*http://www.amse-conference.eu*>.

Papers
We publish articles focused at theoretical and applied statistics, mathematical and statistical methods, conception of official (state) statistics, statistical education, applied economics and econometrics, economic, social and environmental analyses, economic indicators, social and environmental issues in terms of statistics or economics, and regional development issues.

The journal of *Statistika* has the following sections:

The *Analyses* section publishes high quality, complex, and advanced analyses based on the official statistics data focused on economic, environmental, and social spheres. Papers shall have up to 12 000 words or up to twenty (20) 1.5-spaced pages.

The *Discussion* section brings the opportunity to openly discuss the current or more general statistical or economic issues; in short, with what the authors would like to contribute to the scientific debate. Discussions shall have up to 6 000 words or up to 10 1.5-spaced pages.

The *Methodology* section gives space for the discussion on potential approaches to the statistical description of social, economic, and environmental phenomena, development of indicators, estimation issues, etc. Papers shall have up to 12 000 words or up to twenty (20) 1.5-spaced pages.

The *Book Review* section brings reviews of recent books in the fieled of the official statistics. Reviews shall have up to 600 words or one (1) 1.5-spaced page.

In the *Information* section we publish informative (descriptive) texts. The maximum range of information is 6 000 words or up to 10 1.5-spaced pages.

Language
The submission language is English only. Authors are expected to refer to a native language speaker in case they are not sure of language quality of their papers.

Recommended Paper Structure
Title (e.g. On Laconic and Informative Titles) — Authors and Contacts — Abstract (max. 160 words) — Keywords (max. 6 words / phrases) — JEL classification code — Introduction — … — Conclusion — Annex — Acknowledgments — References — Tables and Figures

Authors and Contacts
Rudolf Novak*, Institution Name, Street, City, Country
Jonathan Davis, Institution Name, Street, City, Country
* Corresponding author: e-mail: rudolf.novak@domainname.cz, phone: (+420) 111 222 333

Main Text Format
Times 12 (main text), 1.5 spacing between lines. Page numbers in the lower right-hand corner. *Italics* can be used in the text if necessary. *Do not* use **bold** or underline in the text. Paper parts numbering: 1, 1.1, 1.2, etc.

Headings
1 FIRST-LEVEL HEADING (Times New Roman 12, bold)
1.1 Second-level heading (Times New Roman 12, bold)
1.1.1 Third-level heading (Times New Roman 12, bold italic)

Footnotes
Footnotes should be used sparingly. Do not use endnotes. Do not use footnotes for citing references.

References in the Text
Place reference in the text enclosing authors' names and the year of the reference, e.g. "White (2009) points out that…", "… recent literature (Atkinson et Black, 2010a, 2010b, 2011, Chase et al., 2011, pp. 12–14) conclude…". Note the use of alphabetical order. Include page numbers if appropriate.

List of References
Arrange list of references alphabetically. Use the following reference styles: [for a book] HICKS, J. *Value and Capital: An inquiry into some fundamental principles of economic theory.* Oxford: Clarendon Press, 1939. [for chapter in an edited book] DASGUPTA, P. et al. Intergenerational Equity, Social Discount Rates and Global Warming. In PORTNEY, P., WEYANT, J., eds. *Discounting and Intergenerational Equity.* Washington, D.C.: Resources for the Future, 1999. [for a journal] HRONOVÁ, S., HINDLS, R., ČABLA, A. Conjunctural Evolution of the Czech Economy. *Statistika, Economy and Statistics Journal,* 2011, 3 (September), pp. 4–17. [for an online source] CZECH COAL. *Annual Report and Financial Statement 2007* [online]. Prague: Czech Coal, 2008. [cit. 20.9.2008]. <http://www.czechcoal.cz/cs/ur/zprava/ur2007cz.pdf>.

Tables
Provide each table on a separate page. Indicate position of the table by placing in the text "insert Table 1 about here". Number tables in the order of appearance Table 1, Table 2, etc. Each table should be titled (e.g. Table 1 Self-explanatory title). Refer to tables using their numbers (e.g. see Table 1, Table A1 in the Annex). Try to break one large table into several smaller tables, whenever possible. Separate thousands with a space (e.g. 1 528 000) and decimal points with a dot (e.g. 1.0). Specify the data source below the tables.

Figures
Figure is any graphical object other than table. Attach each figure as a separate file. Indicate position of the figure by placing in the text "insert Figure 1 about here". Number figures in the order of appearance Figure 1, Figure 2, etc. Each figure should be titled (e.g. Figure 1 Self-explanatory title). Refer to figures using their numbers (e.g. see Figure 1, Figure A1 in the Annex).

Figures should be accompanied by the *.xls, *.xlsx table with the source data. Please provide cartograms in the vector format. Other graphic objects should be provided in *.tif, *.jpg, *.eps formats. Do not supply low-resolution files optimized for the screen use. Specify the source below the figures.

Formulas
Formulas should be prepared in formula editor in the same text format (Times 12) as the main text.

Paper Submission
Please email your papers in *.doc, *.docx or *.pdf formats to statistika.journal@czso.cz. All papers are subject to double-blind peer review procedure. You will be informed by our managing editor about all necessary details and terms.

Contacts
Journal of Statistika | Czech Statistical Office
Na padesátém 81 | 100 82 Prague 10 | Czech Republic
e-mail: statistika.journal@czso.cz
web: www.czso.cz/statistika_journal

Managing Editor: Jiří Novotný
phone: (+420) 274 054 299
fax: (+420) 274 052 133
e-mail: statistika.journal@czso.cz
web: www.czso.cz/statistika_journal
address: Czech Statistical Office | Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscription price (4 issues yearly)
CZK 372 (incl. postage) for the Czech Republic,
EUR 117 or USD 174 (incl. postage) for other countries.
Printed copies can be bought at the Publications Shop of the Czech Statistical Office (CZK 66 per copy).
address: Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscriptions and orders
MYRIS TRADE, s. r. o.
P. O. BOX 2 | 142 01 Prague 4 | Czech Republic
phone: (+420) 234 035 200,
fax: (+420) 234 035 207
e-mail: myris@myris.cz

Design: Toman Design
Layout: Ondřej Pazdera
Typesetting: Družstvo TISKOGRAF, David Hošek
Print: Czech Statistical Office

97th year of the series of professional statistics and economy journals of the State Statistical Service in the Czech Republic: ***Statistika*** (since 1964), ***Statistika a kontrola*** (1962–1963), ***Statistický obzor*** (1931–1961) and ***Československý statistický věstník*** (1920–1930).

Published by the Czech Statistical Office
ISSN 1804-8765 (Online)
ISSN 0322-788X (Print)
Reg. MK CR E 4684