STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **102** (4) 2022

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

EXECUTIVE BOARD

MANAGING EDITOR

CONTENTS

ANALYSES

369 Boris Marton, Alena Mojsejová
Macroeconomic Indicators and Subjective Well-Being: Evidence from the European Union

382 Viera Pacáková, Ľubica Sipková, Petr Šild
Factors of Differences in the Highest Wages of Employees in the Slovak Republic (2020 vs. 2010)

396 Jana Cibulková, Barbora Kupková
Review of Visualization Methods for Categorical Data in Cluster Analysis

409 Vladimír Mucha, Ivana Faybíková, Ingrid Krčová
Use of Markov Chain Simulation in Long Term Care Insurance

426 Nataliia Versal, Vasyl Erastov, Mariia Balytska, Ihor Honchar
Digitalization Index: Case for Banking System

443 Guan-Yuan Wang
Churn Prediction for High-Value Players in Freemium Mobile Games: Using Random Under-Sampling

454 Mohamed Saidane
A New Viterbi-Based Decoding Strategy for Market Risk Tracking: an Application to the Tunisian Foreign Debt Portfolio During 2010–2012

INFORMATION

471 Ondřej Vozár
ROBUST 2022 (Volyně) 22nd Event of International Statistical Conference

473 Stanislava Hronová
24th International Conference *Applications of Mathematics and Statistics in Economics (AMSE 2022)*

476 Petr Doucek, Lea Nedomová
International Conference *Interdisciplinary Information Management Talks (IDIMT 2022)*

478 Petra Zýková, Josef Jablonský
Mathematical Methods in Economics (MME 2022) International Conference

480 Tomáš Löster, Jakub Danko
16th Year of the *International Days of Statistics and Economics (MSED 2022)*

481 European Statistical System Peer Reviews

483 Contents 2022 (Vol.102)

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. Its mission is to promote the official statistics as a tool supporting the decision making at the level of international organizations, central and local authorities, as well as businesses. We contribute to the world debate and efforts in strengthening the bridge between theory and practice of the official statistics. Statistika is professional double-blind peer reviewed open access journal included in the citation database of peer-reviewed literature **Scopus** (since 2015), 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

Macroeconomic Indicators and Subjective Well-Being: Evidence from the European Union

Boris Marton[1] | *Technical university of Košice, Košice, Slovakia*
Alena Mojsejová[2] | *Technical university of Košice, Košice, Slovakia*

Received 7.4.2022, **Accepted** (reviewed) 8.6.2022, **Published** 16.12.2022

Abstract

This paper examines the role of factors which could have influenced subjective well-being (SWB) in European countries at a national level between 2010 and 2019. Macroeconomic variables in much of the existing literature have looked at GDP, inflation, government size and expenditure and their relationship to SWB. The current analysis included corruption, property rights, poverty, life expectancy, working time and emissions to enrich the existing body of literature. The World Happiness Index (WHI) is used to measure SWB in this study. The correlation analysis in this study shows a high level of correlation between WHI and the Human Development Index (HDI) which suggests the WHI is a suitable proxy for measuring subjective well-being. Next, the fixed and random effects models were estimated since the dataset was longitudinal, and we have also compared panel regression models with OLS regression models. This analysis revealed positive relationships of GDP, income and property rights on WHI, while poverty and unemployment impact WHI negatively, thus we can conclude positive relationship between material aspects of life and subjective well-being. Corruption and working time impact SWB in a negative way while the impact of life expectancy is positive. The regression models with inflation and emissions were not found to be significant in the research. The results were compared with existing studies based on individual as well as aggregated data. Similarities in results prove that it is possible to analyze determinants of SWB from aggregated data on national level. At the end, we formulate proposals for improving quality of life in the analyzed countries.

Keywords	***DOI***	***JEL code***
Panel data models, quality of life, World Happiness Index, macroeconomic factors, correlation analysis	*https://doi.org/10.54694/stat.2022.19*	*C33, E24, I31, E01*

[1] Department of Regional Sciences and Management, Faculty of Economics, Technical University of Košice, Němcovej 32, 040 01 Košice, Slovakia, Slovakia. E-mail: boris.marton@tuke.sk. ORCID: 0000-0001-7511-4887.
[2] Department of Applied Mathematics and Business Informatics, Faculty of Economics, Technical University of Košice, Němcovej 32, 040 01 Košice, Slovakia. E-mail: alena.mojsejova@tuke.sk. ORCID: 0000-0002-0495-0324.

INTRODUCTION

Subjective well-being is a popular research topic with over 14 000 publications touching on it in 2015 (Diener et al., 2017). As stated in a United Nations report: "The ultimate goal of every individual is happiness (happiness is used to measure SWB as explained in later sections), so then, it must be responsibility of the state, or the government, to create those conditions that will enable citizens to pursue this value, this goal" (Antolini and Simonetti, 2019, p. 264). The government can set policies to help citizens become happy. Diener et al. (2015) recommended creating national accounts of subjective well-being, as an indicator of national progress and growth. However, economic development of a country has set the ultimate goal of many governments. When studying the relationship between happiness and GDP (the macroeconomic indicator for economic growth of countries used since the second world war), the seminar work by Easterlin (1974) is usually the starting point. The main idea of the Easterlin paradox is that after reaching a certain level of income, happiness starts to decrease, contrary to expectations. Since then, many researchers have tried to confirm or refute this paradox. Recent research has confirmed the relationship between happiness, SWB and economic growth (Easterlin, 2015; Veenhoven and Vergunst, 2014). For decades, GDP has been treated as the sole indicator of objective well-being. However, the limitations of GDP were highlighted by the Stiglitz Commission, created by French president Nicolas Sarkozy in 2009. The conclusions of the commission were published (Sen et al., 2010) and concluded that GDP could not be the only measure that reflects people's well-being. It is not sufficient to only study the relationship between GDP and happiness, as there are numerous other variables that can potentially affect happiness. As such, this research is focused on the macro-determinants of subjective well-being at a national level.

The main aim of this paper is to find the macroeconomic variables associated with subjective well-being in the European Union during the period 2010–2019, analyze their relationship to SWB and propose recommendations based on those finding to governments with the aim of enhancing the SWB of EU citizens. The secondary aim of this study is to test whether it achieved comparable results by using aggregated data as other authors have done with individual data.

The paper is organized as follows. Section 1 describes the theoretical background including a review of the literature concerned with topics related to the determinants of subjective well-being. The data and methods used in this study are outlined in section 2 while section 3 discusses the results. The final section concludes the findings and policy implications.

1 THEORETICAL BACKGROUND

One of the most prominent studies on the relationship between income and SWB is the study by Diener et al. (2013) which used over 800 000 individual responses aggregated to a national level in 135 countries in the period 2005–2011. It found a stronger positive relationship between the increase in income and increase in SWB than between GDP and SWB. Yu et al. (2020) also focused on the relationship between income and subjective well-being using individual data in Germany and found a positive relationship. Wealth, as accumulated income, and its effect on SWB was also analyzed by D'Ambrosio et al. (2020) who found a positive relationship between SWB and wealth using OLS regression analysis on a German Socio-Economic Panel dataset. Hochman and Skopek (2013) also found the same (positive) relationship where their OLS regression analysis utilized data from the Survey of Health, Aging and Retirement in Europe aggregated to a national level. Another study by Van der Meer (2014) found a negative relationship between unemployment and SWB, using individual data from the 2004 European Social Survey. This result was later supported by Beja (2020) who used 2004 and 2012 European Social Survey data for the individual part and World Development Indicators dataset for the national level part. The negative impact of unemployment on SWB was found in both datasets. Based on these findings, aggregated macro determinants were used at a national level in this study to analyze the impact of selected variables

on subjective well-being. The study contributes to the existing body of literature by employing panel data analysis on various possible macro determinants of subjective well-being over a relatively long period (10 years). The selection of variables is examined in the following paragraphs.

The relationship between happiness and income is one of the most often studied questions in the field of quality of life (QoL) research. Easterlin (1974) showed that happiness does not always go along with GDP, and after achieving a certain level of GDP or income, further growth in GDP is not necessarily positively related with happiness. However, at the individual level, a positive relationship has been reported in numerous studies (Grable et al., 2013; Lucas and Schimmack, 2009; Yu et al., 2020). This relationship is referred to as the Easterlin paradox. There have been many articles which have studied, analyzed, explained and adjusted this relationship. Antolini and Simonetti (2019) confirmed the existence of this paradox in Italy. On the other hand, similar research in South Korea has not confirmed this (Slag et al., 2019). According to Lim et al. (2020) and Li (2016), the Easterlin paradox holds true even in some East and South Asian countries, although only in the ones that do not favor social values over income. In addition, it is important to keep in mind that the Easterlin paradox was discovered using data on income and happiness, almost 50 years ago. Since then, the American happiness gap between the rich and the poor has widened by about 40% (Okulicz-Kozaryn and Mazelis, 2017). There have been many attempts to model and improve the theory of the Easterlin paradox (Stelzner, 2021). However, these models usually work only under very specific circumstances, during a limited period and only in some countries, which makes them hard to apply in real life QoL research.

It is important to note that there is more to the relationship of income to happiness than just the Easterlin paradox. D'Ambrosio et al. (2020) found that both permanent income and wealth (accumulated income) were better predictors of life satisfaction than the current ones. Moreover, these predictors matter not only in absolute terms but also in comparative terms, so it is important who people compare themselves to. In Cape Town, it was found that income comparisons, both relative to neighbors and relative to oneself, affect subjective happiness differently, depending on age (Tibesigwa et al., 2016). Unfortunately, it was not possible to find a reliable source of data for these indicators, so it was not possible to analyze these relationships. Wealth and income distribution has no effect on QoL based on research carried out in 28 European countries (Zagorski et al., 2014). However, employment plays a role in the income and happiness relationship (Brzezinski, 2019). Based on this finding, Gross Domestic Product (GDP) per capita was chosen as the measurement of economic growth to test the Easterlin paradox. The adjusted gross disposable income of households per capita in the purchasing power standard (2010) from the Eurostat database was used as the measurement of income. Income distribution is not included since the at risk of poverty rate is used as the measurement of poverty. This will be explained further later in this chapter.

Unemployment goes hand in hand with income. Unemployment affects happiness negatively (Glatz and Eder, 2020; Pierewan and Tampubolon, 2015) although this relationship is not so simple. Luo (2020) suggests that there is at least one more important variable in the relationship between unemployment and happiness; material deprivation. Individuals who suffer from material deprivation during the period of unemployment feel unhappy. On the other hand, there are unemployed individuals that do not suffer from material deprivation. Those individuals can feel happy and even happier than during employment. Another important factor for unemployment and happiness is the quality of social institutions and their unemployment policies (Jakubow, 2016). In the current research, the focus is on unemployment, measured as unemployment rate, since it is not the aim to go too deeply into the relationship between unemployment and QoL.

There has not been much research done on the relationship between happiness and inflation. Raising the price of goods and services without raising income can cause a decrease in consumption, savings or both. Chen et al. (2014) have done research supporting this theory. The Harmonized index of customer prices is used in the current research because it uses the same customer basket for all our selected countries.

Papavlassopulos and Keppler (2011) have stated that: "We argue that life is valued for its quality, and, if positive, its extension is an improvement of well-being." In their study, life expectancy had a strong positive relationship to happiness and this relationship was as significant as the relationship of absolute income to individual happiness. Similar results have been presented by Jones and Klenow (2016) who found life expectancy had a positive relationship with subjective well-being. However, its importance varies in each country. Kageyama (2009) found gender was another important factor in the relationship between happiness and life expectancy. It is a fact that women live longer than men on average and this also affects happiness. Men are more stressed than women in general and this lowers their life expectancy. When men die, women become widows which can also lower their happiness. This relationship has not been analyzed in detail however since such findings are not the goal of this paper. However, this study is important as it shows the significant relationship between happiness and life expectancy. Veenhoven (2006) took the next step and combined happiness and life expectancy into one term Happy-Life-Year. The life expectancy at birth indicator from the Eurostat database was used in the current research. A significant positive relationship between life expectancy and happiness is expected.

Many researchers have dealt with the connection between corruption and well-being or happiness. In Lambsdorff (2007), corruption is defined as the misuse of public power for private gains. Corruption can have many consequences such as scaling down the trust among citizens. There is also evidence that trust increases happiness (Growiec and Growiec, 2013; Hommerich and Tiefenbach, 2018). Li (2016) used the happiness index from the WHR and corruption perception index (CPI) to find a negative relationship between a high level of corruption and decrease in the level of happiness. Amini and Douarin (2020) have presented the relationship between corruption and life satisfaction in Central and Eastern Europe (CEE). There is always a happiness gap between the former soviet countries and western European countries and they have concluded that corruption is the reason for this imbalance. Rodríguez-Pose and Maslauskaite (2012) have stated that: "different levels of individual happiness in CEE are therefore mostly determined by institutional factors such as corruption, government spending and decentralization." The transition countries are the focus of interest in a study by Bartolini et al. (2017). Their research is mostly concerned with social trust and the conclusion that social trust is a powerful predictor of the trends of SWB. To our knowledge, there has been no research in which all the EU countries are considered.

There have been various studies related to the work-life balance. It has been found that increasing working hours and overtime have a positive effect on life and job satisfaction, while the desire to reduce working hours has a negative impact on the satisfaction (Holly and Mohnen, 2012). Wirtz and Nachreiner (2010) have referred to the negative effects of long working hours on the subjective work-life balance.

Property rights or homeownership are also connected to happiness. There are relatively small number of studies, but all of them refer to the positive effect on happiness for homeowners (Cheng et al., 2016; Spruk and Kešeljević, 2016).

The importance of air quality as a determinant of life satisfaction is discussed in Ferreira et al. (2012). The impact of climate and air pollution conditions on happiness in the Spanish regions using individual-level data from the European Social Survey has been shown to be significant (Cunado and Perez De Gracia, 2012). In this study, the Eurostat indicator Greenhouse gas emissions per capita is used (Apergis, 2018) and it is assumed that it will have a negative effect on happiness.

2 RESEARCH METHODOLOGY

This study is based on the macroeconomic and social indicators published by Eurostat, the United Nations Development program, the Heritage Foundation and Sustainable Development Solutions Network. The subsequent selection of indicators is based on the previously mentioned studies. The research was focused on the member states of the European Union during the period 2010–2019. The United Kingdom and Malta were excluded from the dataset due to missing values. All the used variables, a brief description of them and the descriptive statistics are reported in Table 2. The variables GDP and DPD were transformed, using logarithmic transformation for further analysis.

First, the relationship between the WHI and HDI was examined using a correlation analysis to find out if the WHI is a suitable proxy for measuring subjective well-being. Spearman's correlation coefficients were used since the data did not have a normal distribution. This can be found in Table 1. The results showed a high correlation (above 0.8) for all the selected years and were statistically significant at 0.001 significance level. Yin et al. (2021) used data from over 150 countries in the period 2005–2018 to study the relationship between SWB and HDI. The authors used multiple OLS regression models and found out that HDI is a suitable indicator for measuring subjective well-being, more cognitive than affective one. This relationship was more prominent in rich western countries. Since the current correlation analysis showed a high positive correlation relationship between the WHI and HDI and the dataset consists of rich member states from the EU, it can be concluded that the WHI is a suitable indicator for measuring SWB.

Next, a regression analysis was carried out. The longitudinal character of the data was suitable to apply econometric methods for panel data. A OLS regression was done although panel regression models fitted the data better according to the Lagrange Multiplier Test (Breusch-Pagan) for unbalanced panels and F test for individual effects. The panel regression analysis was performed according to previous works (Kennedy, 2008; Park, 2011; Torres-Reyna, 2010). Several regression models for panel data were constructed and further analyzed, both with fixed and random effects. However, as suggested by Figure 1, the variables were strongly correlated which might affect the interpretations of the estimated coefficients. As a result, every variable was analyzed separately, by performing partial panel regression models as shown in Table 3. The preferred model selection was based on Hausman test results. The study checked for cross-sectional dependence (contemporaneous correlation) using a Pesaran CD test for cross-sectional dependence in panels, for serial correlation using a Breusch-Godfrey (Wooldridge) test for serial correlation in panel models and a Breusch-Pagan test for heteroskedasticity. The tests indicated the presence of the mentioned problems in the partial panel regression models, so a Panel-Corrected Standard Errors method was used to account for these problems (Bailey and Katz, 2011).

To check whether our results from partial panel regression models are robust, we run multiple panel regression model. To avoid problem with multicollinearity, we decided to run model with only one of the mentioned correlated variables in it. We chose GDP, because it is most commonly used variable associated with SWB. To clarify, we provide model specification in equation below, and characteristics of the complex panel regression model are shown in Table 4.

Table 1 Spearman's correlation coefficient between HDI and WHI for the years 2010–2019

	HDI10	HDI11	HDI12	HDI13	HDI14	HDI15	HDI16	HDI17	HDI18	HDI19
WHI10	0.873	0.884	0.895	0.880	0.875	0.876	0.870	0.875	0.880	0.866
WHI11	0.897	0.900	0.906	0.897	0.891	0.887	0.872	0.870	0.877	0.863
WHI12	0.865	0.868	0.876	0.885	0.879	0.878	0.866	0.867	0.874	0.861
WHI13	0.874	0.881	0.886	0.881	0.873	0.863	0.850	0.848	0.853	0.840
WHI14	0.834	0.842	0.855	0.852	0.848	0.847	0.834	0.839	0.843	0.831
WHI15	0.851	0.859	0.869	0.861	0.854	0.850	0.835	0.840	0.842	0.831
WHI16	0.820	0.824	0.834	0.831	0.821	0.825	0.814	0.816	0.822	0.805
WHI17	0.842	0.840	0.849	0.845	0.837	0.826	0.809	0.811	0.814	0.803
WHI18	0.877	0.877	0.879	0.877	0.872	0.859	0.845	0.846	0.852	0.838
WHI19	0.897	0.899	0.905	0.901	0.893	0.886	0.880	0.876	0.882	0.868

Source: Own research

$$WHI = \alpha + \beta_1 X_1 + \beta_2 ARP + \beta_3 UNE + \beta_4 WOW + \beta_5 GHE + \beta_6 HICP + \varepsilon . \quad (1)$$

Notes: α – intercept, β_i – coefficient, X_1 – one of five correlated variables, ε – error term.

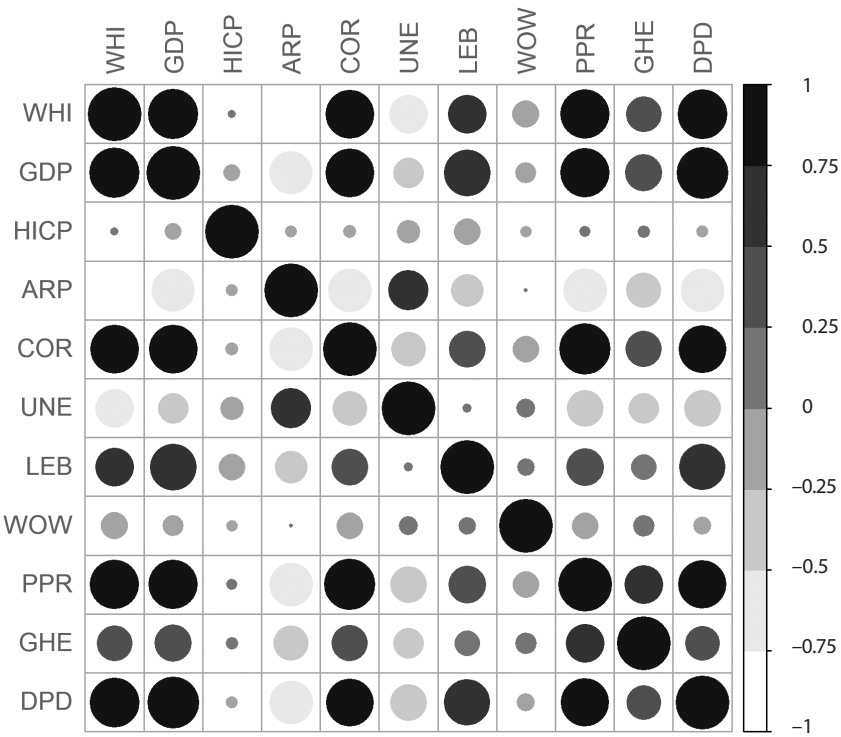
The correlation analysis, panel regression analysis and other tests were performed in R version 4.1.1 and R Studio version 1.4.1717, The descriptive statistics were calculated using Excel 2019.

Table 2 Description and descriptive statistics of variables

Variable	Description	Source	Units of measurement	Min	Max	Average	Median	STDEV	N
HDI	Human development index	United Nations development program	Scale 0–1	0.788	0.955	0.883	0.884	0.040	260
WHI	World happiness index	Sustainable development solutions network	Scale 0–10	3.875	7.858	6.305	6.242	0.860	257
GDP	Real gross domestic product per capita	Eurostat	Chain linked volumes (2010) €	5 080.000	85 030.000	25 632.192	20 325.000	16 991.136	260
HICP	Harmonized index of customer prices	Eurostat	Annual average rate of change	–1.600	6.100	1.493	1.400	1.424	260
ARP	At risk of poverty rate	Eurostat	Total percentage	12.200	49.300	24.130	21.800	7.487	260
COR	Corruption perception index	Eurostat	Scale 0–100	33.000	94.000	63.569	61.000	15.848	260
UNE	Unemployment rate	Eurostat	Percentage of labor force population	2.000	27.500	9.404	8.000	4.859	260
LEB	Life expectancy at birth	Eurostat	In years	73.100	84.000	79.587	80.800	2.885	260
WOW	Average hours worked per week per employed person in a full-time job	Eurostat	In hours	38.400	44.600	41.143	40.900	1.069	260
PPR	Property rights	The Heritage Foundation	Scale 0–100	30.000	95.000	70.916	71.600	17.101	260
GHE	Greenhouse gas emissions per capita	Eurostat	In tons of CO2	5.200	26.600	9.666	8.900	3.646	260
DPD	Adjusted gross disposable income of households per capita	Eurostat	In PPS (2010)	7 880.000	35 012.000	19 202.380	18 518.000	5 651.175	258

Notes: Min – minimum value, Max – maximum value, Average – arithmetic average, STD – standard deviation, N – total number of observations.
Source: Own research

Figure 1 Correlation analysis of variables



Source: Own research

3 RESEARCH RESULTS

In this section the results from the panel regression analysis are discussed. There are only two out of the ten panel regression models which are not statistically significant in any of the considered significance levels. There are three models with a significant fixed effect and five models with a significant random effect (Table 3). The logarithmically transformed adjusted gross disposable income of households per capita explains almost 9 percent more variability of the dependent variable (in this case WHI used for measuring quality of life) than the logarithmically transformed gross domestic product per capita. This result is in line with Sen et al. (2010) which suggests that GDP is not always the most preferred measure for quality of life research. Both estimated coefficients for the income variables are significant and positive, so it can be concluded that there is a positive relationship between the material aspects of life and subjective well-being. Similar facts have also been previously found (Diener and Biswas-Diener, 2002; World Happiness Report, 2020). The coefficients for the logarithmic values of GDP and DPD are positive, which means that these results are not in accordance with the Easterlin Paradox. The current results suggest that an increase in income will cause an increase in happiness.

Poverty is one of the variables related to quality of life, supporting the previously mentioned relationship between the material aspect of life and quality of life. In the current model, the coefficient associated with the poverty variable is negative. This is consistent with other findings (Mood and Jonsson, 2016) where panel data methods were applied on individual longitudinal data from the Swedish Level-of-Living Survey suggesting poverty in general has a negative effect on social life.

Unemployment is closely related to subjective well-being. As expected, a negative relationship was discovered between unemployment and well-being. This is in line with other research (Beja, 2020) which showed that unemployment affects SWB in negative way and that the indirect cost of unemployment is about twice the size of the direct cost of unemployment. Another study (Van der Meer, 2014) also indicated the negative effect of being unemployed.

Inflation, the rising price level of goods and services in the economy, can influence subjective well-being in a negative way if not accompanied by a proportional rise in income. In this case, the real income of individuals decreases which can negatively affect SWB. In the current research, the relationship between inflation and subjective quality of life measured by the WHI is statistically insignificant and therefore no connection can be identified between the variables. This does not mean that there is no relationship between inflation and subjective well-being however. Rather, it means that this relationship was insignificant during the analyzed period in the selected countries. It should be noted that the dataset consisted of rich and financially stable western countries with relatively low levels of inflation. A similar study in China showed the negative effect of inflation on subjective happiness (Chen et al., 2014).

A first sight the results suggests that corruption raises subjective well-being, although when the Corruption perception index calculation is analyzed (0 representing a very high level of corruption and a score of 100 representing a very "clean" country) it can be seen that a rising level of corruption decreases subjective well-being. This is in line with the results from other studies (Amini and Douarin, 2020; Rodríguez-Pose and Maslauskaite, 2012). According to these studies, corruption negatively affects people's lives in two ways: Firstly, corruption mostly takes more money away from people with the fewest contacts in "high places" and acts as a cost of living in corrupt countries. Secondly, there is the psychological impact of paying a bribe for an otherwise free service.

Every individual wants to live a long and happy life. Life expectancy at birth predicts how long a newborn individual will live, so a positive relationship between life expectancy and the WHI was expected. Papavlassopulos and Keppler (2011) have found a strong positive relationship between life expectancy and happiness as have Cervellati and Sunde (2011). They found that rising life expectancy in highly dynamic economies had the potential to stimulate the economy from stagnation to growth by effective work allocation, which can also increase quality of life. The current research also showed the coefficient related to life expectancy at birth to be positive and statistically significant, supporting previous findings.

Everyone has limited time in their life. As a result, individuals need to distribute their time between work and other activities such as socializing with friends and family. In their study on 50 000 people in Italy, Mingo and Montecolle (2014) found leisure activities to be positively interrelated to happiness and well-being. Wirtz and Nachreiner (2010) also found a negative relationship between extended working time and subjective well-being. The current results are in line with these studies. The coefficient related to average working time is negative and significant and it can thus be concluded that there is a negative effect of increasing working time, which in turn lowers free time for individuals, on SWB, so people in general do not like to spend more time in work.

The property rights index is a sub-indicator of the Index of Economic Freedom (IEF) measured by the Heritage Foundation. It measures the extent to which an individual can accumulate capital freely, without restrictions from the government. Capital accumulation is a function of income which is related to SWB as mentioned at the beginning of this chapter. This study found that the possibility of accumulating capital freely had a positive relationship with subjective well-being. This supports previous findings which have shown the positive effect of economic freedom and property rights on subjective quality of life, although this relationship is only positive in the short-term (Spruk and Kešeljević, 2016). Another study examining this interrelationship in China (Cheng et al., 2016) found that property rights had an impact on SWB.

While the material aspects of life do matter, these macroeconomic indicators do not give the whole picture. Wu (1999) revealed a strong negative impact of economic growth side effects like air and water

pollution on subjective well-being in China. This study uses the greenhouse gas emissions per capita indicator to at least touch on the topic of pollution and environment. This was inspired by Apergis (2018) who found a negative association between greenhouse gas emissions and personal well-being, both at an aggregate-country level and regional country level. He used a panel data methodological approach on 58 countries, including but not limited to the countries selected in the current research between 2005 and 2014. Surprisingly, the current results are in contrast to Apergis (2018). The current model is not significant in any of the considered significance levels, and it cannot be concluded that there is a relationship between greenhouse gas emissions per capita and SWB as measured by the WHI.

The results from multiple panel regression model were qualitatively the same as partial panel regression models. Coefficients decreased in absolute term, but this is expected in presence of other significant

Table 3 Characteristics of partial panel regression models

Variable	Preferred model	Coefficient	Significance	Coefficient of determination
Log(DPD)	Random	2.2574 (0.2387)	***	0.4817
ARP	Random	–0.0770 (0.0087)	***	0.4786
UNE	Random	–0.0723 (0.0101)	***	0.4011
Log(GDP)	Fixed	2.2908 (0.3806)	***	0.3934
LEB	Random	0.1934 (0.0426)	***	0.1784
PPR	Fixed	0.0171 (0.0047)	***	0.1493
COR	Fixed	0.0250 (0.0094)	**	0.0838
WOW	Random	–0.2754 (0.1190)	*	0.0653
GHE	Random	0.0271 (0.0396)		0.0045
HICP	Random	0.0004 (0.0184)		0.0001

Notes: Values are rounded mathematically to four decimal places. Standard errors of coefficients are in parentheses. *** denotes $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and · $p < 0.1$.
Source: Own research

Table 4 Characteristics of multiple panel regression model

Variable	Coefficient	Significance
Intercept	2.3376 (2.6715)	
Log(GDP)	0.7316 (0.1121)	***
ARP	–0.0403 (0.0065)	***
UNE	–0.0220 (0.0081)	**
WOW	–0.0498 (0.0530)	
GHE	–0.0095 (0.0119)	
HICP	0.0027 (0.0125)	
WOW	*	0.0653
GHE		0.0045
HICP		0.0001

Notes: Values are rounded mathematically to four decimal places. Standard errors of coefficients are in parentheses. *** denotes $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, and · $p < 0.1$.
Source: Own research

predictors in a complex model. Similarly, to GDP model, we run alternative models with DPD, COR, PPR and LEB. All models were found to be significant on 0.001 significance level. We tested these five models as described in Section 2 of this article. Random and fixed effects were significant, there were no problem with cross-sectional dependence or serial autocorrelation, but GDP and DPD models had problem with heteroskedasticity present. To account for this a robust covariance matrix was used according to (Kennedy, 2008; Park, 2011) studies.

Unemployment and poverty indicators were significant, and their coefficients were negative in all different specification of models, while both income variables and corruption perception index were found to be significant and positively related to the dependent variable. The working time, emissions, inflation and property rights were not significant predictors throughout different specifications of models, so we can conclude, that our results from partial panel regression models are robust.

CONCLUSIONS

This paper shows that macroeconomic indicators are strongly related to subjective well-being by performing a panel data analysis on a dataset gathered from Eurostat, the United Nations Development program, The Heritage Foundation and Sustainable Development Solutions Network. The research was focused on member states of the European Union during the period 2010–2019. The World Happiness Index (WHI) from the World Happiness Report was used as the measure of subjective well-being. A positive relationship was found between income and wealth variables and quality of life. A negative relationship was also found between quality of life, poverty and unemployment. These findings lead us towards setting of a positive relationship between the material aspects of life, such as income or wealth, and subjective well-being. Thus, it can be concluded that this research does not support the Easterlin paradox, although there is more to this relationship than can be seen in the results. Fighting corruption will not only increase freedom but also increase subjective happiness. The amount of time spent at work can influence subjective well-being in negative way, since everyone has limited time in life and spending time at work means less time for family, friends or leisure activities, which can also influence health. Health goes along with environment. In the current research, a positive relationship was detected between life expectancy at birth and subjective happiness while the relationship between emissions and happiness was not significant. It can also be concluded that the relationship between selected variables and SWB does not change when using aggregated data instead of individual data, except emissions. However, this effect needs more attention in a separate study.

According to Azizan and Mahmud (2018), it is important to pay attention to the determinants of quality of life and subjective well-being in order to improve the overall quality of life of citizens. Their review study found income, wealth, employment and health to be the most important predictors of subjective well-being. Ngamaba (2016) has stated: "In order to decide what policies should be pursued in order to improve SWB there is a need to identify what the key drivers of SWB are." He used a cross-sectional multilevel random effects model on nationally representative data for 59 countries over the period 2010–2014. His results indicate that the most significant determinants of subjective well-being are health, household's financial satisfaction and freedom of choice. Diego-Rosell et al. (2018) have also shown that subjective material well-being and its objective determinants, including economic growth and income inequality, should remain at the center of the research and policy agenda. The current research is similar to these studies in that it examines the determinants, mostly macro economic ones, of subjective well-being. In line with those studies, the current results suggest that the most important predictors of subjective well-being are income, wealth, health and freedom in accumulating capital. This study differs from the other ones in variable selection and relatively long period of observation. It did not only include measurements of income like GDP and disposable income of households but also poverty,

unemployment, corruption, working time, life expectancy, inflation, emissions and property rights. To the best of our knowledge, there is no study like this in the existing body of literature.

From a policy perspective, the findings of this study suggest that it is important to focus on quality education in economics since it seems that people in the European Union are materialistically oriented. The negative impact of a materialistic lifestyle on subjective well-being and happiness has been shown by many researchers (Górnik-Durose, 2020; Kasser et al., 2014; Ng and Diener, 2014). In particular, Kasser et al. (2014) found that after a quick educational course people became less materialistic and happier. In other words, a proper education can lead people to a happier life as well as to a healthier one, at least from a psychological perspective.

This study also has its limitations. The first and biggest one is using an aggregated index for measuring subjective well-being and happiness. It is important to keep in mind that the goal was not to analyze the relationship of happiness and other macro-determinant in detail for each country, but rather focus on a longitudinal analysis for the entire European Union at once. The second limitation of this study is analyzing the entire group of European Union countries at once, without differentiating countries by degree of economic development. In future research it would be interesting to examine the relationships between the variables used in the article with attention to regions. However, there are no data that we are aware of for such an analysis at present. The third limitation is that more variables such as distribution of wealth in a country or quality of social institutions are needed to be taken into consideration when examining the impact of income and wealth on SWB. The fourth limitation of this study is its timing. The focus was on the period from 2010 to 2019, which were times of the economic crisis and post-crisis era. Future research should compare the determinants of happiness during such periods with a relatively stable macroeconomic environment, that we hope will come in the near future.

ACKNOWLEDGMENTS

This work was supported by by the Slovak Scientific Grant Agency as part of the research project VEGA 2/0002/19.

References

AMINI, C., DOUARIN, E. (2020). Corruption and Life Satisfaction in Transition: Is Corruption Social Norm in Eastern Europe? [online]. *Social Indicators Research*, 151(2): 723–766. <https://doi.org/10.1007/s11205-020-02389-6>.

ANTOLINI, F., SIMONETTI, B. (2019). The Easterlin Paradox in Italy, or the Paradox in Measuring? Define Happiness Before Investigating It [online]. *Social Indicators Research*, 146(1–2): 263–285. <https://doi.org/10.1007/s11205-018-1890-7>.

APERGIS, N. (2018). The Impact of Greenhouse Gas Emissions on Personal Well-Being: Evidence from a Panel of 58 Countries and Aggregate and Regional Country Samples [online]. *Journal of Happiness Studies*, 19(1): 69–80. <https://doi.org/10.1007/s10902-016-9809-y>.

AZIZAN, N. H., MAHMUD, Z. (2018). Determinants of Subjective Well-Being: a Systematic Review [online]. *Environment-Behaviour Proceedings Journal*, 3(7): 135. <https://doi.org/10.21834/e-bpj.v3i7.1228>.

BAILEY, D., KATZ, J. N. (2011). Implementing Panel-Corrected Standard Errors in R: the PCSE Package [online]. *Journal of Statistical Software*, 42(Code Snippet 1). <https://doi.org/10.18637/jss.v042.c01>.

BARTOLINI, S., MIKUCKA, M., SARRACINO, F. (2017). Money, Trust and Happiness in Transition Countries: Evidence from Time Series [online]. *Social Indicators Research*, 130(1): 87–106. <https://doi.org/10.1007/s11205-015-1130-3>.

BEJA, E. L. (2020). Subjective Well-Being Approach to Valuing Unemployment: Direct and Indirect Cost [online]. *International Journal of Community Well-Being*, 3(3): 277–287. <https://doi.org/10.1007/s42413-019-00053-7>.

BRZEZINSKI, M. (2019). Diagnosing Unhappiness Dynamics: Evidence from Poland and Russia [online]. *Journal of Happiness Studies*, 20(7): 2291–2327. <https://doi.org/10.1007/s10902-018-0044-6>.

CERVELLATI, M., SUNDE, U. (2011). Life expectancy and economic growth: the role of the demographic transition [online]. *Journal of Economic Growth*, 16(2): 99–133. <https://doi.org/10.1007/s10887-011-9065-2>.

CHEN, Y., LI, T., SHI, Y., ZHOU, Y. (2014). Welfare Costs of Inflation: Evidence from China [online]. *Social Indicators Research*, 119(3): 1195–1218. <https://doi.org/10.1007/s11205-013-0553-y>.

CHENG, Z., KING, S., SMYTH, R., WANG, H. (2016). Housing Property Rights and Subjective Wellbeing in Urban China [online]. *European Journal of Political Economy*, 45: 160–174. <https://doi.org/10.1016/j.ejpoleco.2016.08.002>.

CUNADO, J., PEREZ DE GRACIA, F. (2012). Environment and Happiness: New Evidence for Spain [online]. *Social Indicators Research*, 112. <https://doi.org/10.1007/s11205-012-0038-4>.

D'AMBROSIO, C., JÄNTTI, M., LEPINTEUR, A. (2020). Money and Happiness: Income, Wealth and Subjective Well-Being [online]. *Social Indicators Research*, 148(1): 47–66. <https://doi.org/10.1007/s11205-019-02186-w>.

DIEGO-ROSELL, P., TORTORA, R., BIRD, J. (2018). International Determinants of Subjective Well-Being: Living in a Subjectively Material World [online]. *Journal of Happiness Studies*, 19(1): 123–143. <https://doi.org/10.1007/s10902-016-9812-3>.

DIENER, E., BISWAS-DIENER, R. (2002). Will Money Increase Subjective Well-Being? [online]. *Social Indicators Research*, 57(2): 119–169. <https://doi.org/10.1023/A:1014411319119>.

DIENER, E., HEINTZELMAN, S. J., KUSHLEV, K., TAY, L., WIRTZ, D., LUTES, L. D., OISHI, S. (2017). Findings all psychologists should know from the new science on subjective well-being [online]. *Canadian Psychology/Psychologie Canadienne*, 58(2): 87–104. <https://doi.org/10.1037/cap0000063>.

DIENER, E., OISHI, S., LUCAS, R. E. (2015). National accounts of subjective well-being [online]. *American Psychologis*t, 70(3): 234–242. <https://doi.org/10.1037/a0038899>.

DIENER, E., TAY, L., OISHI, S. (2013). Rising income and the subjective well-being of nations [online]. *Journal of Personality and Social Psychology*, 104(2): 267–276. <https://doi.org/10.1037/a0030487>.

EASTERLIN, R. (2015). Happiness and Economic Growth: the Evidence [online]. *Global Handbook of Quality of Life: Exploration of Well-Being of Nations and Continents.* <https://doi.org/10.1007/978-94-017-9178-6_12>.

EASTERLIN, R. A. (1974). Does Economic Growth Improve the Human Lot? Some Empirical Evidence [online]. *Nations and Households in Economic Growth*, Elsevier, 89–125. <https://doi.org/10.1016/B978-0-12-205050-3.50008-7>.

FERREIRA, S., AKAY, A., BRERETON, F., CUNADO, J., MARTINSSON, P., MORO, M. (2012). Life Satisfaction and Air Quality in Europe [online]. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.2114912>.

GLATZ, C., EDER, A. (2020). Patterns of Trust and Subjective Well-Being Across Europe: New Insights from Repeated Cross-Sectional Analyses Based on the European Social Survey 2002–2016 [online]. *Social Indicators Research*, 148(2): 417–439. <https://doi.org/10.1007/s11205-019-02212-x>.

GÓRNIK-DUROSE, M. E. (2020). Materialism and Well-Being Revisited: the Impact of Personality [online]. *Journal of Happiness Studies*, 21(1): 305–326. <https://doi.org/10.1007/s10902-019-00089-8>.

GRABLE, J. E., CUPPLES, S., FERNATT, F., ANDERSON, N. (2013). Evaluating the Link Between Perceived Income Adequacy and Financial Satisfaction: a Resource Deficit Hypothesis Approach [online]. *Social Indicators Research*, 114(3): 1109–1124. <https://doi.org/10.1007/s11205-012-0192-8>.

GROWIEC, K., GROWIEC, J. (2013). Trusting Only Whom You Know, Knowing Only Whom You Trust: the Joint Impact of Social Capital and Trust on Happiness in CEE Countries [online]. *Journal of Happiness Studies*, 15: 1–26. <https://doi.org/10.1007/s10902-013-9461-8>.

HOCHMAN, O., SKOPEK, N. (2013). The impact of wealth on subjective well-being: a comparison of three welfare-state regimes [online]. *Research in Social Stratification and Mobility*, 34: 127–141. <https://doi.org/10.1016/j.rssm.2013.07.003>.

HOLLY, S., MOHNEN, A. (2012). Impact of Working Hours on Work-Life Balance [online]. *SSRN Electronic Journal.* <https://doi.org/10.2139/ssrn.2135453>.

HOMMERICH, C., TIEFENBACH, T. (2018). Analyzing the Relationship Between Social Capital and Subjective Well-Being: the Mediating Role of Social Affiliation [online]. *Journal of Happiness Studies*, 19. <https://doi.org/10.1007/s10902-017-9859-9>.

JAKUBOW, A. (2016). Subjective Well-Being and the Welfare State: Giving a Fish or Teaching to Fish? [online]. *Social Indicators Research*, 128(3): 1147–1169. <https://doi.org/10.1007/s11205-015-1073-8>.

JONES, C., KLENOW, P. (2016). Beyond GDP? Welfare across Countries and Time [online]. *American Economic Review*, 106: 2426–2457. <https://doi.org/10.1257/aer.20110236>.

KAGEYAMA, J. (2009). Happiness and Sex Difference in Life Expectancy [online]. *Journal of Happiness Studies*, 13. <https://doi.org/10.1007/s10902-011-9301-7>.

KASSER, T., ROSENBLUM, K. L., SAMEROFF, A. J., DECI, E. L., NIEMIEC, C. P., RYAN, R. M., ÁRNADÓTTIR, O., BOND, R., DITTMAR, H., DUNGAN, N., HAWKS, S. (2014). Changes in materialism, changes in psychological well-being: Evidence from three longitudinal studies and an intervention experiment [online]. *Motivation and Emotion*, 38(1): 1–22. <https://doi.org/10.1007/s11031-013-9371-4>.

KENNEDY, P. (2008). *A guide to econometrics.* 6th Ed. Blackwell Pub.

LAMBSDORFF, J. (2007). The institutional economics of corruption and reform: Theory, evidence, and policy [online]. *The Institutional Economics of Corruption and Reform: Theory, Evidence, and Policy.* <https://doi.org/10.1017/CBO9780511492617>.

LI, J. (2016). Why Economic Growth did not Translate into Increased Happiness: Preliminary Results of a Multilevel Modeling of Happiness in China [online]. *Social Indicators Research*, 128(1): 241–263. <https://doi.org/10.1007/s11205-015-1028-0>.

LIM, H.-E., SHAW, D., LIAO, P.-S., DUAN, H. (2020). The Effects of Income on Happiness in East and South Asia: Societal Values Matter? [online]. *Journal of Happiness Studies*, 21(2): 391–415. <https://doi.org/10.1007/s10902-019-00088-9>.
LUCAS, R. E., SCHIMMACK, U. (2009). Income and well-being: How big is the gap between the rich and the poor? [online]. *Journal of Research in Personality*, 43(1): 75–78. <https://doi.org/10.1016/j.jrp.2008.09.004>.
LUO, J. (2020). A Pecuniary Explanation for the Heterogeneous Effects of Unemployment on Happiness [online]. *Journal of Happiness Studies*, 21(7): 2603–2628. <https://doi.org/10.1007/s10902-019-00198-4>.
MINGO, I., MONTECOLLE, S. (2014). Subjective and Objective Aspects of Free Time: the Italian Case [online]. *Journal of Happiness Studies*, 15(2): 425–441. <https://doi.org/10.1007/s10902-013-9429-8>.
MOOD, C., JONSSON, J. O. (2016). The Social Consequences of Poverty: an Empirical Test on Longitudinal Data [online]. *Social Indicators Research*, 127(2), 633–652. <https://doi.org/10.1007/s11205-015-0983-9>.
NG, W., DIENER, E. (2014). What matters to the rich and the poor? Subjective well-being, financial satisfaction, and postmaterialist needs across the world [online]. *Journal of Personality and Social Psychology*, 107(2): 326–338. <https://doi.org/10.1037/a0036856>.
NGAMABA, K. H. (2016). Determinants of subjective well-being in representative samples of nations [online]. *The European Journal of Public Health*, ckw103. <https://doi.org/10.1093/eurpub/ckw103>.
OKULICZ-KOZARYN, A., MAZELIS, J. M. (2017). More Unequal in Income, More Unequal in Wellbeing [online]. *Social Indicators Research*, 132(3): 953–975. <https://doi.org/10.1007/s11205-016-1327-0>.
PAPAVLASSOPULOS, N., KEPPLER, D. (2011). Life Expectancy as an Objective Factor of a Subjective Well-Being [online]. *Social Indicators Research*, 104(3): 475–505. <https://doi.org/10.1007/s11205-010-9757-6>.
PARK, H. M. (2011). *International University of Japan Public Management & Policy Analysis Program*, 53.
PIEREWAN, A. C., TAMPUBOLON, G. (2015). Happiness and Health in Europe: a Multivariate Multilevel Model [online]. *Applied Research in Quality of Life*, 10(2): 237–252. <https://doi.org/10.1007/s11482-014-9309-3>.
RODRÍGUEZ-POSE, A., MASLAUSKAITE, K. (2012). Can policy make us happier? Individual characteristics, socio-economic factors and life satisfaction in Central and Eastern Europe [online]. *Cambridge Journal of Regions, Economy and Society*, 5: 77–96. <https://doi.org/10.1093/cjres/rsr038>.
SEN, A., STIGLITZ, J., FITOUSSI, J. (2010). *Mis-measuring our lives: Why GDP doesn't add up?* The New Press.
SLAG, M., BURGER, M. J., VEENHOVEN, R. (2019). Did the Easterlin Paradox apply in South Korea between 1980 and 2015? A case study [online]. *International Review of Economics*, 66(4): 325–351. <https://doi.org/10.1007/s12232-019-00325-w>.
SPRUK, R., KEŠELJEVIĆ, A. (2016). Institutional Origins of Subjective Well-Being: Estimating the Effects of Economic Freedom on National Happiness [online]. *Journal of Happiness Studies*, 17(2), 659–712. <https://doi.org/10.1007/s10902-015-9616-x>.
STELZNER, M. (2021). Growth, Consumption, and Happiness: Modeling the Easterlin Paradox [online]. *Journal of Happiness Studies.* <https://doi.org/10.1007/s10902-021-00402-4>.
TIBESIGWA, B., VISSER, M., HODKINSON, B. (2016). Effects of Objective and Subjective Income Comparisons on Subjective Wellbeing [online]. *Social Indicators Research*, 128(1): 361–389. <https://doi.org/10.1007/s11205-015-1035-1>.
TORRES-REYNA, O. (2010). *Getting Started in Fixed/Random Effects Models using R* [online]. <https://www.princeton.edu/~otorres/Panel101R.pdf>.
VAN DER MEER, P. H. (2014). Gender, Unemployment and Subjective Well-Being: Why Being Unemployed Is Worse for Men than for Women [online]. *Social Indicators Research*, 115(1): 23–44. <https://doi.org/10.1007/s11205-012-0207-5>.
VEENHOVEN, R. (2006). Apparent Quality-of-Life in Nations: How Long and Happy People Live [online]. *Social Indicators Research*, 71: 61–86. <https://doi.org/10.1007/1-4020-3602-7_3>.
VEENHOVEN, R., VERGUNST, F. (2014). The Easterlin illusion: Economic growth does go with greater happiness [online]. *International Journal of Happiness and Development*, 1: 311. <https://doi.org/10.1504/IJHD.2014.066115>.
WIRTZ, A., NACHREINER, F. (2010). The Effects of Extended Working Hours on Health and Social Well-Being – a Comparative Analysis of Four Independent Samples [online]. *Chronobiology International*, 27(5): 1124–1134. <https://doi.org/10.3109/07420528.2010.490099>.
World Happiness Report 2020. (n.d.). [online]. [cit. 23.11.2021]. <https://worldhappiness.report/ed/2020>.
WU, C. (1999). The price of growth [online]. *Bulletin of the Atomic Scientists*, 55(5): 58–66. <https://doi.org/10.1080/00963402.1999.11460375>.
YIN, R., LEPINTEUR, A., CLARK, A. E., D'AMBROSIO, C. (2021). Life Satisfaction and the Human Development Index Across the World [online]. *Journal of Cross-Cultural Psychology*, 002202212110447. <https://doi.org/10.1177/00220221211044784>.
YU, G. B., LEE, D.-J., SIRGY, M. J., BOSNJAK, M. (2020). Household Income, Satisfaction with Standard of Living, and Subjective Well-Being. The Moderating Role of Happiness Materialism [online]. *Journal of Happiness Studies*, 21(8): 2851–2872. <https://doi.org/10.1007/s10902-019-00202-x>.
ZAGORSKI, K., EVANS, M. D. R., KELLEY, J., PIOTROWSKA, K. (2014). Does National Income Inequality Affect Individuals' Quality of Life in Europe? Inequality, Happiness, Finances, and Health [online]. *Social Indicators Research*, 117(3): 1089–1110. <https://doi.org/10.1007/s11205-013-0390-z>.

Factors of Differences in the Highest Wages of Employees in the Slovak Republic (2020 vs. 2010)

Viera Pacáková[1] | *University of Pardubice, Pardubice, Czech Republic*
Ľubica Sipková[2] | *University of Economics in Bratislava, Bratislava, Slovakia*
Petr Šild[3] | *University of Pardubice, Pardubice, Czech Republic*

Received 7.2.2022 (revision received 1.4.2022), **Accepted** (reviewed) 10.5.2022, **Published** 16.12.2022

Abstract

The article offers the results of statistical analysis of data on the highest wages of employees in the Slovak Republic in 2020. Descriptive analysis of sample data is supplemented by generalizing the results to the population of all employees whose salary exceeds the 99th percentile of the sample, by selected methods of statistical inference, which are probability models of the highest wages and analysis of variance. The analysis focuses on assessing the significance of the impact of selected demographic and social factors on the highest salaries of employees in SR in 2020 and their differences. The investigated factors there are gender, level of education, region of residence, the label of occupation, and age category. The article also focuses on inequalities in the number of employees at different levels of the monitored factors. The obtained results of the analysis are compared with the results of similar analysis from 2010.

Keywords	***DOI***	***JEL code***
The highest wages, factors, descriptive characteristics, probability models, analysis of variance, comparisons	*https://doi.org/10.54694/stat.2022.6*	*C46, D31, D33*

INTRODUCTION

Reliable information on inhabitants' and household' incomes is important in each country for many economic and political reasons. Income data are collected by several sample surveys and the data obtained are analysed by various methods. Often the analysis ends at the level of the sample data without generalizing the findings to the whole population. Because the sample survey

[1] Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administration, University of Pardubice, Studentská 95, 532 10 Pardubice, Czech Republic. Corresponding author: v.pacakova@gmail.com.
[2] Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia. E-mail: lubica.sipkova@gmail.com.
[3] Institute of Mathematics and Quantitative Methods, Faculty of Economics and Administration, University of Pardubice, Studentská 95, 532 10 Pardubice, Czech Republic. E-mail: Petr.Sild@student.upce.cz.

is always affected by randomness, we have to take the results of analyses of sample data only with considerable caution. Statistical inference methods using statistical software packages provide an effective tool for generalizing information from a sample to a population.

If the subject of the analysis is one random variable, e.g. wages of employees, the best and most comprehensive generalization of information from sample data to population is the probability distribution or density function of the observed variable. This knowledge will allow the calculation of important basic characteristics of the population, quantiles, probabilities of arbitrary intervals of values, etc.

This article provides the results of such a generalization by selected methods of statistical inference in the analysis of the highest wages of employees in the Slovak Republic (SR) in 2020 based on the Labour Information System sample survey, which has been carried out in the SR since 1992 by company Trexima Bratislava. The observed random variable is the average gross monthly wage (*salary_gm*) of employees in the Slovak Republic in 2020.

The starting point for the analysis there is 11 570 anonymized individual values of the variable *salary_gm* and also personal data of those employees in the Slovak Republic whose average gross monthly wage exceeded the 99th percentile equal to 4 863.17 EUR in the sample of all employees in 2020. The entire sample obtained by stratified random sampling covers more than half of the employees in the Slovak Republic (Trexima, 2022). The survey includes payroll and personnel data on employees.

The analysis also deals with the assessment of the impact of five factors on the inequalities in the highest wages in the Slovak Republic. These factors are personal demographic and social features of employees, specifically gender with categories male-female, together with classification variables as the level of education, the region of residence, label of occupation, and the categorized age.

The main goal of the article is to provide an answer to the question "Which demographic and social characteristics are typical for 1% of employees in the Slovak Republic who received the highest wages in 2020?" In order to meet this goal, it will be necessary to verify inequalities in the number of employees and in the level of their wages at different levels of the monitored factors.

The inspiration for the presented analysis in the article was also the publications Pacáková and Foltán (2011), and Pacáková et al. (2012), which have been used to compare the results of the highest wages analyses in 2020 vs. 2010 in the Slovak Republic. Both years are atypical, post-crisis years. The year 2010 followed the start of the global financial crisis, and 2020 was part of a global pandemic crisis over Covid-19.

1 LITERATURE REVIEW

Differentiation of income of private persons or households is a frequent and important topic of economic research. From various points of view, it is the subject of many publications in the world and domestic scientific economic journals. Deepening income polarization is perceived as one of the economic and social threats of the global world. The seriousness of the problem of income distribution and the associated problems of income inequality and poverty in the world is evidenced by the number of scientific publications. Foreign examples include publications Bell and Van Reenen (2013), Ayyash and Sek (2019), and Tomaskovic-Devey et al. (2020), in which attention is paid to the issue of extreme differences in wages. Conversely, the problem of low wages is the subject of publications Ryczkowski and Maksim (2018) or Skinner et al. (2002). The subject of many publications is the gender pay gap: Hara (2018), Artz and Taengnoi (2019), Whitehouse and Smith (2020).

Several works by Slovak and Czech authors are based on a statistical analysis of available data on employees' wages or household income. Because this is often sample data, it is important to generalize the information obtained to a population. Such publications include e.g. publications focused on modelling wage distributions, for example (Bílková, 2012, 2013; Malá, 2013; Bartošová, 2007; Pacáková and Sipková, 2007; Pacáková et al., 2005). The level and differentiation of employees' wages and households' incomes are examined frequently in a broader context.

The paper of Pauhofová and Martinák (2014) explores changes in income stratification of the Slovak population for determination of the possibility for consumption and propensity to save. The paper primarily examines the regional dimension of income stratification, differences in income stratification of residents in urban and rural areas, and differences in the distribution of income between genders. The analysis uses data based on national accounts data from the Slovak Statistical Office, Eurostat data, and administrative data on individual income from Social Insurance Agency in Slovakia.

Regional differentiation and development of wages after 1989 in the Slovak Republic is the topic of article Michálek (2007). The method of decomposition has facilitated the identification of not only deeper causes and implications of regional wage disparities but also their effects and impact in regions under study. The results point was two important facts, considerable regional wage differences, and their continuous deepening.

Analyses in the paper Pauhofová and Želinský (2015) are based on microdata from the Social Insurance Agency in Slovakia. According to the results, the income polarization in Slovakia is deepening and the economic performance of several districts is lowering. This results in extreme barriers for regional consumption at present, as well as the threat of generation of significantly low levels of pensions in the future.

The peculiarities of the development of the income structure in the Slovak economy examine the article Morvay (2013). Changes in the income structure were driven by shifts in the sectoral composition of the economy (e.g. expansion of branches with low wage share), but also by technological progress within sectors and branches (e.g. growth of capital intensity).

The specifics of the gender pay gap in post-socialist countries are dealt with in several articles. Mysíková (2012) quantifies the basic structure of the gender wage gaps in four Central European countries and finds the highest gender wage gap in the Czech Republic by using a dataset for the year 2008. In her study is mostly explained the observed wage gap by the remuneration effect, but is relatively less explained by the endowment effect in all considered countries. Tartaľová and Sovičová (2013) based on income data from EU SILC in the years 2005–2009 by statistical methods verify the gender pay gap in Slovakia in this period.

The aim of the paper Gottvald et al. (2013) is to capture by the wage equations several determinants affecting the level of wages in the Slovak Republic. In this paper analysis of wages determinants are based on data from the survey Information system on labour cost, which is realized by the company Trexima Bratislava.

Several publications focus on the lowest incomes in the context of the examination of poverty, for example Labudová et al. (2010), Malá (2019), Myslíková and Želinský (2019). On the contrary, in the articles Pacáková and Foltán (2011), and Pacáková et al. (2012), special emphasis is placed on the factors determining one percent of the highest wages in the Slovak Republic.

2 DATA AND METHODS

Data from the sample survey Information System on Labour Prices, which has been implemented in the Slovak Republic since 1992 by the company Trexima, were used for the analysis. The starting point for the analysis is 11570 values of the gross monthly wages of employees in the SR (random variable *salary_gm*), which exceeded the 99^{th} percentile of the sample, equal to 4 863.17 EUR in 2020. The entire sample, obtained by stratified random sampling carried out by Trexima, covers more than half of the employees in the Slovak Republic.

Descriptive characteristics of the central tendency (sampling average, median, quartiles), variability (coefficient of variation), and selected percentiles and their visualization using box plots provided clear information about the sample and its subsets.

Sampling is always influenced by randomness, so it is useful to generalize the information from the sample to the population by methods of statistical inference. The best and the most comprehensive

generalization of information from sample data is to find the probability model of the observed variable in the population. His knowledge will allow the calculation of important characteristics of the population, quantiles, probabilities of any intervals of values, etc.

As appropriate probability model for values exceeding the threshold a high enough (variable X_a), even with the existence of extreme values, it is considered to be 2-parameter Pareto distribution with distribution function in the form:

$$F_a(x) = 1 - (\frac{a}{x})^b, \quad x \geq a \tag{1}$$

where b is the shape parameter.

The basic characteristics of this probability model, that are mean, variance, and skewness, and thus the basic characteristics of the population, express the following formulas:

$$E(X) = \frac{ab}{b-1}, \; b > 1, \tag{2}$$

$$D(X) = \frac{a^2 b}{(b-1)^2(b-2)}, \; b > 2, \tag{3}$$

$$\gamma_1 = \frac{2\sqrt{b-2}(b+1)}{\sqrt{b}(b-3)}, \; b > 3. \tag{4}$$

Selected goodness-of-fit tests run to determine whether the variable X_a - *salary_gm* can be adequately modelled by a 2-parameter Pareto distribution.

Kolmogorov-Smirnov test (*K-S test*) compares the empirical cumulative distribution of the data $Fn(x)$ to the fitted cumulative distribution $F(x)$. The test statistic is given by the formula:

$$d_n = \sup_x |F_n(x) - F(x)|. \tag{5}$$

Cramer-Von Mises W^2 and *Watson* U^2 tests compare the empirical distribution function to the fitted CDF in different ways (see Pacáková et al., 2015). Since the smallest p-value amongst the tests performed is greater than 0.05, we cannot reject the idea that *salary_qm* comes from a 2-parameter Pareto distribution with 95% confidence.

The method analysis of *variance* for *salary_gm* has been used to compare the mean values of *salary_gm* for the different levels of monitored factors (gender, education, region, occupation, age category). The F-test verifies whether there are any significant differences amongst the means. If there are, the multiple range tests will tell which means are significantly different from which others. Especially for the existence of extreme values of the variable *salary_qm*, the conditions of this method are not met so it is convenient to choose the Kruskal-Wallis test which compares medians instead of means.

The *Kruskal-Wallis test* performs the null hypothesis that the medians of *salary_gm* within each of the levels of monitored factors are the same. The data from all the levels of factor is first combined and ranked from smallest to largest value. The average rank is then computed for the data at each level. Since the p-value is less than 0.05, there is a statistically significant difference amongst the medians at the 95.0% confidence level. The various plots can help to present the results of the comparison of the means (more detailed interpretation e.g. in Labudová et al., 2021).

3 RESULTS AND DISCUSSIONS

3.1 Comparison of the basic characteristics and distributions of variable *salary_gm*

We will start the analysis with calculate the basic descriptive statistics of gross monthly wages higher than the 99th percentile in the sample (variable *salary_gm*). Table 1 includes measures of central tendency, variability, and shape for *salary_gm* in the year 2020 and their comparison with the same characteristics in 2010.

Table 1 Comparison of the basic characteristics of *salary_gm* in 2010 and 2020

Year	Count	Average	Median	Coefficient of variation	Minimum	Maximum	Lower quartile	Upper quartile
2010	9 900	6 109.57	4 662.10	88.65 %	3 434.86	165 970	3 915.86	6 387.11
2020	11 570	7 726.92	6 193.77	78.86 %	4 863.17	218 333	5 366.60	8 061.09
Diff. Δ	–	1 617.35	1 531.67	–9.79 %	1 428.31	52 363	1 450.74	1 673.98

Source: Own calculation, output from Statgraphics Centurion

In the sample of 11 570 employees in the Slovak Republic with 1% of the highest gross monthly wages (above 4 863.17 EUR) in 2020. The average gross monthly wage was 7 726.92 EUR. Half of the employees had a lower wage and half higher than the median equal to 6 193.77 EUR and the gross monthly salary of a quarter of these employees exceeded 8 061.09 EUR. With a maximum salary of 218 333 EUR. The existence of extreme values in the set of the highest wages is also confirmed by the high value of the coefficient of variation (CV) of 78.86%.

The minimum value of wages exceeding the upper percentile increased in 2020 to 4 863.17 EUR, which is 1 428.31 EUR more than in 2010. The average value of 1% of the highest wages increased by 1 617.35 EUR compared to 2010. The increase was recorded in all characteristics except for the coefficient of variation, which indicates a lower variability in 2020 compared to 2010. The values of all percentiles also increased by the value of difference Δ (Table 2).

Table 2 Comparison of percentiles of *salary_gm* in 2010 and 2020

Year	1%	10%	25%	50%	75%	90%	99%	Upper quartile
2010	3 451.24	3 610.59	3 915.86	4 662.10	6 387.11	9 226.49	26 536.4	6 387.11
2020	4 879.13	5 044.07	5 366.60	6 193.77	8 061.09	11 252.90	27 350.1	8 061.09
Diff. Δ	1 427.89	1 433.48	1 450.74	1 531.67	1 673.98	2 026.41	813.7	1 673.98

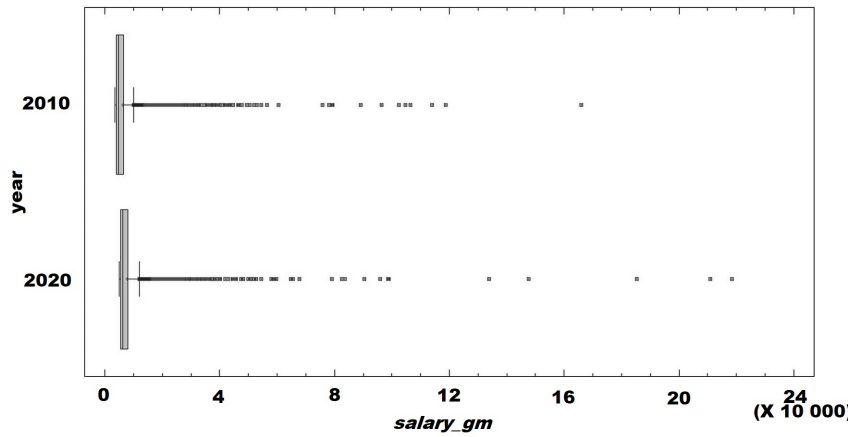
Source: Own calculation, output from Statgraphics Centurion

A graphic view of the basic characteristics of the values of gross monthly wages above the 99th percentile in 2010 and 2020 and their comparison is provided by the box plot in Figure 1.

Information from the sample has been generalized to the set of all employees in the Slovak Republic in 2020, whose *salary_gm* exceeded the value of 4 863.17 EUR. By applying goodness-of-fit tests. Table 3 shows the results of tests run to determine whether *salary_gm* can be adequately modelled by a 2-parameter Pareto distribution (1) with lower threshold a = 4 863.17 EUR and estimated shape parameter b = 2.75362.

Since the smallest p-value amongst the tests performed is greater than or equal to 0.05, we cannot reject the hypothesis that sample data of variable *salary_gm* comes from a 2-parameter Pareto distribution with 95% confidence. Figure 2 shows visually how adequately the 2-parameter Pareto distribution fits the data.

Figure 1 Box plots of gross monthly wages above the 99th percentile in 2010 and 2020



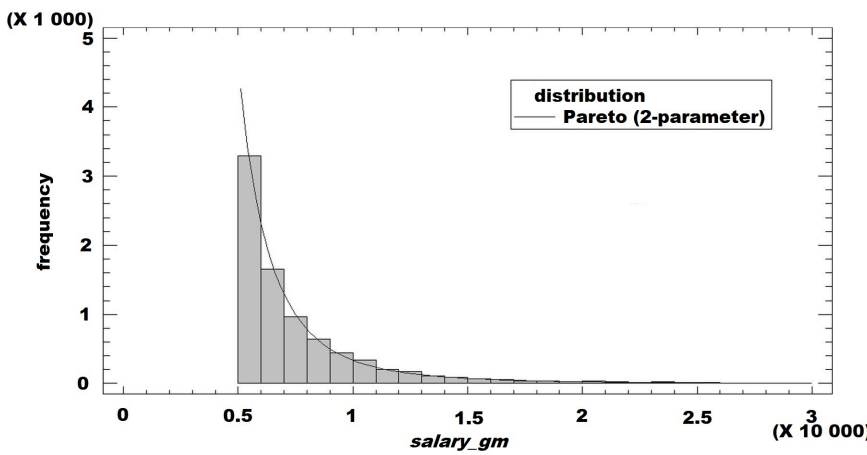
Source: Ma and Hellerstein (1999)

Table 3 The results of goodness-of-fit tests for *salary_gm*

	Pareto (2-parameter)		Pareto (2-parameter)
W^2	0.0000072044	**U^2**	–2 892.5
Modified form	–0.0000273656	**Modified form**	–2 892.7
P-Value	> = 0.10	**P-Value**	> = 0.10

Source: Output from Statgraphics Centurion

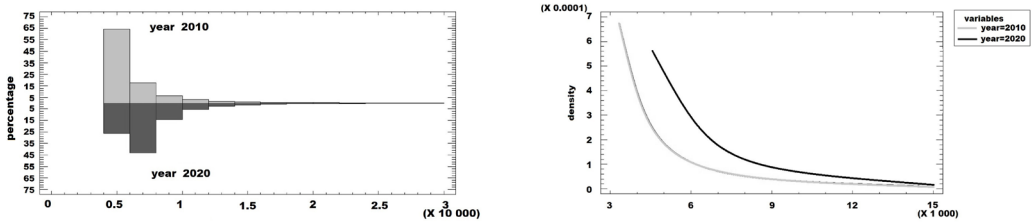
Figure 2 Graphical verification of good fit with 2-parameter Pareto distribution



Source: Output from Statgraphics Centurion

Because the Pareto distribution with lower threshold a = 3 434.86 and shape parameter b = 2.26487 was found as a suitable probability model for *salary_gm*. Exceeding the 99th percentile in 2010, we can visually compare the distribution of *salary_gm* in 2010 and 2020 (Figure 3).

Figure 3 Comparison of histograms and density functions of *salary_gm* in 2010 and 2020



Source: Output from Statgraphics Centurion

3.2 Factors of inequalities in the highest gross monthly wages

We will further focus on the assessment of factors of *gender, education, region of residence, employment classification* and *age category* for the wages of workers with salaries exceeding the 99th percentile in 2020. Table 4 contains the values of basic characteristics of *salary_gm* for men and women. They are all higher for men than for women and there is a difference in the last row of the table. The fact that among the best-earning employees is about 6 398, respectively 3.47 times more men than women indicate significant gender inequality. Also shift to lower-wage values for women is evident from the box plots in Figure 4.

Table 4 Comparison of descriptive statistics of *salary_gm* by gender in 2020

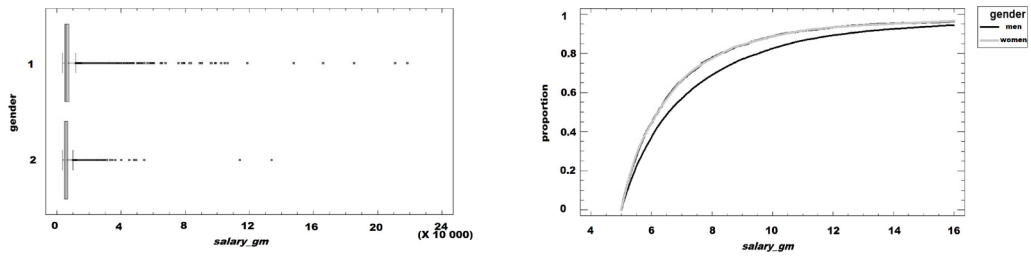
Gender	Count	Count (%)	Average	Median	Coeff. of variation	Minimum	Maximum	Lower quartile	Upper quartile
1-men	8 984	78%	7 908.38	6 260.61	82.19%	4 863.17	218 333	5 397.69	8 282.49
2-women	2 586	22%	7 096.53	5 979.41	61.13%	4 863.33	133 909	5 269.46	7 426.61
Diff.	6 398	56%	811.85	281.20	21.06	0	84 424	128.23	855.88

Source: Own calculation, output from Statgraphics Centurion

Kolmogorov-Smirnov tests (K-S tests) have been used to check up whether the 2-parameter Pareto distribution fits adequately the data of *salary_gm* for men and women. Since the smallest p-value = 0.696624 is higher than 0.05, we cannot reject the idea that *salary_gm* of men with 95% confidence comes from a 2-parameter Pareto distribution with an estimated lower threshold 4 863.17 and shape parameter 2.63536. In the same way has been verified by K-S test suitability of the 2-parameter Pareto distribution with estimated lower threshold 4 863.33 and shape parameter 3.26251 as probability model of *salary_gm* of women. The distribution functions of both fitted distribution models presents Figure 4, according to which for each value of *salary_gm* the probability of lower values is higher for women than for men.

Knowledge of the probability distribution can be used to calculate the mean (2), variance (3) and skewness (4) and compare some percentiles of gross monthly wages of employees, men and women, whose wages are higher than the 99th percentile of the sample. By calculating according to relation (2) we obtained the mean value of *salary_gm* of men equal to 7 836.93 EUR and of women equal to 7 012.859 EUR.

Figure 4 Box plots of the highest monthly wages by gender



Source: Output from Statgraphics Centurion

Both mean values increased significantly compared to 2010, the mean of men's wages by 1 406.08 EUR and the mean of women's wages by 1 542.66 EUR. The calculation of the variance according to (3) made it possible to compare the variability in the highest wages of men and women in 2020 using coefficients of variation with values of 77.3% for men and 49.3% for women. Because the shape parameter $b = 2.63536$ for men is less than 3, it is not possible to calculate the skewness γ_1 according to (4). For women the value of the skewness γ_1 is equal to 20.2, so the wages' distribution of women is strongly right-hand side, low wage values predominate. A comparison of selected percentiles by gender in 2020 in SR contains Table 5.

Table 5 Comparison of selected percentiles x_p of *salary_gm* (in EUR) by gender in 2020

CDF = p	Men	Women
0.10	5 061.54	5 022.95
0.25	5 424.10	5 311.64
0.50	6 326.26	6 014.55
0.75	8 229.51	7 438.28
0.90	11 651.30	9 850.23

Source: Own calculation, output from Statgraphics Centurion

The percentiles x_p have been found as critical values for the Pareto (2-parameter) model. The critical value x_p is defined as the value for the Pareto (2-parameter) such that Probability ($salary\_gm \leq x_p$) = $CDF = p$.

The *education factor* has 11 monitored levels: 0 – unspecified, 1 – basic, 2 – apprenticeship, 3 – secondary without GCSE, 4 – apprentices with graduation, 5 – complete secondary general, 6 – complete secondary vocational, 7 – higher professional, 8 – undergraduate 1st degree, 9 – undergraduate 2nd degree, 10 – undergraduate 3rd degree (at least PhD). The basic selection characteristics, as well as the percentage of employees with different levels of education, are shown in Table 6.

The most numerous group among the best-earning employees in the Slovak Republic in 2020 was the group 9 – undergraduate 2nd degree. However, the average wage in this group of employees in the population is significantly lower than in the groups of employees at levels 5 – complete secondary general and level 7 – higher professional (Figure 5).

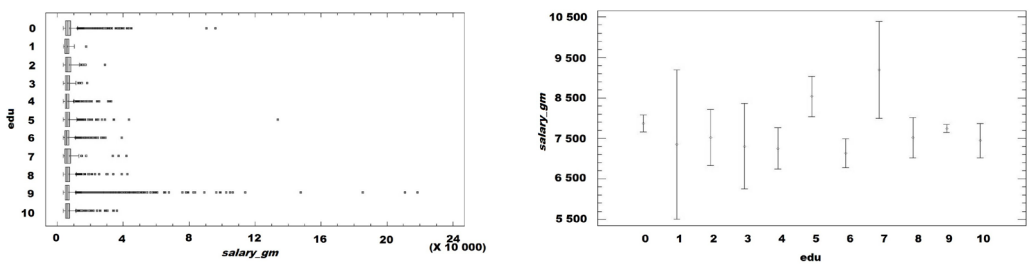
The significance of differences in the gross monthly wages of employees with a gross monthly wage above the 99th percentile in the Slovak Republic in 2020 is also caused by a factor the *employee's region*

Table 6 Descriptive statistics of the highest wages (*salary_gm*) by the level of education

Level of education	Count	Count (%)	Average	Median	Coeff. of variation (%)	Maximum	Lower quartile	Upper quartile
0	1 622	14%	7 864.55	6 336.90	65.58	95 825.5	5 434.34	8 428.14
1	21	0%	7 352.98	6 527.16	38.41	17 469.5	5 669.77	8 193.62
2	149	1%	7 523.70	6 280.06	44.27	31 223.5	5 400.79	8 731.85
3	64	1%	7 306.83	6 539.25	36.32	18 023.9	5 466.02	7 897.03
4	278	2%	7 252.61	6 019.33	53.30	32 527.6	5 259.63	7 459.20
5	286	2%	8 539.19	6 320.66	102.73	133 909.0	5 436.11	8 729.73
6	543	5%	7 133.60	6 186.15	39.91	28 476.4	5 365.06	7 987.72
7	50	0%	9 190.83	6 641.27	83.44	41 932.8	5 675.97	9 472.03
8	282	2%	7 517.94	6 260.56	51.29	42 448.6	5 478.61	8 204.68
9	7 875	68%	7 747.56	6 164.54	84.73	218 333.0	5 350.49	8 002.53
10	400	3%	7 443.83	6 053.48	55.04	34 006.0	5 266.20	7 783.14
Total	11 570	100%	7 726.92	6 193.77	78.86	218 333.0	5 366.60	8 061.09

Source: Own calculation, output from Statgraphics Centurion

Figure 5 Box plots and 95% intervals for means of *salary_gm* by level of education in 2020 in SR



Source: Output from Statgraphics Centurion

of residence. In the sample gross wages above the upper percentile are most often in the Bratislava region (61.7%) and the lowest representation there are in the Prešov (3.0%) and Banská Bystrica (4.1%) regions. In other regions the percentage of employees with a gross monthly wage above the upper percentile is approximately equal. More detailed information is provided in Table 7 and Figure 6.

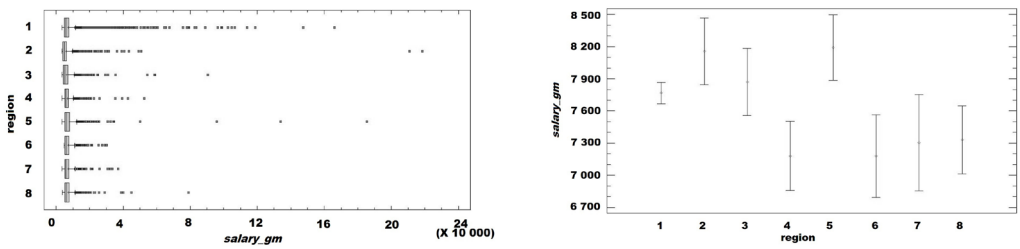
Significant variability in the amount of 1% of the highest gross monthly wages (*salary_gm*) in the regions of the Slovak Republic in 2020 is in the values beyond their upper quartile. From the box plot in Figure 7 it is clear that wages under the third quartile do not differ much, but there is great variability in a quarter of the highest values of *salary_gm* and the occurrence of extreme wages is highest in the Bratislava, Trnava and Žilina regions. The mean wages are the highest in the regions 2 – Trnava and 5 – Žilina, and in regions 4 – Nitra, 6 – Banská Bystrica, 7 – Prešov and 8 – Košice the mean wages are significantly lower than in regions 1 – Bratislava, 2 – Trnava, 3 – Trenčín and 5 – Žilina (Figure 6).

Table 7 Descriptive statistics of the highest wages (*salary_gm*) by region of residence

Region	Count	Count (%)	Average	Median	Coeff. of variation (%)	Maximum	Lower quartile	Upper quartile
1-Bratislava	7 134	62%	7 767.50	6 246.45	68.06	147 603.0	5 397.65	8 111.65
2-Trnava	735	6%	8 156.95	6 303.30	140.38	218 333.0	5 382.29	8 203.49
3-Trenčín	719	6%	7 870.23	6 177.00	79.12	90 367.6	5 335.46	8 381.28
4-Nitra	680	6%	7 179.99	5 926.10	54.62	52 568.8	5 282.19	7 479.06
5-Žilina	766	7%	8 191.20	6 201.74	115.86	185 375.0	5 354.81	8 257.49
6-B. Bystrica	478	4%	7 178.59	6 070.28	45.51	30 161.8	5 307.47	7 669.54
7-Prešov	351	3%	7 303.20	5 823.48	57.38	36 994.7	5 224.60	7 698.25
8-Košice	707	6%	7 328.78	6 110.37	62.26	79 040.2	5 265.73	7 825.19
Total	11 570	100%	7 726.92	6 193.77	78.86	218 333.0	5 366.60	8 061.09

Source: Own calculation, output from Statgraphics Centurion

Figure 6 Box plots and 95% intervals for means of *salary_gm* by region of residence in 2020



Source: Output from Statgraphics Centurion

Another important factor that affects the amount of gross monthly earnings, exceeding the upper percentile in Slovakia in 2020 was followed by a factor Classification of Occupations (isco1). Table 8 contains basic descriptive statistics for the values of *salary_gm*, corresponding to each of the 9 levels of this factor.

The most numerous groups in the sample there are the group 1 – Legislators, executives, up to 53.36 %, then the group 2 – Specialists (25.12 %) and more numerous is the group 3 – Technical and professionals (4.8%). Representation of categories 4 – Administrative staff, 5 – Service and trade workers, 7 – Skilled workers and craftsmen and 8 – Operators and fitters of machinery and equipment is less than 1%, while the gross wage of no employee with the classification 6 – Skilled workers in agriculture, forestry and fishing and 9 – Auxiliary and unskilled workers did not exceed the upper percentile of *salary_gm* in the Slovak Republic in 2020, just like in 2010. Category 0 – Unspecified employment included 14.77% of workers with a high average wage, even with several extreme wage values. The highest values of all descriptive characteristics in Table 8 there are concentrated for the values of *salary_gm* in category 1. These facts are also confirmed by plots in Figure 7. The mean wages in categories 0 – Unspecified employment and 1 – Legislators, executives are significantly higher compared to all other categories.

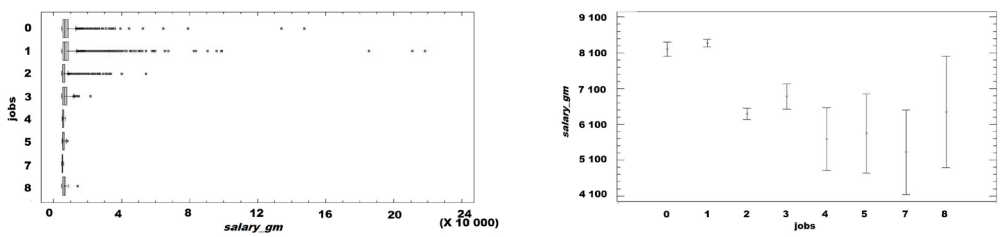
Significant differences in the numerical representation, as well as in the level of gross monthly wages, were also found in different age categories of employees of the monitored sample, as shown Table 9 and Figure 8.

Table 8 Descriptive statistics of the *salary_gm* by classification of occupations in 2020 in SR

Isco1	Count	Count (%)	Average	Median	Coeff. of variation (%)	Maximum	Lower quartile	Upper quartile
0	1 708	14.8%	8 199.94	6 444.50	83.01	147 603.00	5 468.65	8 556.92
1	6 174	53.4%	8 371.35	6 633.65	85.62	218 333.00	5 533.81	8 830.59
2	2 906	25.1%	6 396.57	5 631.29	42.26	54 440.50	5 158.63	6 504.82
3	556	4.8%	6 877.63	6 196.12	30.83	21 888.90	5 257.63	7 919.67
4	90	0.8%	5 687.10	5 610.62	10.25	6 990.02	5 147.92	6 117.91
5	57	0.5%	5 844.88	5 527.36	15.43	7 972.95	5 162.44	6 193.52
7	50	0.4%	5 318.98	5 224.72	6.21	5 960.94	5 065.38	5 636.65
8	29	0.3%	6 446.42	6 169.63	27.90	14 159.70	5 163.16	6 976.44
Total	11 570	100%	7 726.92	6 193.77	78.86	218 333.0	5 366.60	8 061.09

Source: Own calculation, output from Statgraphics Centurion

Figure 7 Box plots and 95% interval for means of *salary_gm* by classification of occupations in 2020 in SR



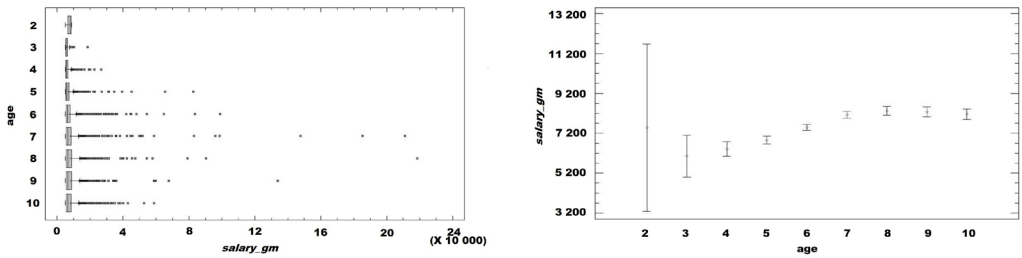
Source: Output from Statgraphics Centurion

Table 9 Descriptive statistics of the *salary_gm* by age category in 2020

Age category and interval	Count	Count (%)	Average	Median	Coeff. of variation (%)	Maximum	Lower quartile	Upper quartile
2: 20–24	4	0.03%	7 479.42	8 036.37	22.65	8 835.89	6 444.09	8 514.75
3: 25–29	64	0.55%	6 056.20	5 355.80	33.02	18 579.50	5 119.66	6 250.33
4: 30–34	522	4.51%	6 404.95	5 691.15	36.43	26 715.20	5 167.97	6 577.12
5: 35–39	1 767	15.27%	6 854.78	5 936.45	52.27	82 539.10	5 296.14	7 177.98
6: 40–44	2 979	25.75%	7 499.79	6 131.90	62.15	98 932.90	5 362.97	7 887.90
7: 45–49	2 579	22.29%	8 133.09	6 394.08	98.04	210 882.00	5 415.34	8 502.64
8: 50–54	1 550	13.40%	8 324.13	6 463.78	93.09	218 333.00	5 459.65	8 833.56
9: 55–59	1 073	9.27%	8 277.69	6 463.77	80.04	133 909.00	5 415.20	8 709.49
10: 60 and over	1 032	8.92%	8 164.50	6 393.39	67.15	58 811.30	5 424.33	8 559.19
Total	11 570	100%	7 726.92	6 193.77	78.86	218 333.00	5 366.60	8 061.09

Source: Own calculation, output from Statgraphics Centurion

Figure 8 Box plots and 95% interval for means of *salary_gm* by age category in 2020 in SR



Source: Output from Statgraphics Centurion

CONCLUSION

The results of the analysis in the article show that each factor whose impact on the level of the highest wages we examined significantly affects the wage differentiation of workers with wages above the 99th percentile in the Slovak Republic in 2020. Each monitored factor also causes large inequalities in the number of employees according to the levels of these factors.

The article includes also an assessment of changes in the highest wages of employees in the Slovak Republic in the time period from 2010 to 2020. The results in Section 2.1 show that over the course of ten years, the distribution of wages above the upper percentile of employees in the Slovak Republic has shifted by less than 1 500 EUR, but the shape of the probability distribution has hardly changed. This is absolutely clearly confirmed by the probability densities of the Pareto distributions for both years 2010 and 2020 in Figure 3. Knowledge of these distributions and their parameters allows to calculate and to compare the means $E(X)$ (by Formula 2) and the variances $D(X)$ (by Formula 3) of the one percent the highest wages of employees in the Slovak Republic in 2010 and 2020.

The mean of the highest wages increased from 6 150.44 EUR in 2010 to 7 636.39 EUR in 2020, which represents an average annual growth of 2.19%. For comparison, the average monthly wage of an employee in the Slovak economy increased from the value of EUR 769 in 2010 to EUR 1 133 in 2020, i.e. by an average of 3.95% per year. With annual average inflation of 1.63% over this period, this average wage growth is very modest (ŠÚ SR, 2022). The mean wage of 1% of the best earning employees is eight times the average wage in the SR in 2010 and the 6.74 times the average wage in the SR in 2020. Unfortunately, this applies to only 1% of employees.

On the contrary, variability of the highest wages decreased during the period from 2010 to 2020. The value of the coefficient of variation decreased from 129.1% in 2010 to 69.4% in 2020. Because the values of the shape parameters b of the Pareto distributions were less than the value 3 in both years, it was not possible to calculate and compare the skewness according to Formula (4) for these years.

The analysis of the highest wages in the crisis years of 2010 and 2020 in the Slovak Republic discovers the level of these wages and specified the groups of employees earned them.

The article may be inspiring for researchers focusing on the same issues for further research in this area.

ACKNOWLEDGMENT

This article was supported by grant No SGS_2021_011 *Utilization of data and information technologies as means supporting evidence-based decision making in development of a smart region* supported by the Student Grant Competition at the University of Pardubice, Faculty of Economics and Administration

and by grant No 1/0561/21 *The impact of the COVID-19 crisis on business demography and employment in the Slovak Republic and the EU* of the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic.

References

ARTZ, B., TAENGNOI, S. (2019). The Gender Gap in Raise Magnitudes of Hourly and Salary Workers [online]. *Journal of Labor Research*, 40(1): 84–105. <https://doi.org/10.1007/s12122-018-9277-8>.

AYYASH, M., SEK, S. K. (2019). Multilevel Analysis of Wage Inequality in Palestine [online]. *Statistika: Statistics and Economy Journal*, 99(3): 317–335. <https://www.czso.cz/documents/10180/88506454/32019719q3_317_ayyash_methodology.pdf/8c8fd1e3-04a9-4de0-a970-76c4f1f2fc7a?version=1.0>.

BARTOŠOVÁ, J. (2007). Pravděpodobnostní model rozdělení příjmů v České republice [online]. *Acta Oeconomica Pragensia*, 15(1): 7–12. <https://doi.org/10.18267/j.aop.30>.

BÍLKOVÁ, D. (2012). Recent Development of the Wage and Income Distribution in the Czech Republic [online]. *Prague Economic Papers*, 21(2): 233–250. <https://doi.org/10.18267/j.pep.421>.

BÍLKOVÁ, D. (2013). Modelování mzdových rozdělení posledních let v České republice s využitím l-momentů a predikce mzdových rozdělení podle odvětví. *E & M Ekonomie a management*, 16(4): 42–54.

BELL, B. D., VAN REENEN, J. (2013). Extreme Wage Inequality: Pay at the Very Top [online]. *American Economic Review*, 103(3): 153-157. <https://doi: 10.1257/aer.103.3.153>.

GOTTVALD, J., RIEVAJOVÁ, E., ŠIPIKALOVÁ, S. (2013). Determinants of Individual Wages in the Slovak Republic. *Ekonomický časopis/Journal of Economics*, 61(7): 672–689.

HARA, H. (2018). The gender wage gap across the wage distribution in Japan [online]. *Labour Economics*, 53: 213–229. <https://doi. org/10.1016/j.labeco.2018.04.007>.

LABUDOVÁ, V., VOJTKOVÁ, M., LINDA, B. (2010). Aplikácia viacrozmerných metód pri meraní chudoby. *E & M Ekonomie a management*, 13(1): 6–22.

LABUDOVÁ. V., PACÁKOVÁ. V., SIPKOVÁ. Ľ., ŠOLTÉS. E., VOJTKOVÁ. M. (2021). *Štatistické metódy pre ekonómov a manažérov.* Bratislava: Wolters Kluwer.

MALÁ, I. (2013). Použití konečných směsí logaritmicko-normálních rozdělení pro modelování příjmů českých domácností [online]. *Politická ekonomie*, 61(3): 356–372. <https://doi. org/10.18267/j.polek.902>.

MALÁ, I. (2019). Modelling Deprivation of the 50+ Population of the Czech Republic Based on the Share Survey [online]. *Statistika: Statistics and Economy Journal*, 99(2): 117–128. <https://www.czso.cz/documents/10180/88506448/32019719q2_117_mala_analyses.pdf/c1cab79e-840d-449f-b011-27a40ed8eecc?version=1.0>.

MORVAY, K. (2013). Osobitosti vývoja štruktúry príjmov v slovenskej ekonomike. *Ekonomický časopis*, 61(4): 327–343.

MICHÁLEK, A. (2007). Regionálne mzdové nerovnosti na Slovensku. *Geografický časopis*, 59(2): 181–209.

MYSLÍKOVÁ, M. (2012). Gender Gap in the Czech Republic and Central European Countries [online]. *Prague Economic Papers*, 21(3): 328–346. <https://doi. org/10.18267/j.pep.427>.

MYSLÍKOVÁ, M., ŽELINSKÝ, T. (2019). On the Measurement of the Income Poverty Rate: the Equivalence Scale across Europe [online]. *Statistika: Statistics and Economy Journal*, 99(4): 383–397. <https://www.czso.cz/documents/10180/88506452/32019719q4_383_mysikova_analyses.pdf/dbb49c82-8789-46be-8f40-e0de659dea35?version=1.0>.

PACÁKOVÁ, V. et al. (2015). *Štatistická indukcia pre ekonómov a manažérov.* Bratislava: Wolters Kluwer.

PACÁKOVÁ, V., FOLTÁN, F. (2011). Analysis of the highest wages in the Slovak Republic. *Scientific Papers of the University of Pardubice*, Series D, XVI(19): 172–180.

PACÁKOVÁ, V., SIPKOVÁ, Ľ., SODOMOVÁ, E. (2005). Štatistické modelovanie príjmov domácností v Slovenskej republike. *Ekonomický časopis,* 53(4): 427–439.

PACÁKOVÁ, V., SIPKOVÁ, Ľ. (2007). Generalized Lambda Distributions of Household's Incomes. *E+M Ekonomie a Management*, X(1): 98–107.

PACÁKOVÁ, V., LINDA, B., SIPKOVÁ, Ľ. (2012). Rozdelenie a faktory najvyšších miezd zamestnancov v Slovenskej republike. *Ekonomický časopis/Journal of Economics*, 60(9): 918–934.

PAUHOFOVÁ, I., MARTINÁK, D. (2014). Súvislosti príjmovej stratifikácie populácie Slovenskej republiky. *Ekonomický časopis/Journal of Economics:* 62(8): 842–860.

PAUHOFOVÁ, I., ŽELINSKÝ, T. (2015). Regionálne aspekty príjmovej polarizácie v Slovenskej republike [online]. *Politická ekonomie*, 63(6): 778–796. <https://doi. org/10.18267/j.polek.1026>.

RYCZKOWSKI, M., MAKSIM, M. (2018). Low wages – Coincidence or a result? Evidence from Poland [online]. *Acta Oeconomica*, 68(4): 549–572. <https://doi.org/10.1556/032.2018.68.4.4>.

SKINNER, C., STUTTARD, N., BEISSEL-DURRANT, G., JENKINS, J. (2002). The measurement of low pay in the UK Labour Force Survey [online]. *Oxford Bulletin of Economics and Statistics*, 64(S1): 653–676. <https://doi.org/10.1111/1468-0084.64.s.5>.

ŠÚ SR. (2022). *Potvrdenie pre infláciu a priemernú mesačnú mzdu* [online]. Bratislava. [cit. 10.3.2022]. <https://slovak.statistics.sk/wps/portal/ext/services/infoservis/confirmation/!ut/p/z0/04_Sj9CPykssy0xPLMnMz0vMAfIjo8ziw3wCLJycDB0NDMwszA0c_V0dLcwDPQy83U31C7IdFQHp6c-x>.

TARTAĽOVÁ, A., SOVIČOVÁ. T. (2013). Analýza príjmovej diferenciácie mužov a žien na Slovensku. *E&M Ekonomie a management*, 16(2): 54–65.

TOMASKOVIC-DEVEY, D. et al. (2020). Rising between-workplace inequalities in high-income countries [online]. *Proceedings of the National Academy of Sciences*, 117(17): 9277–9283. <https://doi:10.1073/pnas.1918249117>.

TREXIMA. (2022). *Mzdy a pracovné podmienky* [online]. Bratislava. <https://www.trexima.sk/portfolio/mzdy-pracovne-podmienky>.

WHITEHOUSE, G., SMITH, M. (2020). Equal pay for work of equal value. wage-setting and the gender pay gap [online]. *Journal of Industrial Relations*, 62(4). <https://doi.org/10.1177/0022185620943626>.

ŽELINSKÝ, T., HUDEC, O. (2008). Odhad subjektívnej chudoby na Slovensku založený na distribučnej funkcii rozdelenia príjmov. *Forum statisticum slovacum*, 4(7): 152–157.

Review of Visualization Methods for Categorical Data in Cluster Analysis

Jana Cibulková[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Barbora Kupková[2] | *Prague University of Economics and Business, Prague, Czech Republic*

Received 25.1.2022 (revision received 9.8.2022), **Accepted** (reviewed) 1.9.2022, **Published** 16.12.2022

Abstract

The paper focuses on visualization methods suitable for outcomes of cluster analysis of categorical data (nominal data, specifically). Since nominal data have no inherent order, their graphical representation is often challenging or very limited. This paper aims to provide a list of common visualization methods in the domain of cluster analysis of objects characterized by nominal variables. Firstly, the various plot types (such as clustering scatter plot, dendrogram, icicle plot) for cluster analysis are presented, and their suitability for presenting clusters of nominal data is discussed. Then, we study approaches of sorting nominal values on chart axes in such a way that would improve visualization of the data. Lastly, we introduce a simple alternative to cluster scatter plot for nominal data, that makes the final visualization of clustering solution more efficient since the pattern and groups in data are now more apparent. The suggested method is demonstrated in illustrative examples.

Keywords	***DOI***	***JEL code***
Cluster analysis, nominal data, hierarchical clustering, visualization	*https://doi.org/10.54694/stat.2022.4*	*C38, C18*

INTRODUCTION

Cluster analysis belongs to a group of unsupervised learning methods. Usually, its objective is to divide a set of objects into groups, called clusters. The aim is to define clusters in such a way that objects would be homogeneous within one cluster and heterogeneous among different clusters. There have been many clustering algorithms developed in very different fields: artificial intelligence, information technology, image processing, biology, psychology, marketing, etc. In this paper we focus mainly on hierarchical cluster analysis (HCA) and explain how its clustering solutions may be visualized. However, many visualization methods (including newly proposed ones) do not limit themselves to HCA only.

[1] Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: jana.cibulkova@vse.cz.
[2] Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: kupb01@vse.cz.

Often, it is required to use cluster analysis on categorical data, especially nominal data. This situation might occur in the field of biology (taxonomic category), in medicine (the listing of drugs), in marketing (marketing personas), etc. Nominal data clustering has not yet been studied to the same extent as quantitative data clustering has been, which also applies to its visualization. In this paper, we focus on visualization methods suitable for outcomes of cluster analysis of nominal data since visualization is becoming increasingly important for the analysis and research of multidimensional data (Ma and Hellerstein, 1999).

Working with nominal data has its specifics: it is not possible to sort nominal values in any objective way; we can only state whether the values are identical or not. This causes problems with the visualization procedures commonly used for quantitative data as there is no clear right way to place variables on axes of these graphs. Technically, any order is correct. However, as shown in this paper, there are certain ways of sorting these variables on the axes, which can significantly increase the quality of the information that the graph provides.

Several authors introduced new methods and approaches suitable for the visualization of categorical data in cluster analysis. Andrews (1972) presented a smoothed version of a parallel coordinate plot. Chernoff (1973) introduced a method that produces different images of human faces. Later, more advanced techniques were proposed. Hofmann and Buhmann (1995) proposed three strategies derived in the maximum entropy framework for visualizing data structures. Pölzlbauer et al. (2006) and Vesanto (1999) presented visualization methods using self-organizing maps, while the use of a minimum spanning tree for visualizing of HCA is discussed by Kim et al. (2000). Itoh et al. (2004) proposed an algorithm that can provide overviews of structures and the content of the hierarchical data. Chang and Ding (2005) proposed a method for visualizing clustered categorical data based on three-dimensional space.

Other authors, such as Ma and Hellerstein (1999) or Rosario et al. (2004) explored possibilities of ordering nominal data in order to improve visualization. These methods may be useful if one wishes to apply a well-known visualization technique that is commonly used for quantitative data on nominal data.

This paper aims to provide an overview of various available visualization methods in the domain of cluster analysis of objects characterized by nominal variables. Firstly, the most common visualization methods for cluster analysis are presented, and their suitability for presenting clusters of nominal data is discussed. Then, we study approaches of sorting nominal values.

Lastly, we present our alternative of cluster scatter plot for nominal data, which makes the final visualization of clustering solution more efficient since the pattern and groups in data are now more apparent. The suggested methods are demonstrated in illustrative examples, and all the computations and graphs, unless stated otherwise, were prepared by authors in R (R Core Team, 2021); HCA of nominal data was done using package *nomclust* (Šulc et al. 2021).

1 OVERVIEW OF VISUALIZATION METHODS FOR HCA

This section contains an overview of visualization methods suitable for cluster analysis. Each method is briefly described, a simple example is provided, and a method's suitability for presenting clusters of nominal data is discussed.

1.1 Dendrogram

A *dendrogram* is a diagram, that illustrates clusters creation in the process of HCA. Figure 1 shows an example of simple dendrogram on a dataset of 17 observations. According to Sibson (1973), we define a dendrogram to be function:

$$c\colon [0,\infty] \to E(D), \tag{1}$$

where D is the dataset, $E(D)$ is the set of equivalence relations on D, δ represents distance between objects, and c satisfies these conditions:

$$h \leq h' \text{ implies } c(h) \subseteq c(h'), \tag{2}$$

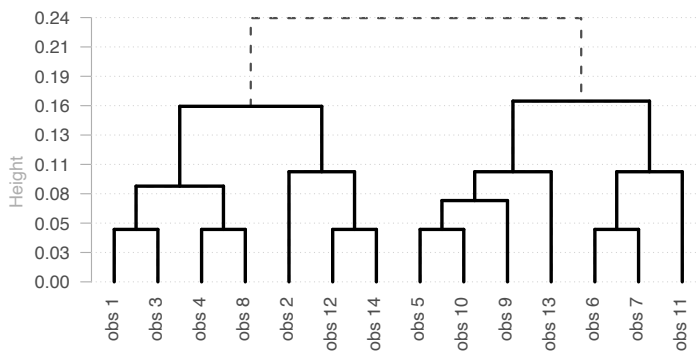
$$c(h) \text{ is eventually } D \times D, \tag{3}$$

$$c(h + \delta) = c(h) \text{ for all small enough } \delta > 0. \tag{4}$$

Hence, a dendrogram is a nested sequence of partitions with associated numerical levels. A dendrogram is usually represented as a tree diagram, but there is a great deal of freedom, and various alterations exist.

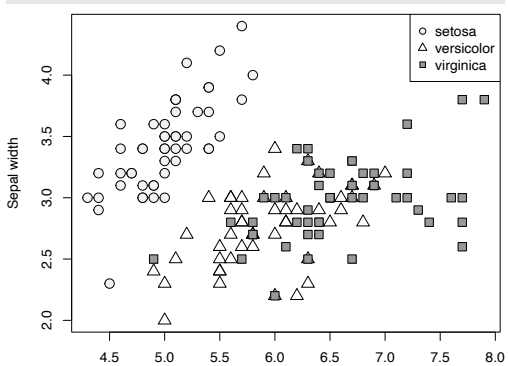
As mentioned above, the dendrogram represents the HCA process. Therefore, dataset with any type (including nominal) may be visualized, which is a big advantage. The disadvantage of this visualization method is that with an increasing number of observations, the dendrogram becomes difficult to read.

Figure 1 Example of a simple dendrogram



Source: Authors

Figure 2 Example of a simple cluster scatter plot on dataset iris



Source: Authors

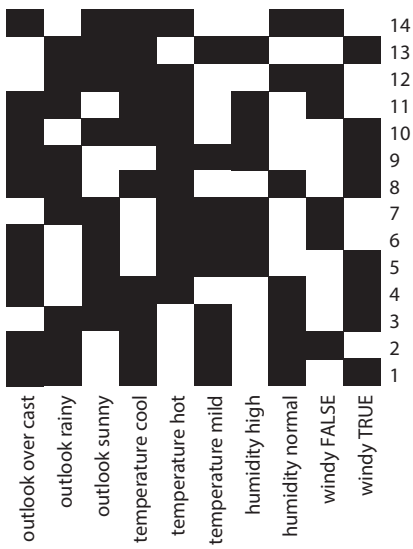
1.2 Cluster scatter plot

A *scatter plot* uses dots to represent values for two different numeric variables. The position of each dot relative to the horizontal and vertical axes indicates values for an individual data point. Scatter plots are used to observe relationships between variables. Up to four variables can be plotted in a scatter plot: two numerical variables on the *x*- and *y*-axis, a numerical or ordinal variable for the definition of point size, and a nominal variable for color definition (Rahlf, 2019).

Cluster scatter plot usually represents three variables: two chosen numerical variables on the *x*- and *y*-axis a nominal variable for

a color definition of an assigned cluster. This is a powerful visualization tool since it shows objects assignation into a cluster with respect to two chosen variables; it shows how well separated the clusters are and how many observations are within each cluster. Figure 2 shows such a cluster scatter plot for dataset iris (Anderson, 1935; Fisher, 1936). However, there is no clear right way to place variables on axes of a scatter plot when dealing with nominal data. Moreover, nominal data usually takes value from a relatively small number of categories; hence problem with overlapping points most likely occur even if we know how to place nominal variables on axes.

Figure 3 Example of a heat map



Source: Authors

1.3 Heatmap

A *heat map* is a two-dimensional matrix in which the cells are colored depending on their value. When visualizing outcomes of cluster analysis, color represents an assignation into a given cluster. It may be a table with individual data or aggregated values (Rahlf, 2019).

Heatmaps can be used for the visualization of all data types and their assignations into clusters. Before constructing a heatmap of clustering solution of nominal data, all nominal attributes are usually transformed into binary variables that are then treated as numeric. Hence, if the nominal attribute has k possible values, it is replaced by $k - 1$ synthetic binary variable, the i-th being 0 if the value is one of the first i in the ordering and 1 otherwise (Witten, 2011). Heatmap becomes hard to read easily when there are many variables with numerous categories, and binary transformation is performed, see Figure 3.

There is no rule stating how the rows or columns should be arranged (Rahlf, 2019). Rearranging rows or columns in a meaningful way would increase the overall readability of the plot, and it could help to discover hidden patterns in the data.

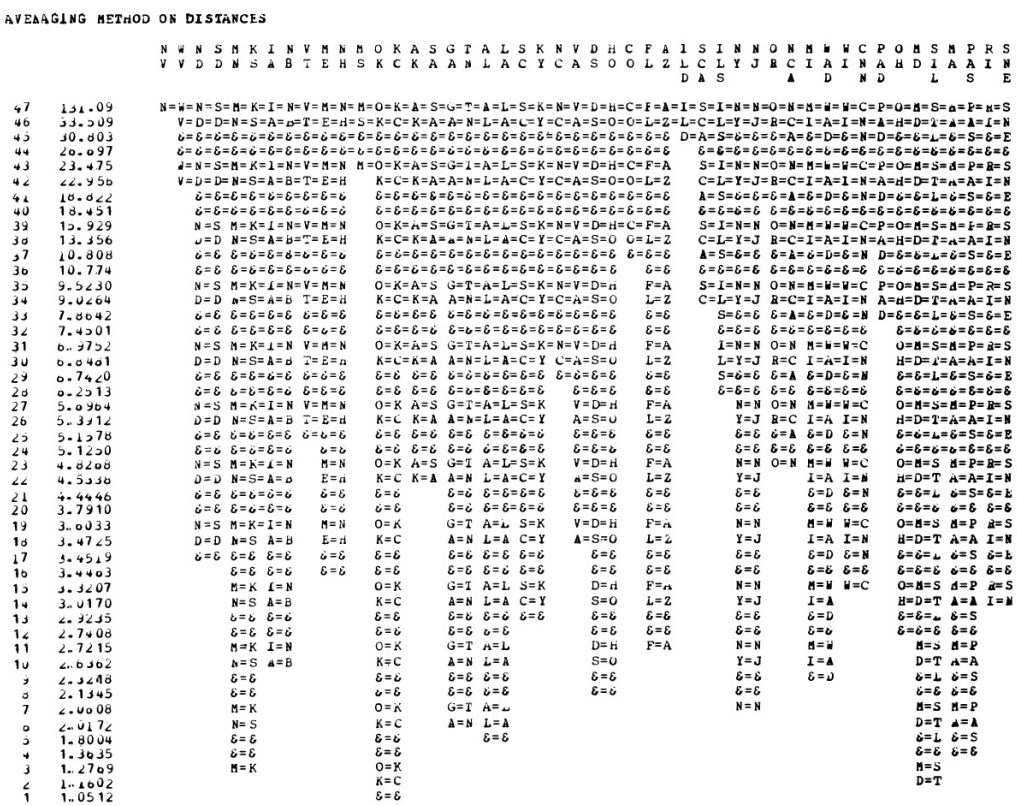
1.4 Icicle plot

According to Kruskal and Landwehr (1983), an *icicle plot* should be easier to read off which object belongs to which clusters and which objects join or drop out from a cluster as we move up or down the levels of hierarchy. However, this method was proposed in 1983, and it took into consideration the very limited capability of printers to print plots, see Figure 4. This plot is basically an upside-down variation of a dendrogram; therefore, it is suitable for visualization of the process of hierarchical clustering of any data type, including nominal one.

1.5 Andrew's plot

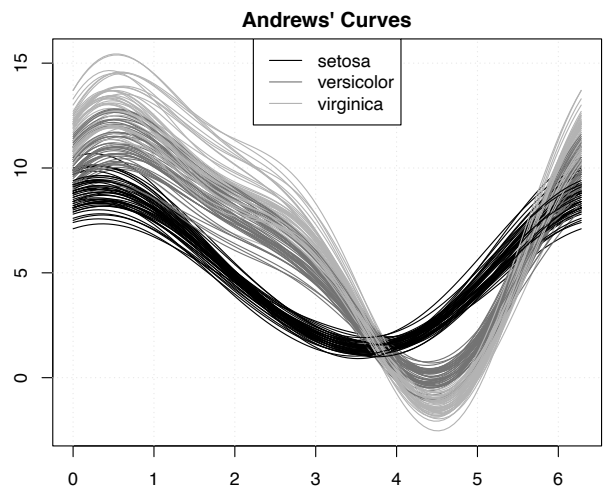
Andrew (1972) introduced a method to plot high-dimensional data with curves. This plot is based on the same principles as a parallel coordinate plot, and it is called *Andrew's plot* or *Andrew's curve*. Each curve represents one object, obtained by using the components of the data vectors as coefficients of orthogonal sinusoids, which are then added together pointwise. Figure 5 shows Andrew's plot for dataset iris (Anderson, 1935; Fisher, 1936). This graph surely can be used for the visualization of clustering solutions. However, it doesn't solve the problem with sorting categories, and overlaps of curves make it difficult to see the inherent structure in the data (especially if clusters are not completely homogeneous).

Figure 4 Example of an icicle plot



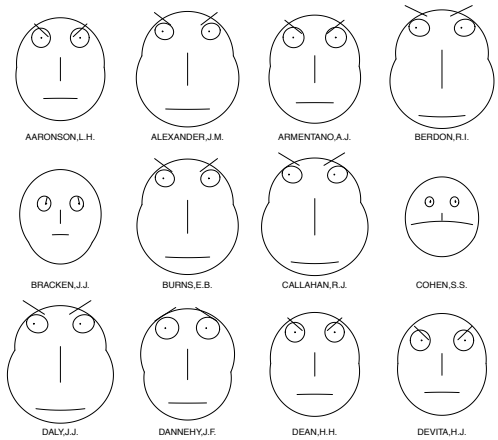
Source: Kruskal and Landwehr (1983)

Figure 5 Example of Andrews plot on dataset iris



Source: Authors

Figure 6 Example of Chernoff faces on dataset USJudgeRatings



Source: Authors

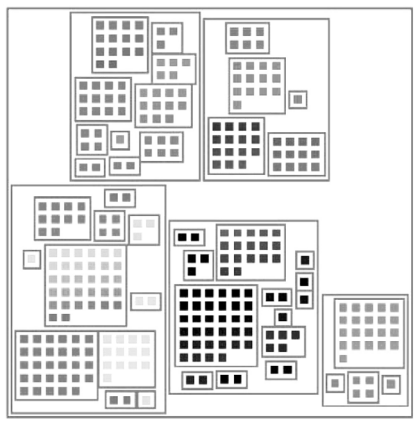
1.6 Chernoff's faces

Chernoff (1973) famously introduced a method to visualize multivariate data with faces. Each point in a d-dimensional space ($d \leq 18$) is represented by a facial caricature whose features such as length of nose, eyebrow position, mouth size, and shape are determined by the value of the corresponding variable. Figure 6 shows *Chernoff's* faces on dataset USJudgeRatings using library TeachingDemos in R (Snow, 2020).

1.7 Itoh's nested rectangles

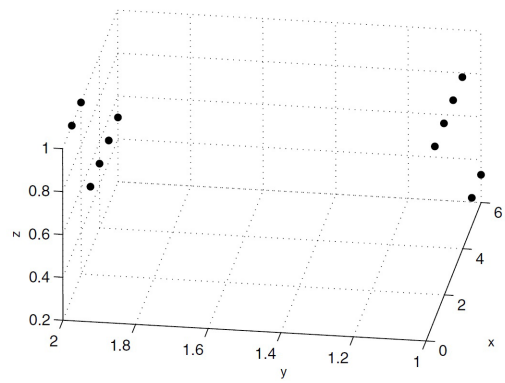
Itoh et al. (2004) presented a technique for representing large-scale hierarchical data by using nested rectangles. It first packs icons or thumbnails of the lowest-level data and then generates rectangular borders that enclose the packed data.

Figure 7 Hierarchical data using Strip Squarified Treemap



Source: Itoh (2004)

Figure 8 Plot of the two clusters



Source: Chang and Ding (2005)

It repeats the process of generating rectangles that enclose the lower-level rectangles until the highest-level rectangles are packed. It provides good overviews of complete structures and the content of the data in one display space. The approach refers to Delaunay triangular meshes connecting the centers of rectangles to find gaps where rectangles can be placed, see Figure 7.

1.8 Chang's and Ding's three-dimensional scatter plot

Chang and Ding (2005) proposed a method for visualizing clustered categorical data. Their method allows users to adjust the clustering parameters based on the visualization, so it is a new visualization method as well as a new clustering method. In this method, a special three-dimensional coordinate system is used

to represent the clustered categorical data. The three-dimensional coordinate system to plot a variable's value is constructed such that the x-axis represents the variables, the y-axis represents the variable's values, and the z-axis represents the probability that the variable's value is in the cluster. To display a set of clusters, the methods construct a coordinate system such that interference among different clusters can be minimized in order to observe closeness.

The visualization of the two categorical clusters using this method is shown in Figure 8. However, this approach doesn't provide information about a cluster's size, and it may be confusing when a large number of variables and their categories are visualized.

2 SORTING VARIABLES

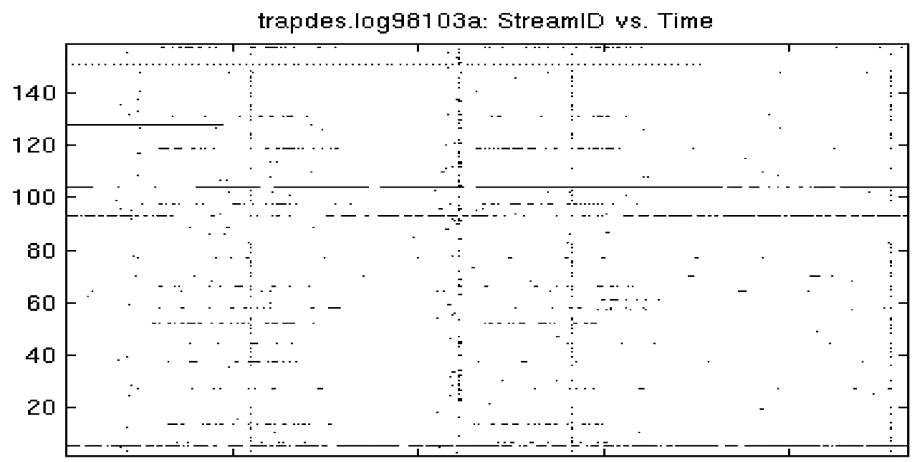
In this section, we focus on reordering categories of nominal variables in such a way that would:

- improve visualization of the data,
- allow us to adjust well-known methods for clusters visualization of quantitative data and to extend them nominal data as well.

Although the importance of sorting categorical values on chart axes is obvious, there are not many approaches to sorting them yet. Among the best known is the manual sorting of categories, which requires an experienced user. Another option to sort values of the nominal variables is to sort by an auxiliary quantitative variable, such as time. This method is not intended for visualizations and generally does not lead to satisfactory results (Ma and Hellerstein, 1999).

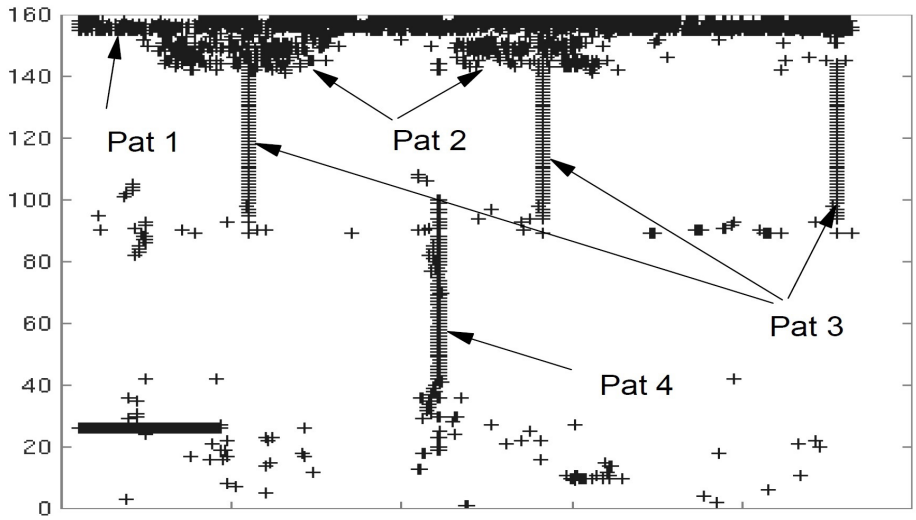
Ma and Hellerstein (1999) proposed an algorithm for ordering nominal data by constructing natural clusters, sequencing these clusters to minimize order conflicts (situations in which the same category must have two or more positions in the ordering) and ordering the values within the clusters to eliminate the above-mentioned order conflicts. The effect of this approach can be seen in Figures 9 and 10, where x-axis represents time and y-axis represents a host name in a production network. We can see which hosts generate events at specific time in Figure 10, while there are no obvious patterns visible in Figure 9. Even though the algorithm seems effective, it is computationally demanding.

Figure 9 Scatter plot example of randomly ordered data



Source: Ma and Hellerstein (1999)

Figure 10 Scatter plot example of data ordered by Ma's and Hellerstein's algorithm



Source: Ma and Hellerstein (1999)

Rosario et al. (2004) proposed another method, called the *Distance-Quantification-Classing approach* (DQC), to preprocess nominal variables before being imported into a visual exploration tool. This method solves the problem of order-and-spacing assignation among nominal values and reduces the number of distinct values to display. It works in three steps:

1 *Distance Step:* We identify a set of independent dimensions that can be used to calculate the distance between nominal values.
2 *Quantification Step:* We use the independent dimensions and the distance information to assign order and spacing among the nominal values.
3 *Classing Step:* We use results from the previous steps to determine which values within the domain of a variable are similar to each other and thus can be grouped together.

Each step in the DQC approach can be accomplished by a variety of techniques.

3 ALTERNATIVE OF CLUSTER SCATTER PLOT FOR NOMINAL DATA

In this section, we present an alternative of cluster scatter plot for nominal data because cluster scatter plot is such an essential visualization tool for data analysis.

The alternative of cluster scatter plot for nominal data shows a relationship between two nominal variables. Let's assume, that we have a dataset with nominal variables var_1, var_2, … var_m. Let's also assume that cat_l is a number of unique categories of a variable var_l; $l = 1, 2 \ldots, m$. Moreover, let's assume that we know an assignation to a cluster of each observation, let's assume k is number of clusters of our final clustering solution, that we wish to visualize on the plot.

To create a good visualization of clusters of objects represented by nominal variables *var_i* and *var_j*; where $1 \leq i, j \leq m$; $i \neq j$, we need to solve following three problems:

- *Sorting and spacing,*
- *Cardinality,*

- *Clarity of the graph.*

3.1 Sorting and spacing

As mentioned in Section 1.2, a scatter plot uses dots to represent values for two different numeric variables. The position of each dot on the horizontal and vertical axis indicates values for an individual data point. We aim to create an alternative to a scatter plot that uses dots to represent values for two different *nominal* variables var_i and var_j; $1 \leq i, j \leq m$; $i \neq j$. Hence, we need to assign a numerical value to each nominal value. In order to simplify the problem, we assume that distances among values are equal. Hence, we only need to sort values of var_i and var_j on *x*-axis and *y*-axis in such a way that clusters will be the most homogeneous and compact. These are the step we are proposing:

1. We calculate contingency tables of relative frequencies of var_i and var_j for each of k clusters. Let's denote C is an array of k contingency tables.
2. Sorting of var_i values: We calculate the sum of relative frequencies across all var_i values for each cluster. This way a new temporary dataset of k columns and cat_i rows is created. Each row corresponds to a vector of relative frequencies of the corresponding value of var_i of clusters 1 to k. We perform average-linkage HCA with Euclid distance on this temporary dataset. We obtain a dendrogram and from that dendrogram we can easily get its sorted leaves, that correspond to desired sorted values of var_i.
3. Sorting of var_j values, analogously to var_i.
4. We re-arrange contingency tables C according to the obtained new order for values of var_i and var_j.

3.2 Cardinality and clarity of the graph

Rahlf (2019) says that up to four variables can be plotted in a scatter plot: two numerical variables on the x- and y-axis, a numerical or ordinal variable for the definition of dot size, and a nominal variable for color definition. In our case, the first and second variables on the x- and y-axis need to be nominal variables. Color (and dot shape) definition must correspond to the assigned cluster. Dot size must correspond to the overall relative frequency of given values combination of examined two variables. Moreover, the relative frequencies need to be reasonably scaled so all the dots on the plot would be visible.

Hence *min-max normalization* is applied to contingency tables, to set point size within a range from a to b. Normalized contingency tables C' follow this formula:

$$C' = \frac{C - min(C)}{max(C) - min(C)} \times (b - a) + a\,, \tag{5}$$

where C is an array of k contingency tables, $\min(C)$ is minimal relative frequency across all k contingency tables, $\max(C)$ is maximal relative frequency across all k contingency tables. A range of point sizes in the final graph, defined by a and b, can be set based on personal preferences, we suggest $a = 1$ and $b = 6$.

3.3 Illustrative example of newly proposed visualization method applied on data

We demonstrate the newly proposed visualization method on a dataset from Cortez and Silva (2008). The dataset was obtained in a survey of students' math courses in a secondary school. It contains a lot of interesting social, gender, and study information about students. For our analysis, we performed average-linkage HCA with Eskin distance (Eskin et al., 2002) measure on three nominal variables:

- *Mjob* – mother's job;
  nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other',
- *Fjob* – father's job;

nominal: 'teacher', 'health' care related, civil 'services', 'at_home' or 'other',

- *Address* – student's home address type;
  nominal: 'U' – urban or 'R' – rural.

The above mentioned Eskin distance is a distance measure which can express a dissimilarity between two objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ that correspond to rows of given categorical data matrix $X = [\boldsymbol{x}_{ic}]$, where $i = 1, 2, \ldots, n$ and $c = 1, 2, \ldots, m$. It can be calculated as follows:

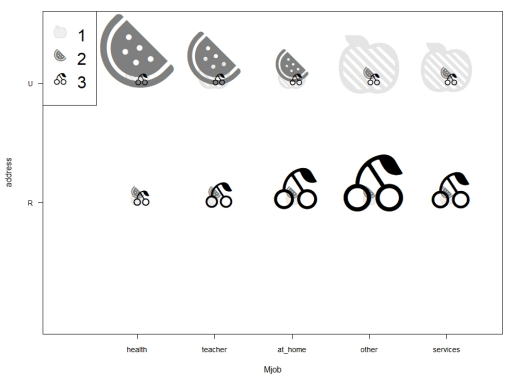
$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{S(\boldsymbol{x}_i, \boldsymbol{x}_j)} - 1 , \tag{6}$$

where $S(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is total similarity between the objects $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ calculated as arithmetic mean of similarities $S(x_{ic}, x_{jc})$ given by formula:

$$S(x_{ic}, x_{jc}) = \begin{cases} 1; \; if \; x_{ic} = x_{jc} \\ \dfrac{K_c^2}{K_c^2 + 2} \;\; if \; x_{ic} \neq x_{jc} \end{cases} , \tag{7}$$

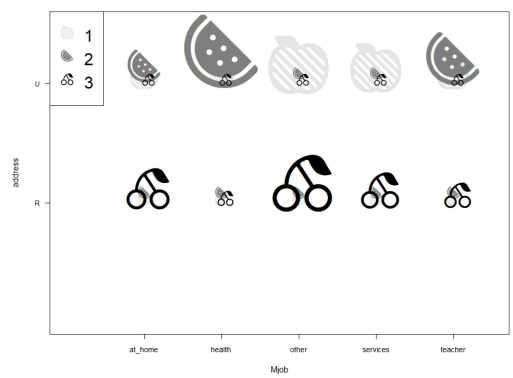
where the number of categories of the c-th variable is denoted as K_c.

Figure 11 Sorted alternative of cluster scatter plot for nominal data



Source: Authors

Figure 12 Unsorted alternative of cluster scatter plot for nominal data



Source: Authors

The Eskin distance measure is implemented in R package nomclust (Šulc et al., 2021), that was used to perform HCA. Then, observations were divided into three clusters.

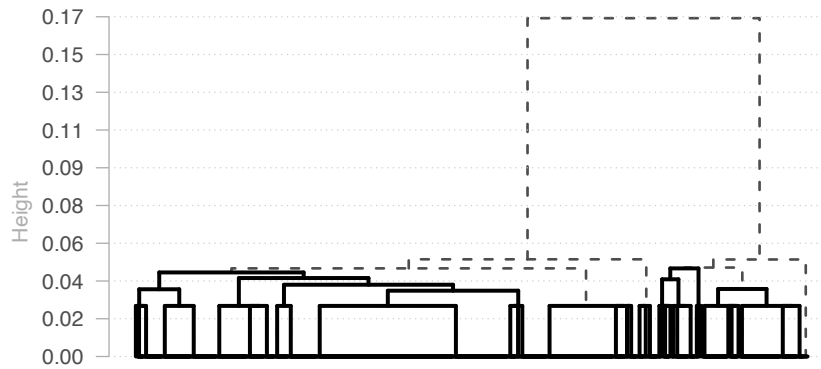
Hence, we can visualize the relationship between variable *Mjob* and *address*. Three contingency tables were calculated, and the new order of values of given variables was obtained. Contingency tables were re-arranged based on the new order and then normalized to range from 1 to 6. Lastly, Figure 11 was created.

The plot re-arranges the nominal values based on frequencies, so it is easier to spot any pattern in the data. For example, we can clearly see, that there are main families living in rural area in the third cluster, while urban area is occupied by families from the first and second cluster. Mothers working in health

care and in education are dominating in the second cluster. Mothers working in services or elsewhere are more prominent in the third cluster. The same interpretation would be harder to spot if we did not perform any sorting; see Figure 12 for comparison.

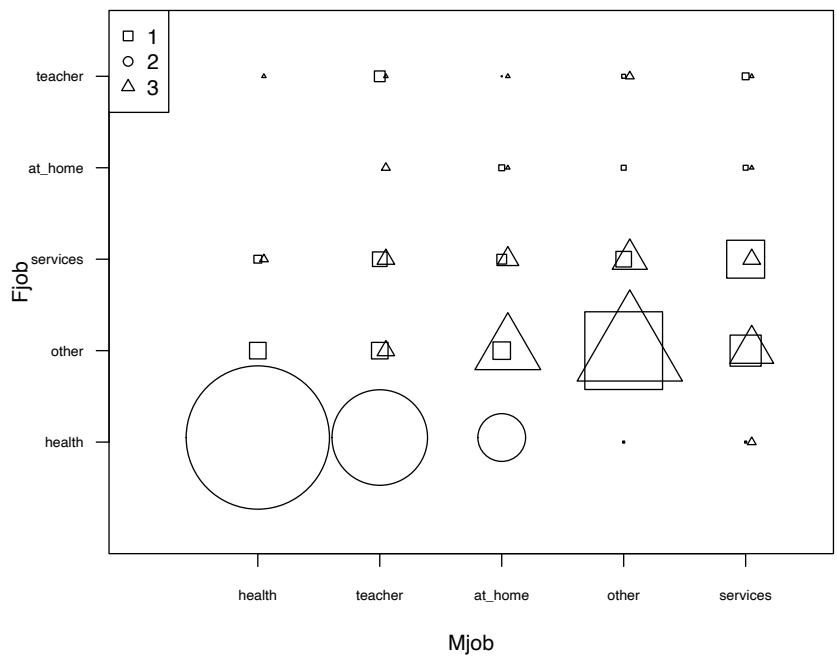
While a dendrogram illustrates a process of HCA, agenesis of clusters, and closeness of clustered objects, it does not provide any information about clusters' internal composition. However, the newly

Figure 13 Dendrogram of HCA of Mjob and Fjob



Source: Authors

Figure 14 Alternative of cluster scatter plot for clusters of Mjob and Fjob



Source: Authors

proposed alternative of cluster scatter plot provides such inside, and thus it is a useful complementary tool to dendrogram. This may be illustrated in the case of clustering objects in the aforementioned dataset from Cortez and Silva (2008). Based on the dendrogram in Figure 13, we can see how three clusters were formed. Figure 11 and Figure 13 show the inner structure of these clusters. Figure 11 shows the relationship between address and mother's job, while Figure 14 shows the relationship between mother's and father's job, with respect to the prevalence of category combinations within clusters. Due to the new visualization method, the patterns within data are visible. Hence it is obvious that the first and second cluster is formed by students who live in an urban area, and the third cluster is formed by students living in rural areas. There are almost no fathers who work as teachers or who stay at home, and very few of them work in civil services. Civil services or other areas are the most dominant for mothers in the first and third clusters. Stay-at-home mothers are the most prominent in the third cluster with students from rural areas, and in the second cluster with students living in an urban area but with a father who works in healthcare. Not all the mothers of students from the second cluster stay at home; they usually work in healthcare or as teachers.

CONCLUSION

Data visualization is vital in the final step of data mining applications. However, clustering and visualization of nominal data have not been explored to such an extent as clustering and visualization of the quantitative data has been. This paper provided an illustrative overview of available visualization methods in the area of cluster analysis with a focus on the visualization of nominal data. Two methods for sorting nominal data in order to improve visualization were also briefly presented. The goal of this paper was to provide an overview of existing methods but also to identify opportunities to improve the clustering visualization on nominal data. We proposed a new and very simple method to visualize the relationship between two nominal variables and assignation of observations into clusters as an alternative of cluster scatter plot for nominal data.

ACKNOWLEDGMENT

This work was supported by the Prague University of Economics and Business under Grant IGA F4/22/2021.

References

ANDERSON, E. (1935). The Irises of The Gaspe Peninsula [online]. *Bulletin of the American Iris Society*, 59: 2–5. <https://doi.org/10.2307/2394164>.

ANDREWS, D. (1972). Plots of high-dimensional data [online]. *Biometrics*, 28(1): 125–136. <https://doi.org/10.2307/2528964>.

CHANG, C., DING, Z. (2005). Categorical Data Visualization and Clustering Using Subjective Factors [online]. *Data & Knowledge Engineering*, 53(3): 243–262. <https://doi.org/10.1007/978-3-540-30076-2_23>.

CORTEZ, P., SILVA, A. (2008). Using Data Mining to Predict Secondary School Student Performance. *FUture BUsiness TEChnology Conference*, 5: 5–12.

ESKIN, E. et al. (2002). *A Geometric Framework for Unsupervised Anomaly Detection* [online]. Boston: Springer, 77–101. <https://doi.org/10.7916/D8D50TQT>.

FISHER, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems [online]. *Annals of Eugenics.* 7(2): 179–188. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>.

HOFMANN, T., BUHMANN, J. (1995). Multidimensional Scaling and Data Clustering [online]. *The MIT Press: Advances in Neural Information Processing Systems*, 7: 459–466. <https://doi.org/10.5555/2998687.2998744>.

ITOH, T. et al. (2004). Hierarchical data visualization using a fast rectangle-packing algorithm [online]. *IEEE Transactions on Visualization and Computer Graphics*, 10(3): 302–313. <https://doi.org/10.1109/TVCG.2004.1272729>.

KIM, S. et al. (2000). Interactive Visualization of Hierarchical Clusters Using MDS and MST [online]. *Metrika*, 51(1): 39–51. <https://doi.org/10.1007/s001840000043>.

KRUSKAL, J. B., LANDWEHR, J. M. (1983). Icicle Plots: Better Displays for Hierarchical Clustering [online]. *The American Statistician*. 37(2): 162–168. <https://doi.org/10.1080/00031305.1983.10482733>.

MA, S., HELLERSTEIN, J. L. (1999). Ordering Categorical Data to Improve Visualization. *IEEE Symposium on Information Visualization*, 15–18.

POLZLBAUER, G. et al. (2006). Advanced visualization of self-organizing maps with vector fields. *Neural Networks*, 19(6–7): 911–922.

R CORE TEAM (2021). *R: a Language and Environment for Statistical Computing* [online]. Vienna, Austria. <http://www.R-project.org>.

RAHLF, T. (2019). *Data Visualisation with R: 111 Examples* [online]. 2nd Ed. Cham: Springer. <https://doi.org/10.1007/978-3-030-28444-2>.

ROSARIO, G. E. et al. (2004). Mapping Nominal Values to Numbers for Effective Visualization [online]. *Information Visualization*. 3(2): 80–95. <https://doi.org/10.1057/palgrave.ivs.9500072>.

SIBSON, R. (1973). SLINK: an Optimally Efficient Algorithm for the Single Link Cluster Method [online]. *The Computer Journal*. 16(1): 30–34. <https://doi.org/10.1093/comjnl/16.1.30>.

SNOW, G. (2020). *TeachingDemos: Demonstrations for Teaching and Learning* [online]. R package Version 2.12. <https://CRAN.R-project.org/package=TeachingDemos>.

ŠULC, Z. et al. (2021). *Nomclust: an R package for hierarchical clustering of objects characterized by nominal variables*. Version 2.5.0.

VESANTO, J. (1999). SOM-based Data Visualization Methods [online]. *Intelligent Data Analysis*, 3(2):111–126. <https://doi.org/10.1016/S1088-467X(99)00013-X>.

WITTEN, I. H. et al. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd Ed. Burlington: Morgan Kaufmann publications.

Use of Markov Chain Simulation in Long Term Care Insurance

Vladimír Mucha[1] | *University of Economics in Bratislava, Bratislava, Slovakia*
Ivana Faybíková[2] | *University of Economics in Bratislava, Bratislava, Slovakia*
Ingrid Krčová[3] | *University of Economics in Bratislava, Bratislava, Slovakia*

Received 13.4.2022, **Accepted** (reviewed) 27.6.2022, **Published** 16.12.2022

Abstract

The aim of this paper is to present the use of simulations of non-homogeneous Markov chains in discrete time in the context of the problem of long-term care delivery. The object of investigation is to model the distribution of clients into different states during specified time steps, then to estimate the average time a client stays in a given state, as well as to estimate the insurance premiums. Within the use of the Monte Carlo simulation method, the focus is on providing approaches that ensure more accurate results in the context of the number of simulations performed. Based on the statistical processing of the data obtained from the simulations, it is possible to obtain the information necessary for the provision of resources for the provision of health care and for the determination of the aforementioned premiums. For the implementation of the above techniques and their graphical presentation available packages such as markovchain, ggplot2 or custom code created using the R language were used.

Keywords	***DOI***	***JEL code***
Long-term care insurance, Markov Chains, multi-state models, simulations, Monte Carlo Method, markovchain package	*https://doi.org/10.54694/stat.2022.20*	*C63, G22*

INTRODUCTION

We currently see an increase in average life expectancy and we can assume that this trend will continue in the future. It is the older age group that suffers from various chronic illnesses or physical limitations, and it is the older age group that makes the most use of the services of healthcare providers. For this reason, health care institutions pay considerable attention to estimating the number of clients who will need health care later on. They focus particularly on the issue of Long-Term Care (LTC), which is provided to people who have reached a state of non-self-sufficiency. The increased number of people who become incapacitated due to illness also represents an increase in health care costs. Looking at the other side of the issue, people are also thinking about the capital that they would have available in the event that they need long-term care. Without long-term care insurance, the cost of providing these

[1] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: vladimir.mucha@euba.sk.
[2] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: ivana.faybikova@euba.sk.
[3] Department of Mathematics and Actuarial Science, Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: ingrid.krcova@euba.sk.

services can quickly deplete an individual's or family's savings. Markov chains are a popular tool used to make estimations as regards to the incidence of critical illnesses and the provision of long-term care. The issue of using these random processes to model the evolution of different illnesses in the context of multi-state models is currently being addressed by many authors. They use various software support to implement them, one suitable possibility being the R language with its available packages. For example, the available R language package markovchain (Spedicato, 2017) can be used to create Markov chain objects, the implementation of probabilistic operations using them, statistical analysis and simulation of homogeneous and non-homogeneous Markov chains in discrete time. Another tool for dealing with Markov chains in discrete time is the DTMCPack package (Nicholson, 2013). Functions to implement Markov Chain Monte Carlo (MCMC) using the Metropolis algorithm, for example, are contained in the package mcmc (Geyer and Johnson, 2013). The open-source software MARCH, which is a set of functions from the MATLAB programming environment (Berchtold, 2001), is also available to model Markov chains in discrete time. If necessary, custom code can be developed in the above programming languages to simulate Markov chains in discrete and continuous time according to algorithms available in various publications, e.g. (Janková, et al., 2014: 85–87). As mentioned before Markov chains are also used in the context of multi-state models for modelling in the field of life and non-life insurance, e.g. for estimating the costs associated with the provision of health care, estimating premiums, predicting the evolution of various illnesses, as well as for modelling the number of insured lives in a bonus -malus system in the framework of compulsory car insurance. The problem of planning financial resources for health care is presented by Garg et al. (2010), using non-homogeneous Markov chains in discrete time to model the number of patients, as well as by Diz and Query (2012). Methods for estimating transition probabilities or transition intensities, the use of Markov Chain Monte Carlo simulation and modelling with Markov Chains also in continuous time in the field of long-term care are discussed by other authors such as (Sato and Zouain, 2010; Esquível et al., 2021; Fleischmann, Hirz, Sirianni, 2021; Xie, Chaussalet, Millard, 2005). Modelling with Markov chains also allows estimation of the expected or average stay time of an individual in the healthy and sick states, respectively (Dudel and Myrskylä, 2020). Another area of interest in the implementation of Markov chains in critical illness modelling in the context of healthcare is critical illness insurance. A lump sum benefit will be paid to the insured in case of a critical illness diagnosis. The above issue is presented by Pasaribu et al. (2019), using the continuous-time Markov chain apparatus to estimate the premiums for specified age categories of insureds. Many authors use Markov chains to predict the evolution of various infectious diseases (Li, Dushoff and Bolker, 2018; Twumasi, Asiedu and Nortey, 2019). An alternative stochastic modeling approach that can be implemented in this area is represented by Hawkes processes (Maciak, Okhrin and Pešta, 2021; Unwin et al., 2021). In non-life insurance, homogeneous Markov chains in discrete time are used to model the distribution of the number of policyholders in a bonus-malus system (Fernandez-Morales, 2015). The present paper focuses on the simulation of trajectories of non-homogeneous Markov chains in discrete time using the R language. Based on the processing of the generated data, we will analyze the modelling of the distribution of the number of clients in each state during the specified years, as well as the estimation of the average time a client stays in a given state, and the estimation of the premiums in the case of long-term care insurance. To present the above techniques, we have used data obtained from the markovchain package, which were for the male population in Italy. In the three-state model, the sick state represents the state of unfitness into which the client has fallen due to, for example, Alzheimer's disease.

1 METHODS OF ANALYSIS

If the insurer has real data on the health status of insured lives, it can obtain transition probabilities between the different states, which can be used to model the evolution of the insured's state over

the analysed time periods. Since these probabilities depend on the age of the insured in our dataset, we use non-homogeneous Markov chains and their simulations for this purpose.

1.1 Non-homogenous Markov chains in discrete time

A random chain $\{X_t\}_{t\in T}$ is called a Markov chain, if for each $h = 0, 1, 2, ...$, for all times $0 \le t_0 \le t_1 ...$, $t_h \le t_{h+1}$, $t_0, t_1 ..., t_h, t_{h+1} \in T$ and for all states $s \in S$ we have

$$P(X_{t_{h+1}} = s_{t_{h+1}} \mid X_{t_h} = s_{t_h}, \dots, X_{t_1} = s_{t_1}, X_{t_0} = s_{t_0}) = P(X_{t_{h+1}} = s_{t_{h+1}} \mid X_{t_h} = s_{t_h}), \quad (1)$$

assuming that the random variable $X_{t_{h+1}}$ is independent from $X_{t_0}, X_{t_1}, \dots, X_{t_h}$ (Janková, et al., 2014).

This means that in the case of Markov chains, the probability of transition to the next state depends only on the current state and not on previous states, hence they are also called "memoryless" chains (Dobrow, 2016). We consider Markov chains in discrete time, so T is the set of natural numbers with zero. The values taken by the random variables X_t, $t \in T$, are called states, we denote their set by $S = \{s_1, s_2, \dots s_m\}$. We call the Markov chain $\{X_t\}_{t\in T}$ *non-homogenous* (in time), unless we have that as follows:

$$\forall i, j \in S, \forall k \in N: P(X_{t+k+1} = j \mid X_{t+k} = i) = P(X_{t+1} = j \mid X_t = i). \quad (2)$$

Transition probabilities from state i to state j after one time step from the time t we denote by

$$p_{i,j}(t, t+1) = P(X_{t+1} = j \mid X_t = i), \quad (3)$$

and arrange them for a given t into the transition probability matrix

$$P(t; t+1) = \left(p_{i,j}(t; t+1)\right)_{i,j\in S}, \quad (4)$$

for which we have $\sum_{j\in S} p_{i,j}(t; t+1) = 1$, which means, that each row of this matrix is a probability distribution, we call it a stochastic matrix (Jones and Smith, 2018).

Let $\{X_t\}_{t\ge 0}$ be a Markov chain. The probability distribution $\alpha = \{\alpha_k\}_{k\in S}$ such that $P(X_{t_0} = s_k) = \alpha_k = p_{s_k}(0)$ for $s_k \in S$, we call *the initial distribution of the chain* $\{X_t\}_{t\ge 0}$.

The vector $\mathbf{p}(0) = (p_{s_1}(0); \dots ; p_{s_k}(0); \dots ; p_{s_m}(0))$ will be called the *vector of initial probabilities*. The probability of transition from the initial state k to state j, $j \in S$ in h time steps from time 0, i.e. from the beginning of the Markov chain, is called *the absolute probability of the states of the Markov chain* and is denoted as follows

$$p_{k,j}(0,h) = p_j^{(k)}(h), \quad (5)$$

whereby we will call the vector $\mathbf{p}^{(k)}(h) = (p_j^{(k)}(h))_{j\in S}$ *the vector of absolute probabilities*.

We get the following expression for the vector of absolute probabilities $\mathbf{p}^{(k)}(h)$ using the *Chapman-Kolmogorov equality* (Fecenko, 2018).

$$\mathbf{p}^{(k)}(h) = \mathbf{p}^{(k)}(h-1) \cdot P(h-1; h) = \mathbf{p}(0) \cdot \dots \cdot P(h-1; h). \quad (6)$$

1.2 Generating trajectories of a non-homogeneous Markov chain in discrete time in R

For a non-homogenous Markov chain with transition matrices $P(t; t+1) = (p_{ij}(t; t+1))_{i,j\in S}$ for $t \in T$, we define a random variable Z_r.

The values of the probability function $P(Z_r = j)$ represent in the corresponding matrix $P(t; t + 1)$ the values $p_{s_{ij}}(t; t + 1)$, $j \in S = \{s_1, s_2, \dots s_m\}$, which appear in its r-th row. We write this discrete distribution using the notation $p_{Z_r}(j) = \{p_{s_{ij}}(t; t + 1)\}_{j \in S}$. The algorithm for generating the random variable values Z_r by the inverse transformation method can be written as follows in the given context:

1. generate the value u of random variable $U \sim Unif(0; 1)$
2. transform the value u to the value of the random variable Z_r as follows

$$Z_r = s_1, \text{ if } u \le p_{Z_r}(s_1) \quad \text{or} \quad Z_r = j, \text{ if } \sum_{l=s_1}^{j-1} p_{Z_r}(l) < u \le \sum_{l=s_1}^{j} p_{Z_r}(l) . \tag{7}$$

We will simulate a non-homogeneous Markov chain with transition matrices $P(t; t + 1)$ and initial distribution $\alpha=\{\alpha_k\}_{k \in S}$ on a set of states $S = \{s_1, s_1, \dots s_m\}$ in discrete time, i.e. construct its trajectory, until time t_h using the following steps:

1. From the discrete initial distribution $\{\alpha_k\}_{k \in S}$ we generate the value $s_{t_0} = s_k$ of the random variable X_{t_0} at the initial point of time.
2. From the discrete distribution $\{p_{s_{ij}}(t_0;t_1)\}_{j \in S}$, i.e. from the k-th row of the transition matrix $P(t_0; t_1)$, we generate the value s_{t_1}, which represents a value of the random variable X_{t_1}.
3. If $t_c < t_h$ and we have generated the value of the random variable X_{t_c}, then from the distribution. $\{p_{s_{i,j}}(t_c; t_{c+1})\}_{j \in S}$, i.e., from the row corresponding to the state s_{t_c} in the transition matrix $P(t_c; t_{c+1})$, we generate the value $s_{t_{c+1}}$, which represents the value of the random variable $X_{t_{c+1}}$.

If $t = t_h$, we stop the generation. The result will be the realisation of a set of h states $s_{t_0}, s_{t_1}, \dots , s_{t_h}$, which we get after h time steps. By repeating this algorithm n times, we get n trajectories of the Markov non-homogeneous chain in discrete time (Janková et al., 2014).

1.3 Accuracy of Monte Carlo estimation of the probability of an event occurring

To estimate the probability of occurrence of an event we use the law of large numbers, or Bernoulli's theorem, according to which as the number n of repeated independent trials increases, the relative frequency of occurrence of the observed event f_n approaches the theoretical probability p of occurrence of this event in each trial, which we can express as:

$$\lim_{n \to \infty} P(|f_n - p| < \varepsilon) = 1, \varepsilon > 0. \tag{8}$$

Thus, the number of occurrences of the observed event in a series of n independent simulation steps follows a binomial distribution $Y_n \sim B_i(n; p)$ with characteristics $E(Y_n) = n \cdot p$ and $D(Y_n) = n \cdot p \cdot q$. Using the Moivre – Laplace theorem we get:

$$P\left(\left|\frac{Y_n}{n} - p\right| < \varepsilon\right) \approx 2 \cdot \Phi\left(\varepsilon \cdot \sqrt{\frac{n}{p \cdot q}}\right) - 1, \tag{9}$$

and hence we can determine with probability $(1 - \alpha)$ the accuracy of the theoretical probability estimate p using the relative frequency $f_n = \frac{Y_n}{n}$ (Mucha and Páleš, 2018) by means of the confidence interval $(p - \varepsilon; p + \varepsilon)$, for which:

$$\frac{Y_n}{n} \in \left(p - u_{1-\frac{\alpha}{2}} \cdot \sigma; p + u_{1-\frac{\alpha}{2}} \cdot \sigma\right), \text{where } \sigma = \sqrt{D\left(\frac{Y_n}{n}\right)} = \sqrt{\frac{p \cdot q}{n}} . \tag{10}$$

The accuracy, or error, ε therefore depends on the chosen level of confidence $(1 - \alpha)$ and from the standard deviation, the value of which can be bounded by the expression (Horáková and Mucha, 2002).

$$\sigma = \sqrt{\frac{p \cdot q}{n}} \leq \frac{1}{2} \cdot \sqrt{\frac{1}{n}} \ . \tag{11}$$

We can thus estimate more generally the maximum deviations of the simulated values $f_n = \frac{Y_n}{n}$ from the theoretical probability p for a given number of simulations from the equation:

$$\varepsilon = u_{1-\frac{\alpha}{2}} \cdot \frac{1}{2} \cdot \sqrt{\frac{1}{n}} \ . \tag{12}$$

Table 1 gives the calculated maximum errors ε for probability $1 - \alpha = 0.9$ and for different numbers of simulations n.

Table 1 Accuracy of the probability estimate p for a given number of simulations n with confidence $1 - \alpha = 0.9$

n	$\varepsilon_{0.9}$
1 000	0.0260
10 000	0.0082
100 000	0.0026

Source: Own construction

It should be noted that if the theoretical probability is close to $p = 0.5$, for a given number of simulations, the estimation error would be close to the values given in Table 1. Therefore, to obtain more accurate results, it is advisable to perform the order of tens or hundreds of thousands of simulations when estimating the probability of an event using relative frequency.

2 DATA DESCRIPTION AND MODEL BUILDING

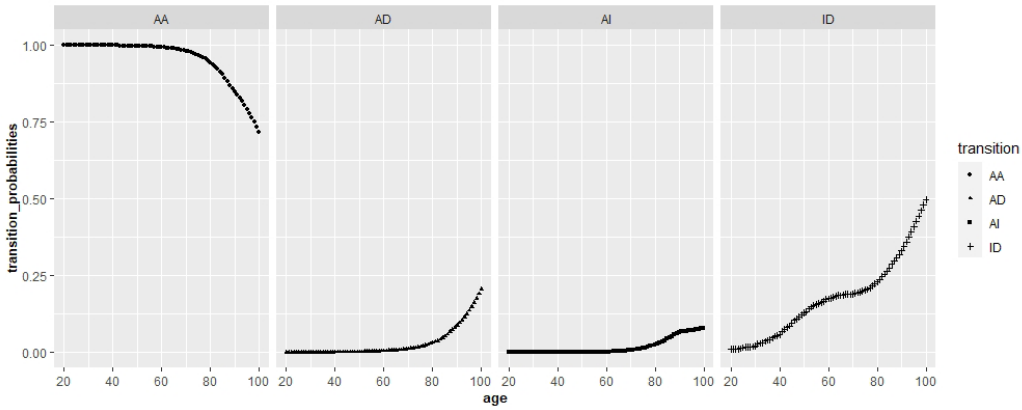
We will use a multi-state model to solve the problem and focus on a unidirectional model with three states: healthy/active A, (terminally) ill I and dead D. From the healthy state it is possible to transition to the ill state and to the dead state. After leaving the healthy state, it is not possible to return to it again. From the ill state it is only possible to transition to the absorbing dead state. We can use this to model the situation of an illness for which there is no cure. This is also called the permanent disability model (Škrovánková and Simonka, 2021).

2.1 Data description

The dataset that we use to present the possibilities of using discrete-time simulation of non-homogeneous Markov chains in long-term care insurance was obtained in a text file from the *Markovchain* package, available in R. These data, in the form of transition probabilities between states depending on the age of the insured, were obtained from *Assicurazioni Sulla Salute*: *Caratteristiche, modelli attuariali e basi tecniche* by Paolo de Angelis and Luigi di Falco (2016). The data presented refer to the male population in Italy, whereby the status of ill I is considered, according to the author of the mentioned package (Spedicato, 2017), as a disability leading to the insured life's incapacity to work, for example Alzheimer's disease. We display graphically the obtained transition probabilities in Figure 1 using the *ggplot2* package in the R language environment (Wickham, 2016). This allows us to visually analyse the individual

transition probabilities depending on the age of the insured. We have plotted their values for the age interval from 20 to 100 years.

Figure 1 Transition probabilities $p_{i,j}(t, t+1)$, $i,j \in \{A, I, D\}$ by age t of males in Italy



Source: Own construction, customized in R

2.2 Model building

We will consider the model as a system of generated trajectories of non-homogeneous Markov chains in discrete time, from which we obtain the desired results based on their statistical processing. Since the algorithm for simulating Markov chains uses transition matrices, we created them in the context of the rules of the given three-state model using the *Markovchain* package (Spedicato et al., 2017). We display the transition probability matrix in general for age t of the insured life:

$$P(t; t+1) = \begin{pmatrix} p_{A;A}(t; t+1) & p_{A;I}(t; t+1) & p_{A;D}(t; t+1) \\ 0 & p_{I;I}(t; t+1) & p_{I;D}(t; t+1) \\ 0 & 0 & 1 \end{pmatrix}. \tag{13}$$

In this way, we have modified the original data into the format of individual transition matrices $P(t; t+1)$, which we will use to create the final model for solving the presented problem. By simulating non-homogeneous Markov chains, we will create a model through the generation of their trajectories, which will mimic the real evolution of the states of the insured during h time steps. The results can be written into n rows and h columns of the matrix ${}^{(z)}M = [m_{ij}]_{n \times h}$, $z \in \{A, I\}$, where z represents the initial status of the insured life. For practical reasons, we will consider only the initial states healthy and ill. The elements of this matrix will be of interest to us in the context of carrying out analyses in the area of long-term care insurance.

Based on the statistical processing of a sufficient amount of n generated data in the h-th column of the matrix ${}^{(z)}M$ it is possible to estimate the percentage distribution of insured lives in each state $g \in \{A, I, D\}$ after h time steps according to the equation:

$$perc_g^{(z)}(h) = p_g^{(z)}(h) \cdot 100 \approx \frac{\sum_{i=1}^{n} I[m_{ih}=g]}{n} \cdot 100, g \in \{A, I, D\}, z \in \{A, I\}, \tag{14}$$

where $p_g^{(z)}(h)$ represents a particular component of the absolute probability vector, which we estimate from the generated values in the h-th column of the matrix ${}^{(z)}M$.

In the presented model, we consider a portfolio composed of K insured lives, where we denote the initial number of insured lives in the healthy state by K_A and the number of insured lives in the ill state by K_I, thus:

$$K = K_A + K_I \,. \tag{15}$$

The absolute distribution of the number of insured in each state after h time steps can be written in the form of a vector $\mathbf{k}$, which is a linear combination of absolute probability vectors $\mathbf{p}^{(A)}(h)$ and $\mathbf{p}^{(I)}(h)$, which we write as:

$$\mathbf{k} = K_A \cdot \mathbf{p}^{(A)}(h) + K_I \cdot \mathbf{p}^{(I)}(h) \,. \tag{16}$$

By substituting the mentioned vectors into the equation for vector $\mathbf{k}$, we get the following expression in the considered three-state model:

$$\mathbf{k} = (k_1; k_2; k_3) = (K_A \cdot p_A{}^{(A)}(h); K_A \cdot p_I{}^{(A)}(h) + K_I \cdot p_I{}^{(I)}(h); K_A \cdot p_D{}^{(A)}(h) + K_I \cdot p_D{}^{(I)}(h)), \tag{17}$$

where k_1 represents the number of insured lives in the healthy state, k_2 represents the number of insured lives in the ill state and k_3 the number of insured lives in the dead state in the considered portfolio after h time steps. The above distribution of the number of policyholders into the different states makes sense, given Bernoulli's law of large numbers, if the numbers of policyholders K_A and K_I are large enough, i.e., in the order of tens of thousands or hundreds of thousands. We estimate the individual probabilities $p_A{}^{(A)}(h)$, $p_I{}^{(A)}(h)$, $p_I{}^{(I)}(h)$, $p_D{}^{(A)}(h)$, $p_D{}^{(I)}(h)$ using the relative frequencies from the generated matrices ${}^{(z)}M = [m_{ij}]_{n\times h}$, $z \in \{A, I\}$.

By generating the trajectories of non-homogeneous Markov chains, it is also possible to determine the percentage distribution of the number of insured lives K into the different states after h time steps, which we write using the vector:

$$\mathbf{perc} = (perc_1; perc_2; perc_3), \tag{18}$$

where $perc_1$ represents the percentage of insured lives in the healthy state, $perc_2$ represents the percentage of insured lives in the ill state and $perc_3$ the number of insured lives in the dead state in the considered portfolio after h time steps.

To determine the individual components $perc_i$, $i = 1, 2, 3$ in the considered three-state model we used a weighted arithmetic average with weights $w_1 = K_A$, $w_2 = K_I$, whereby:

$$perc_1 = \frac{K_A}{K} \cdot perc_A{}^{(A)}(h) + \frac{K_I}{K} \cdot perc_A{}^{(I)}(h) = \frac{K_A}{K} \cdot perc_A{}^{(A)}(h) \,, \tag{19}$$

$$perc_2 = \frac{K_A}{K} \cdot perc_I{}^{(A)}(h) + \frac{K_I}{K} \cdot perc_I{}^{(I)}(h) \,, \tag{20}$$

$$perc_3 = \frac{K_A}{K} \cdot perc_D{}^{(A)}(h) + \frac{K_I}{K} \cdot perc_D{}^{(I)}(h) \,, \tag{21}$$

If we do not consider a specific portfolio of insured lives, but the population in general, the above condition of a sufficiently large number of K_A and K_I is automatically satisfied and the predicted absolute and percentage distributions can be considered relevant without verification.

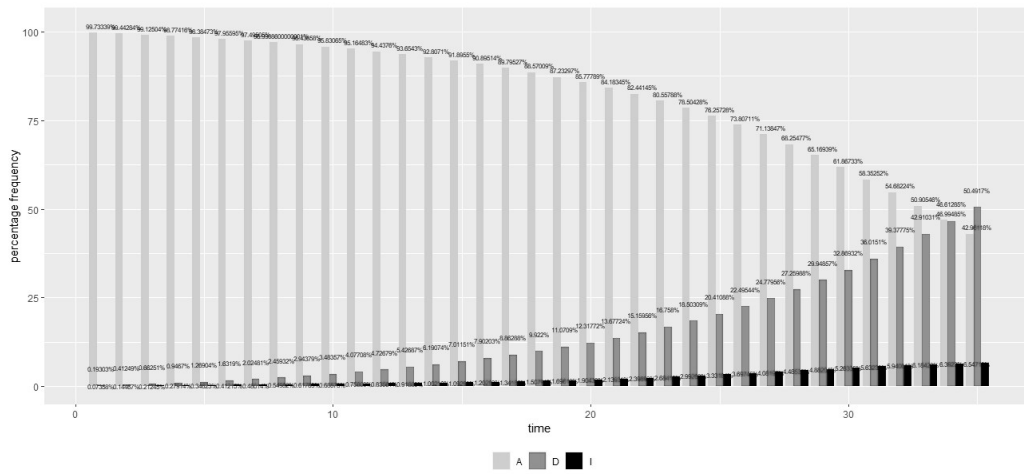
3 RESULTS AND DISCUSSION

In this part of the paper, we will use the described data set to model and analyse the development for a particular critical illness (for example Alzheimer's disease), which requires long-term care in the event of its occurrence. We will use the simulation of the trajectories of non-homogeneous Markov chains, which we will implement using the R language. Using the statistical data in the generated matrices ${}^{(z)}M = [m_{ij}]_{n\times h}$, $z \in \{A, I\}$, the insurance company can obtain information necessary for the provision of health care and for premium calculation.

3.1 Estimation of the distribution of insured lives in the separate states

Due to the nature of the database, we will focus on predicting the percentage distribution of the number of insured lives in the separate states on a yearly basis for a certain number of years. If the data were recorded differently, for example monthly, we could use that as our time interval for modelling purposes. First, we will show the evolution of the percentage distribution of the number of initially healthy insured lives aged 50, which we illustrate graphically in Figure 2 using the R language package *ggplot2* (Wickham, 2016).

Figure 2 Percentage distribution of initially healthy lives aged 50 in the different states A, I, D over time



Source: Own construction, customized in R

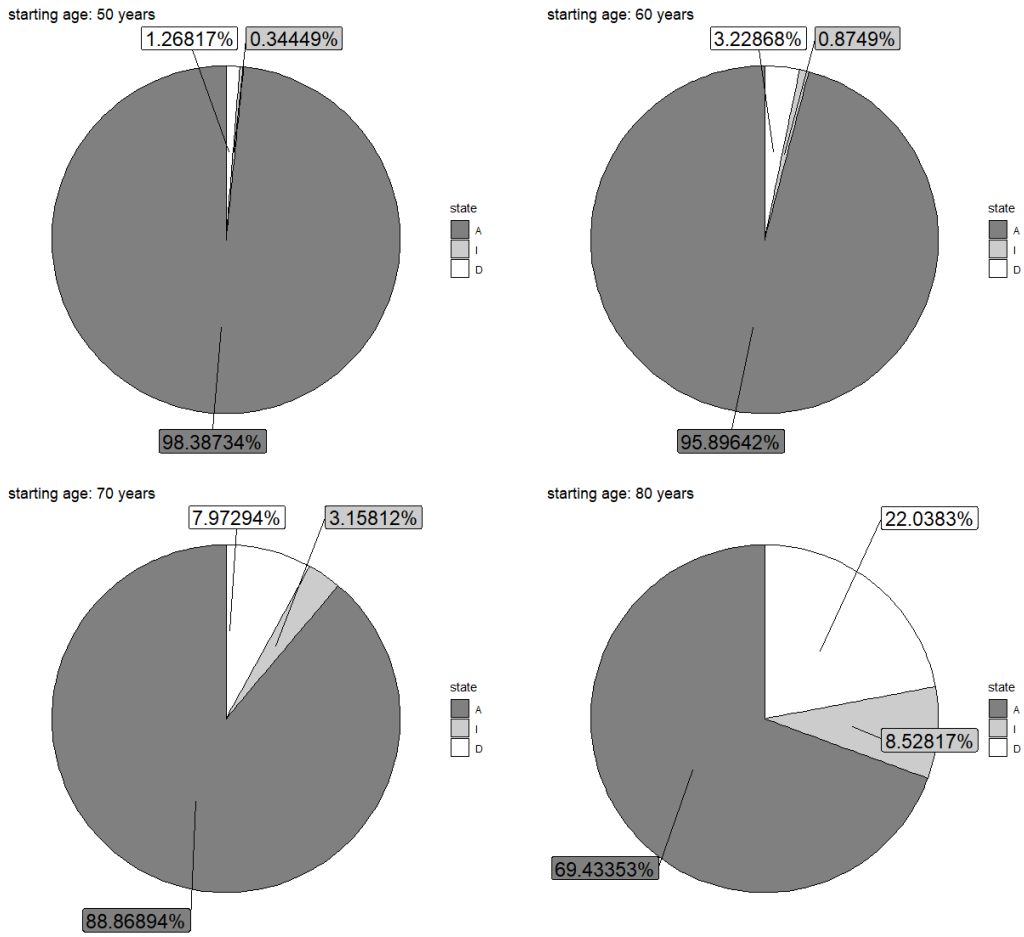
The graphical presentation above shows the trends in the percentages of the separate states over time. To achieve more accurate results, we have carried out $n = 100\,000$ simulations, whereby we repeated this scenario using the R language 100 times and calculated the results of the percentage distribution in each year as the arithmetic average.

From the presented results we see that, for example, after 20 years 85.77789% of the initially healthy insured lives aged 50 will still be in the healthy state, 1.90439% will be in the ill state and 12.31772% will be in the dead state. Of course, the above statement is in general only true if the initial portfolio of healthy males aged 50 was sufficiently large, i.e., in terms of the law of large numbers, it consists of the order of a few tens of thousands or hundreds of thousands of lives. In the case of the male population in Italy, this condition is of course met.

Figure 3 shows graphically the percentage distribution after 5 years for males initially aged 50, 60, 70 and 80 who were initially in the healthy state. In the case of males aged 70 and 80 who were initially healthy, there is a significant increase in the number in the ill state (illness requiring long-term care)

after 5 years, compared to males aged 50 and 60. Given the nature of the illness (for example Alzheimer's disease), this increase is to be expected.

Figure 3 Distribution of the number of insured lives after 5 years as a percentage for males initially aged 50, 60,70 and 80 who started in the healthy state after running 100 000 simulations



Source: Own construction, customized in R

So far, we have been modelling assuming that the insured lives were in the healthy state at the start. We will now model the evolution of the number of insured lives for a specific portfolio that is composed of K_A in the healthy state and K_I in the ill state. Based on the simulation trajectories, we present in Tables 2 and 3 the distribution of the absolute and percentage number of insured lives initially aged 50 during a period of 10 years, where $K_A = 300\ 000$ and $K_I = 20\ 000$.

When expressing the number of insured lives as a percentage, based on the data generated in the matrices ${}^{(z)}M = [m_{ij}]_{n\times h}$, $z \in \{A, I\}$, it is not necessary to specify the absolute number of policyholders K_A and K_I. It is enough to enter the relative or percentage frequency of the considered states $\{A, I\}$ at the start of modelling.

Table 2 Distribution of the number of insured lives during 10 years for males aged 50, of which at the beginning $K_A = 300\ 000$ were in the healthy state and $K_I = 20\ 000$ in the ill state

	State			
Time	**healthy**	**ill**	**dead**	**Total**
1	299 211	17 649	3 140	320 000
2	298 339	15 551	6 110	320 000
3	297 386	13 628	8 986	320 000
4	296 329	11 900	11 771	320 000
5	295 166	10 407	14 427	320 000
6	293 895	9 140	16 965	320 000
7	292 519	8 072	19 409	320 000
8	291 014	7 185	21 801	320 000
9	289 364	6 460	24 176	320 000
10	287 541	5 871	26 588	320 000

Source: Own construction

Table 3 Percentage distribution of the number of insured lives during 10 years for males aged 50 of which $K_A = 300\ 000$ were initially in the healthy state and $K_I = 20\ 000$ in the ill state

	State			
Time	**healthy**	**ill**	**dead**	**Total**
1	93.503%	5.515%	0.982%	100%
2	93.231%	4.860%	1.909%	100%
3	92.933%	4.259%	2.808%	100%
4	92.603%	3.719%	3.678%	100%
5	92.239%	3.252%	4.509%	100%
6	91.842%	2.856%	5.302%	100%
7	91.412%	2.523%	6.065%	100%
8	90.942%	2.245%	6.813%	100%
9	90.426%	2.019%	7.555%	100%
10	89.857%	1.834%	8.309%	100%

Source: Own construction

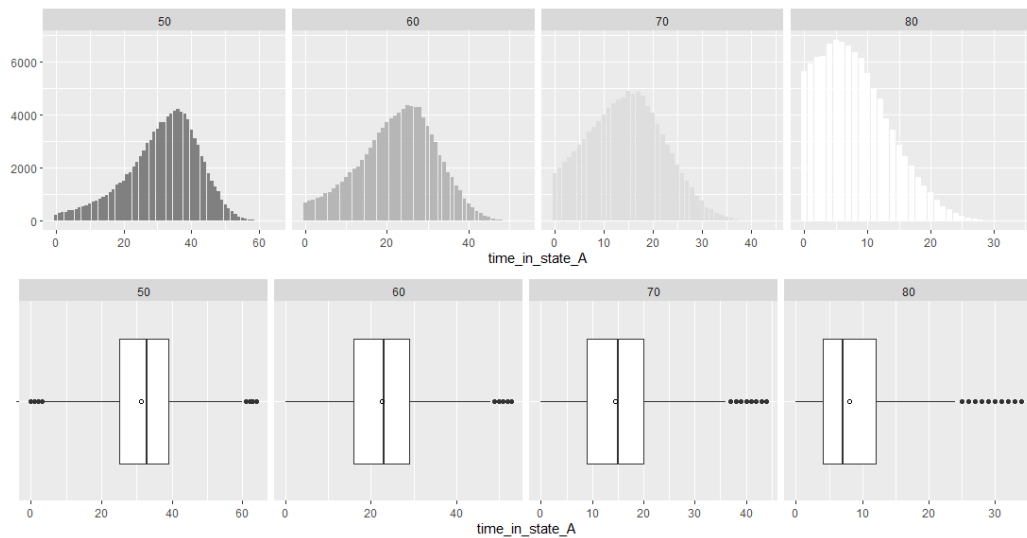
By using the simulation of non-homogeneous Markov chains in discrete time, it is possible to estimate the evolution of the representation in each of the separate states during the modelled years. By comparing the values obtained from the simulations with the values obtained from the absolute probability vectors, we can conclude that the presented simulation model provides relevant results for the described number of simulations. We will therefore use it to further model and obtain information that is relevant for the insurance of critical illnesses that require long-term care.

3.2 Estimation of time remaining healthy and remaining ill

In this part of the paper, we will analyse the estimation of the time during which the insured life remains in the healthy and ill state, respectively. We assume that the insured lives are healthy at ages 50, 60, 70,

and 80 at the beginning of the modelling period. Using 1 000 simulations of the trajectories of non-homogeneous Markov chains in the matrix ${}^{(A)}M = [m_{ij}]_{n\times h}$ we recorded data on the number of years the insured life remained in the healthy state. We will use the full range of available transition probability matrices and model the states up to age 120. We present these data for each age category in the form of a bar plot and box plot in Figure 4. The circle in the box plot denotes the estimated mean value of the number of years the insured life remained in the healthy state and the line in the box denotes the median value.

Figure 4 Analysis of the number of years the insured life remained healthy for the initial ages 50, 60, 70 and 80 using a bar plot and a box plot



Source: Own construction, customized in R

For the sake of illustration, we list the selected values in Table 4.

Table 4 Estimated values for the number of years the insured life has been in the healthy state

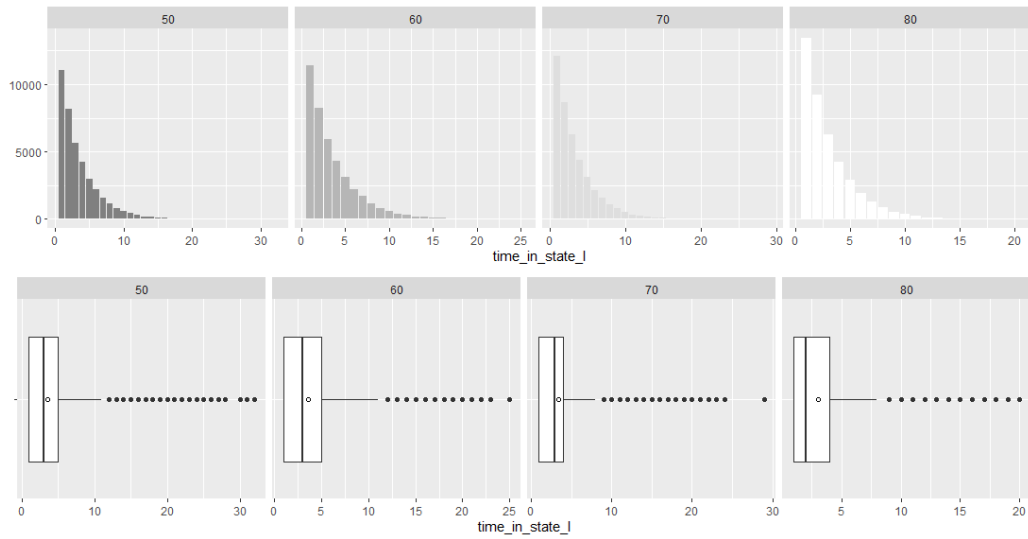
Age	$x_{0.25}$	Median	$x_{0.75}$	Mean
50	25	33	39	31.39801
60	16	23	29	22.53211
70	9	15	20	14.54286
80	4	7	12	8.10224

Source: Own construction

Thus, for example, in the case of healthy insured lives aged 60, 75% of them have the number of years they remain healthy less than or equal to 29, and 25% of them have the number of years they remain healthy greater than 29 years. The average number of years for a healthy insured life aged 50 is equal to 31.39801 years and for a 70 year old is equal to 14.54286 years. Given that the estimated mean value is an arithmetic mean, it is generally necessary to consider the dispersion of the values on the left and right sides of the mean. This may ultimately affect the relevance of the information thus obtained, despite a sufficiently large set of values. In this case, the median can be used for estimation.

Another important element for insurance calculations is the time during which the insured life remains ill. Again, we assume that the lives are healthy at ages 50, 60, 70 and 80 years at the start of the modelling period. Using 100 000 simulations of non-homogeneous Markov chains, the matrix ${}^{(A)}M = [m_{ij}]_{n\times h}$ records the data on the number of years the insured life remained ill. We present these data for each age category in the form of a bar plot and a box plot in Figure 5.

Figure 5 Analysis of the number of years during which the insured life remained ill for ages 50, 60, 70 and 80 using a bar plot and a box plot



Source: Own construction, customized in R

For the sake of illustration, we will list the selected values in Table 5.

Table 5 Estimated values for the number of years during which the insured life was in the ill state

Age	$x_{0.25}$	Median	$x_{0.75}$	Mean
50	1	3	5	3.599501
60	1	3	5	3.532215
70	1	3	4	3.365925
80	1	2	4	3.002761

Source: Own construction

Out of the 100 000 simulations in 59 754 cases an insured life aged 50 subsequently died whilst remaining in the healthy state. This means that, from the data available to us, he was not registered in the three-state model described above as an insured life in need of intensive long-term care because of illness. In the remaining 40 246 cases represented by the trajectory of the considered Markov chain we found that if an insured life entered the ill state, he stayed in this state 3.599501 years on average before moving to the dead state. This compares with a value of 3.365925 years for the 70 year old insured life as shown in Table 5.

3.3 Calculation of long-term care insurance premiums

Finally, we will deal with the determination of the single premium P, which a life aged x has to pay in order to receive an annual payment of C while in a state of non-self-sufficiency. We assume, of course, that the life is in the healthy state at the start of the policy. We use the generated trajectories of the non-homogeneous Markov chains, which we have written into the matrix ${}^{(A)}M = [m_{ij}]_{n\times h}$, to determine the above insurance premium, where $n = 100\,000$ and $h = 120 - x$. For the purpose of determining the premium, we transform all elements of this matrix indicating the ill state I to the amount C and its other elements to zero values. We denote the resulting matrix by $M^C = [m^C_{ij}]_{n\times h}$. The single premium P is then determined using the equation:

$$P = M^C \cdot U\,, \tag{22}$$

where for the matrix elements $U = [u_{ij}]_{h\times 1}$ it holds that $u_{ij} = (1+u)^{-i}$, where u is the annual rate of interest.

The individual elements of the matrix $P = [p_{ij}]_{n\times 1}$ can be interpreted as representing the given premium determined for a particular modelled scenario of the insured life represented by the corresponding trajectory of the non-homogeneous Markov chain. The arithmetic average was then used to calculate the single premium P, to be paid by the life aged x. For more accurate results, we repeated this scenario using R 1 000 times and for the premium P we again used the arithmetic average as the estimated value. For example, a healthy life aged 50 would have to pay a premium of $P = €12\,583.42$ at the interest rate used of $u = 0.01$, in order to receive an annual payment $C = €12\,000$ at the beginning of each year if he falls ill. For comparison, we have also calculated the premium using a standard life insurance formula

$$P = \sum_{t=1}^{\omega-x} {}_{t-1}p_x^{AA} \cdot q^{AI}_{x+t-1} \cdot v^t \cdot \pi(\ddot{a}^{(I)}_{x+t}),\ t = 1,\ 2,\ \ldots,\ \omega\,, \tag{23}$$

ω – the highest age in the relevant mortality table,
${}_{t-1}p_x^{AA}$ – the probability that a life x years old remains healthy for $t - 1$ years,
q^{AI}_{x+t-1} – the probability that a life aged $x + t - 1$ years in the healthy state becomes ill within one year, i.e., at age $x + t$ is in the ill state,
v – is the discount factor, i.e. $v = \dfrac{1}{1+u}$, where u is the annual interest rate,
$\pi(\ddot{a}^{(I)}_{x+t})$ – the whole life annuity-due for a life aged $x + t$ years, if he is then in the ill state, for an annual payment of C payable in advance, i.e.

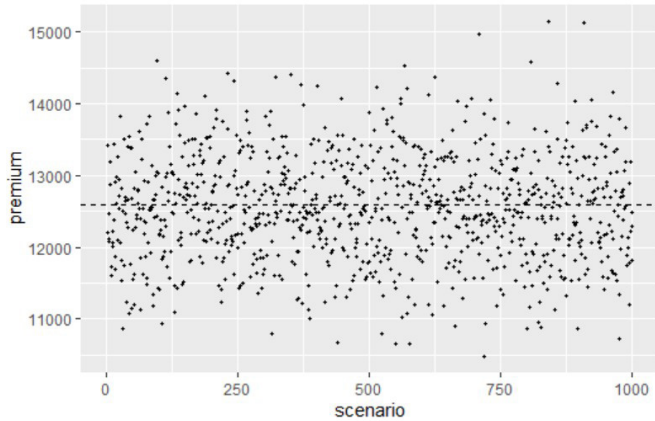
$$\pi(\ddot{a}^{(I)}_{x+t}) = \sum_{k=1}^{\omega-(x+t)} C \cdot {}_{k-1}p^{II}_{x+t} \cdot v^{k-1}, \tag{24}$$

where ${}_{k-1}p^{II}_{x+t}$ is the probability a life in the ill state at age $x + t$ remains in that state for a further $k - 1$ years (Dickson, Hardy and Waters, 2013).

Using this formula we calculated the value of the single premium for our male life aged 50 as $P = €12\,584.37$, which is comparable to the amount obtained by using simulations. However, the ability to determine the premium based on the generation of the trajectory of the non-homogenous Markov chain, represents a more flexible and efficient approach.

We now review the importance of creating multiple scenarios and a sufficient number of simulations in the situation described in order to obtain relevant results for the insurance premium estimation. If we were to implement only one scenario in the form of $n = 1\,000$ simulations, a sufficient accuracy of the results might not be achieved. We have therefore determined the premium as the average value of the 1 000 created premium scenarios whose variability can be seen in Figure 6.

Figure 6 Premium modelling based on the creation of 1 000 scenarios for 1 000 simulations of a non-homogenous Markov chain

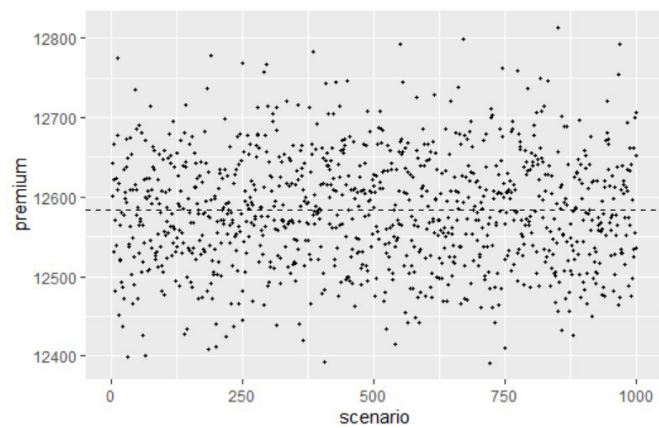


Source: Own construction, customized in R

For comparison, we have shown as a dashed line the value of the premium P = €12 584.37 as determined by the standard equation. The average premium calculated from the presented 1 000 values, each of which was itself calculated as the average from the 1 000 simulated trajectories of the insured life, is €12 552.58. If we choose P = €12 584.37 as a comparative premium value, then with a number of simulations n = 1 000 from the number of 1 000 created scenarios only 508 values of the premium are located in the interval (P – 500, P + 500). So, in 492 cases, the premium differed from the comparative value by more than €500.

If the number of simulations is increased to n = 100 000 the average premium is P = €12 583.42 and there is significantly less variability in the premiums as can be seen in Figure 7. For this number of simulations all 1 000 estimated premium values are in the interval (P – 500, P + 500).

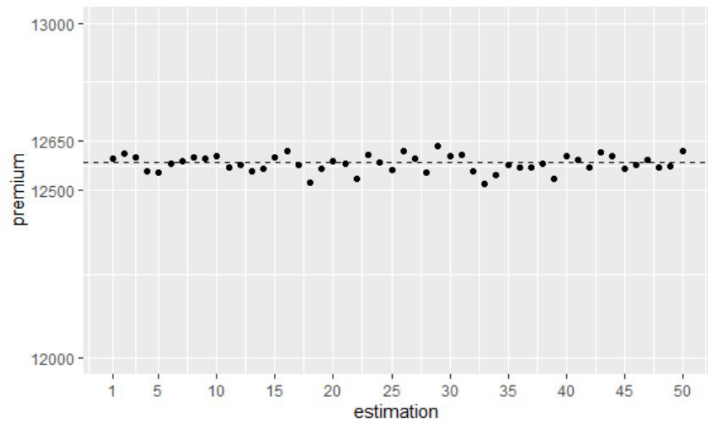
Figure 7 Modelling of the premiums based on 1 000 scenarios for 100 000 non-homogeneous Markov chain simulations



Source: Own construction, customized in R

Therefore, it is important for n = 1 000 simulations to estimate premiums as an average from values obtained from a sufficient number of created scenarios. Of course increasing the number of simulations will ensure more accurate results and we recommend implementing, for example 100 000 simulations. However, one needs to note that when creating 1 000 scenarios with n = 100 000 simulations the calculations using R were time-consuming. On the other hand they provide a sufficiently accurate result. If we were to carry out only one scenario with n = 1 000 simulations we could get an inaccurate estimate of the premium. The solution is to create enough scenarios for the given number of simulations. The result obtained in this way can then be considered as sufficiently accurate. For illustration, in Figure 8 we show 50 possible premium estimates for an alternative 1 000 scenarios with 1 000 simulations in comparison with the value P = €12 584.37.

Figure 8 50 premium estimates for 1 000 scenarios for 1 000 non-homogeneous Markov chain simulations



Source: Own construction, customized in R

It can be noted that the values presented in Figure 8 are comparable to the benchmark premium shown by the dashed line.

CONCLUSION

The use of Markov chains in the context of multi-state models is a frequently used tool for modelling the evolution of conditions in relation to disease incidence and long-term healthcare delivery. By simulating non-homogeneous Markov chains through the generation of their trajectories, we created a model that mimics the real evolution of insured lives' states over time. In the context of the Monte Carlo method, we also discussed in the paper the impact of the number of simulations on the accuracy of the obtained results. Due to the nature of the data in the context of recording a given disease, we performed our calculations in discrete time on an annual basis. The data presented here refer to the male part of the Italian population, where by the ill I(*ill*) state, according to the author of the markovchain package (Giorio Alfredo, Spedicato), we mean disability in the sense of the so-called non-self-sufficiency of the insured life, i.e., disability similar to that of Alzheimer's disease. Using the above modelling implemented using the R language, we have presented the absolute and percentage distribution of insured lives into different states over several years, based on the statistical processing of the generated data, and we have also described it by means of graphical and vector representations. We addressed the analysis of the illness state for the four selected ages, following the trend of its evolution over time. The results obtained could

be used to estimate the costs of a health care institution. Another aspect of the use of the simulation model developed was the estimation of the average number of years that an insured life remains in the healthy state and the estimation of the time during which he/she remains in the ill state. This analysis was also carried out for selected ages, and the situation was presented graphically using bar plots and box plots. The information obtained may be important not only in the context of health care costing, but also in analyses for long-term care insurance contracts. In the last part of the paper we have dealt with the calculation of the premiums for such contracts, presenting in the context of simulations an approach that provides results at a sufficient level of accuracy. We pointed out that insufficient simulations in the premium calculation can provide inaccurate results. This shortcoming can be remedied by creating a sufficient number of scenarios and averaging the premium values we obtained from each scenario. The above analysis was also supported by a graphical presentation of the results of the individual simulation scenarios. The premium values obtained from the simulations were compared with those calculated using a standard life insurance formula and it can be concluded that they are comparable. However, the advantage of the simulation approach lies in greater computational flexibility and the possibility of interactive response when the parameters entering the premium calculation are changed. Modelling the evolution of states over time in the presented domain using Markov chain simulations represents a suitable and efficient solution tool.

ACKNOWLEDGMENT

The paper was supported by a grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic VEGA No 1/0431/22 *Implementation of innovative approaches of modeling and managing risks in internal models of insurance companies in accordance with the Solvency II.*

The paper was supported by a grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic VEGA No 1/0410/22 *Analysis of insurance risks in relation to the economy of the life insurance company.*

References

BERCHTOLD, A. (2001). Markov Chain Computation for Homogeneous and Non-homogeneous Data: MARCH 1.1 Users Guide [online]. *Journal of Statistical Software*, 6(3): 1–81. <https://doi.org/10.18637/jss.v006.i03>.

DE ANGELIS, P., DI FALCO, L. (2016). *Assicurazioni sulla salute:caratteristiche, modelli attuariali e basi tecniche.* Il Mulino.

DICKSON, D. C. M., HARDY, M. R., WATERS, H. R. (2013). *Actuarial Mathematics for Life Contingent Risk.* New York: Cambridge University Press.

DIZ, E., QUERY, J. T. (2012). Applying a Markov model to a plan of social health provisions. *Insurance Markets and Companies*, 3(2): 27–34.

DOBROW, R. (2016). *Introduction to Stochastic Processes with R.* John Wiley & Sons.

DUDEL, C., MYRSKYLÄ, M. (2020). Estimating the number and length of episodes in disability using a Markov chain approach [online]. *Popul Health Metr*., 18(1): 15. <https://doi.org/10.1186/s12963-020-00217-0>.

ESQUÍVEL, L. M., GUERREIRO, R. G., OLIVEIRA, C. M., REAL, C. P. (2021). Calibration of Transition Intensities for a Multistate Model: Application to Long-Term Care. Risks [online]. *MDPI*, 9(2): 1–17, <https://doi.org/10.3390/math9131496>.

FECENKO, J. (2018). *Teória pravdepodobnosti II v MAXIME.* Bratislava: Letra Edu.

FERNANDEZ-MORALES, A. (2015). Application of a Discrete-time Markov Chain Simulation in Insurance [online]. *International Journal of Recent Contributions from Engineering, Science´& IT (iJES)*, 3(3): 27–32. <https://doi.org/10.3991/ijes.v3i3.4929>.

FLEISCHMANN, A., HIRZ, J., SIRIANNI, D. (2021). A long-term care multi-state Markov model revisited: a Markov chain Monte Carlo approach [online]. *European Actuarial Journal*. <https://doi.org/10.1007/s13385-021-00285-y>.

GARG, L., MCCLEAN, S. et al. (2010). A non-homogeneous discrete time Markov model for admission scheduling and resource planning in a cost or capacity constrained healthcare systém [online]. *Health Care Manag Sci*, 13: 155–169. <https://doi.org/10.1007/s10729-009-9120-0>.

GEYER, C. J., JOHNSON, L. T. (2013). *mcmc:Markov Chain Monte Carlo* [online]. <http://CRAN.R-project.org/package=mcmc>.

HORÁKOVÁ, G., MUCHA, V. (2002). Určenie rozdelenia celkových škôd s využitím metódy Monte Carlo a jeho porovnanie s numericky presným výpočtom v danom portfóliu poistných zmlúv. *Managing and Modelling of Financial Risks*, VŠB – Technical University of Ostrava, 77–81.

JANKOVÁ, K., KILIANOVÁ, S., BRUNOVSKÝ, P., BOKES, P. (2014). *Markovove reťazce a ich aplikácie.* Bratislava: Epos.

JONES, W. P., SMITH, P. (2018). *Stochastic Processes. An Introduction.* Taylor & Francis Group.

LI, M., DUSHOFF, J., BOLKER, B. M. (2018). Fitting mechanistic epidemic models to data: a comparison of simple Markov chain Monte Carlo approaches [online]. *Statistical Methods in Medical Research*, 27(7): 1956–1967. <https://doi:10.1177/0962280217747054>.

MACIAK, M., OKHRIN, O., PEŠTA, M. (2021). Infinitely stochastic micro reserving [online]. *Insurance: Mathematics and Economics*, 100: 30–58. <https://doi.org/ 10.1016/j.insmatheco.2021.04.007>.

MUCHA, V., PÁLEŠ, M. (2018). *Teória pravdepodobnosti pre ekonómov. S podporou jazyka R.* Bratislava: Letra Edu.

NICHOLSON, W. (2013). *DTMCPack: Suite of functions related to discrete-time discrete-state Markov Chains* [online]. <https://CRAN.R-project.org/package=DTMCPack>.

PASARIBU, S. U., HUSNIAH, H., SARI, N. K. R., YANTI, R. (2019). Pricing Critical Illness Insurance Premiums Using Multiple State Continous Markov Chain Model [online]. *Journal of Physics*, 1366. <https://doi:10.1088/1742-6596/1366/1/012112>.

SATO, R., ZOUAIN, D. (2010). Markov Models in health care [online]. *Einstein*, 8(3): 376–379. <https://10.1590/S1679-45082010RB1567>.

SPEDICATO, A. G. (2017). Discrete Time Markov Chains with R [online]. *The R Journal*, 9(2): 84–104. <https://doi.org/10.32614/RJ-2017-036>.

SPEDICATO, A. G. et al. (2017). *The markovchain Package: a Package for Easily Handling Discrete Markov Chains in R* [online]. <https://cran.rproject.org/web/packages/markovchain/vignettes/an_introduction_to_markovchain_package.pdf>.

UNWIN, H. J. T., ROUTLEDGE, I., FLAXMAN, S., RIZOIU, M.-A., LAI, S., COHEN, J. et al. (2021). Using Hawkes Processes to model imported and local malaria cases in near-elimination settings [online]. *PLoS Comput Biol*, 17(4). <https://doi.org/10.1371/journal.pcbi.1008830>.

ŠKROVÁNKOVÁ, L., SIMONKA, Z. (2021). *Aktuárske metódy a modely v penzijnom, zdravotnom a nemocenskom poistení.* Brno: H.R.G. s.r.o.

TWUMASI, C., ASIEDU, L., NORTEY, E. (2019). Markov Chain Modeling of HIV, Tuberculosis, and Hepatitis B Transmission in Ghana [online]. *Interdisciplinary Perspectives on Infectious Diseases.* <https://doi.org/10.1155/2019/9362492>.

XIE, H., CHAUSSALET, T. J., MILLARD, P. H. (2005). A continuous time Markov model for the length of stay of elderly people in institutional long-term care [online]. *Journal of the Royal Statistical Society: Series A*, 168(1): 51–61. <https://doi.org/10.1111/j.1467-985X.2004.00335.x>.

WICKHAM, H. (2016). *Ggplot2, Elegant Graphics for Data Analysis.* New York: Springer-Verlag.

Digitalization Index: Case for Banking System

Nataliia Versal[1] | *Taras Shevchenko National University of Kyiv, Kyiv, Ukraine*
Vasyl Erastov[2] | *Taras Shevchenko National University of Kyiv, Kyiv, Ukraine*
Mariia Balytska[3] | *Taras Shevchenko National University of Kyiv, Kyiv, Ukraine*
Ihor Honchar[4] | *Taras Shevchenko National University of Kyiv, Kyiv, Ukraine*

Received 4.4.2022 (revision received 1.6.2022), **Accepted** (reviewed) 14.6.2022, **Published** 16.12.2022

Abstract

Economy digitalization has become a trend during the pandemic. The banking sector was also one of the first to face the need to accelerate digitalization. This work is devoted to developing a digitalization index for both the banking sector and an individual bank based on a set of indicators calculated according to data from the World Bank and data from commercial banks. At a macro level, the study concluded that the pandemic has accelerated the digitalization of the banking sector in all the monitored countries; however, a significant increase was observed in countries with lower index values in the pre-pandemic period. At the micro-level, the study showed that digital banks had benefited from digitalization more during the pandemic, unlike classical banks.

Keywords	***DOI***	***JEL code***
Digitalization, financial innovation, digital transformation, digital banking, fintech	*https://doi.org/10.54694/stat.2022.16*	*G20, G21, O33*

INTRODUCTION

The development of innovative technologies has significantly impacted the financial sector. New trends affected financial institutions' business processes and provided financial services. Banks often earlier dominated the financial markets of many countries, now increasing competition in its three segments – the payment market, the deposit market, and the credit market – leads to fintech companies winning back an increasing part of them. The digitalization of financial services has become a dominant idea, further spurred by the Covid-19 pandemic. Thus, according to Codebase Technologies (2021), there was a decrease in cash settlements at points of sale compared to 2019 by 32.1%. In turn, the World Bank and Cambridge Center for Alternative Finance, 2020, based on a survey of financial market regulators, provides the following data on the growth

[1] Department of Insurance, Banking and Risk-Management, Faculty of Economics, Taras Shevchenko National University of Kyiv, Vasylkivska, 90 A, 03022, Kyiv, Ukraine. Also Faculty of Business and Economics, Mendel University, Zemědělská 1, 613 00, Brno, Czech Republic. E-mail: natalia_versal@knu.ua.
[2] Department of Insurance, Banking and Risk-Management, Faculty of Economics, Taras Shevchenko National University of Kyiv, Vasylkivska, 90 A, 03022, Kyiv, Ukraine. E-mail: v_erastov@knu.ua.
[3] Department of Insurance, Banking and Risk-Management, Faculty of Economics, Taras Shevchenko National University of Kyiv, Vasylkivska, 90 A, 03022, Kyiv, Ukraine. E-mail: m.balytska@knu.ua.
[4] Department of Statistics, Information and Analytical Systems and Demography, Faculty of Economics, Taras Shevchenko National University of Kyiv, Vasylkivska, 90 A, 03022, Kyiv, Ukraine. E-mail: igonchar@knu.ua.

of digital financial services (DFS) under the influence of the Covid-19 pandemic. 65% of respondents in emerging market & developing economies and 50% in advanced economies stated an increase of usage of digital payments and remittances. Accordingly, considering digital savings or deposits: 22% vs. 12%, digital lending: 14% vs. 12%, insurtech (incorporation concept of blockchain, artificial intelligence, digitalization, and the sharing economy in the insurance industry according to Lisowski, Chojan, 2020): 10% vs. 24%, digital capital raising: 10% vs. 6%, wealthtech (incorporation concept of using digital technologies as well as tailor-made products and services in investment and client portfolio management (Dziawgo, 2021)): 6% vs. 24%.

Banks have taken up this challenge, though not immediately. One of the reasons for the slow response of banks to innovation is the high degree of regulation, which, in turn, is necessary for bank customers' protection. In addition, the principles of due diligence and knowing your client (KYC) play a crucial role in banking. Most likely, banks would have continued to adapt to the era of digitalization at their own pace, but the Covid-19 pandemic has changed everything. It accelerated the application of innovative technologies in the banking sector, and DFS became a trend for banks. Thus, according to Deloitte Digital (2020), under the influence of Covid-19, 41% of banks increased contactless payment limits, 34% implemented fully digital processes, and 18% launched contactless payment methods. Regulators, in turn, made certain concessions, such as opening accounts without the need to physically visit bank branches, expanding the possibilities of digital ID, etc. (World Bank and Cambridge Center for Alternative Finance, 2020).

Banks' digitalization of financial services can be considered at different levels and perspectives. However, today, the concept of DFS (financial services provided using digital technologies (Pazarbasioglu et al., 2020)) is rather vaguely disclosed. So, the current definition of digital banks in the context of providing financial services specifically to retail clients requires more precise criteria. That is why this study aims to develop approaches to determining indicators based on which it is possible to calculate the index of digitalization of the banking sector in the retail segment, and if data is available, the index of digitalization of banks, as well as to demonstrate the features of the functioning of digital banks in the countries of Central and Eastern Europe in time of Covid-19 pandemic.

The study is structured as follows. The first section contains a literature review, the second section covers data description and methodology, the third section provides main research findings, and the final section concludes.

1 LITERATURE REVIEW

Based on the purpose of our study, we reviewed works that were focused on defining digital transformation (DT) processes. Banks themselves cannot be digital if there are no prerequisites for this, such as a reliable Internet, the presence of gadgets for communication with the bank, in other words, an infrastructure that allows the development of digital banking. We also drew attention to studies that examine the issues of DFS and financial inclusion, including digital financial inclusion. And finally, the main areas of study of digital banking. Based on our goal, we need to understand how it is possible to determine the digitalization level of the banking sector in the segment of retail services for individuals of a particular country and how digital banks can be identified.

The vision of digital transformation is quite a broad definition. Maheshwari (2019) reveals a holistic picture of digital transformation, particularly the digital transformation factors, methods, and technologies. He emphasizes that DT is about introducing new technologies to all business segments. He also mentions the creation of a 'post-personal computers era,' which is extremely important for digital banks since the development of banking applications must consider the variety of devices used by households. Verhoef et al. (2021) look at digital transformation in great detail, highlighting drivers, phases, and strategic directions. In the context of our study, it is vital to stress such digital transformation factors: digital technology, digital competition, and digital customer behavior. Digital banking is also shaped by these drivers and should be considered when building a digitalization index. Zaoui and Souissi (2020)

go further and offer a roadmap for digitalization, specifying the stages of its implementation. Finally, Mergel et al. (2019) point to the example of the public sector that digital transformation is not just a transition to online; it is a genuinely holistic process that cannot be stopped at any stage due to new opportunities that are constantly emerging.

Wewege and Thomsett (2019) emphasize that banks were not ready for the FinTech revolution, so, today, they are trying to catch up, which is often associated with significant capital investments. The authors also point out that certain groups of households are not ready to go online (in particular, older adults), which again underlines the importance of research and issues of digital financial inclusion. In addition, the authors emphasize the need to change banks' infrastructure and business models. Thus, the provision of DFS is accompanied by quite serious challenges. Still, at the same time, it provides significant benefits in the long term and contributes to the growth of financial inclusion.

A comprehensive study by Pazarbasioglu et al. (2020) reveals the DFS need for overcoming poverty and ensuring economic growth. In particular, it is pointed out that DFS can speed up payments, make savings and investments more accessible, and reduce lending costs. In this paper, DFS are presented as financial services provided using digital technologies. DFS models include mobile money, platform eco-systems (BigTech platforms), and Application Programming Interfaces. In addition, the authors reveal the gaps in traditional financial service delivery models (speed, cost, transparency, access, security) and show how DFS can overcome them. Banna and Alam (2021) also confirm that the development of digital finance contributes to sustainable development and contributes to the achievement of Sustainable Development Goals.

According to Agur et al. (2020), DFS are financial services delivered through digital channels. These authors emphasize that Covid-19 has only spurred the development of DFS. The main indicators that the authors pay attention to are the value of digital payments transactions, the number of users of digital payments, the value of digital lending, the number of digital loans, digital remittances. Some of these indicators can be used to assess the digitalization of the banking sector or an individual bank. However, the authors also draw attention to the fact that the rapid digitalization of financial services can also have negative consequences, in particular, the gap between the rural and urban population, youth and older adults, etc., may increase.

In the study by Riley et al. (2020), digital channels of DFS specify 'the Internet, mobile phones, ATMs, point-of-sale terminals, electronically enabled cards, and biometric devices.' It allows taking into account the number of POS terminals, ATMs, and, of course, issued bank cards in the digitalization index. Lyons and Kass-Hanna (2021) show in great detail the transition from classic financial services to DFS in the context of four blocks of financial services: payments and transfers (mobile payments, mobile money, mobile PoS, P2P, B2B, digital and virtual money), savings and investments (mobile banking, mobile trading, etc.), borrowing and financing, and risk management (digital insurance).

Thus, digital transformation, and in particular the very rapid competitive development of DFS by technology companies, has led to the fact that classical banks began to change and become, to one degree or another, digital banks. However, it can be stated that at the moment, the concept of digital banking is not clearly defined. Thus, Ehrentraud et al. (2020) define digital banks as banks that use new technologies and build their business models on these technologies and provide services remotely with a minimal number of branches or no branches at all. Thus, the digital banking definition covers a vast range of banks, particularly neo-banks such as Revolut, Vialet, Bunq, etc. At the same time, in the study of Deloitte Digital (2020), the analysis of digital banks is carried out among banks that provide DFS. In this paper, the authors divide digital banks into four groups: digital latecomers, digital adopters, digital smart followers, and digital champions. This division into groups is carried out according to three criteria: functionalities benchmarking, including, among others, analysis of core banking services digitalization; customer needs, including customer preferences between main channels – branches, internet, mobile;

user experience. However, this analysis will be inaccessible to many since the databases are closed. At the same time, some of the conclusions of this work are as follows: digital champions are retail-focused; mobile channels and Internet channels are peer-to-peer DFS delivery channels; the main services that are of interest to the clients of such banks are transactions (payments), saving & investment, while banks are trying more to promote digital loan services.

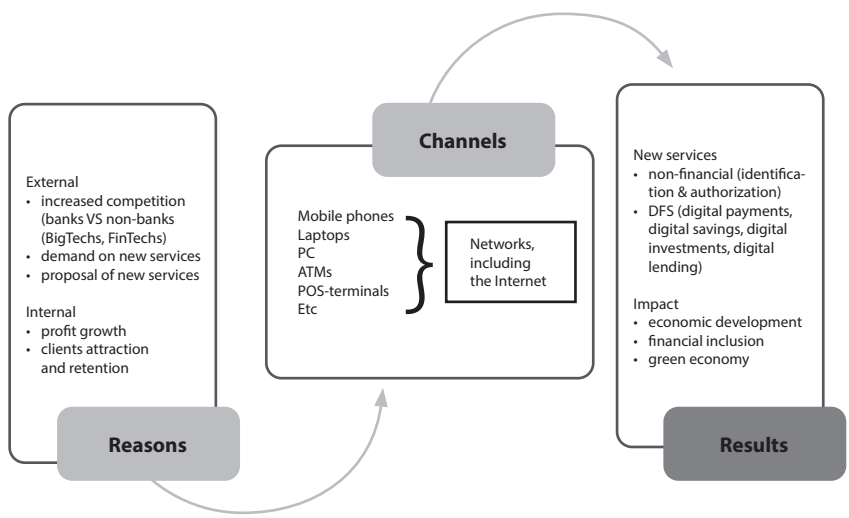
Carletti et al. (2020), in this regard, indicate that many banks wishing to adapt to new realities will implement digital technologies, but not many banks will be successful in this. Therefore, for example, developing an index of digitalization of the banking sector or an individual bank will make it possible to find dependencies with indicators of financial stability and performance efficiency. Also, the development of this index will allow us to analyze the level of financial inclusion and answer questions about whether the digitalization of banking services can affect vulnerable groups.

2 MATERIAL AND METHODS

The literature review allows us to come to certain conclusions that are important from the point of view of methodology. For example, suppose we accept the definition of DFS as financial services that are provided and used through digital channels. In that case, one of the signs of digital banking in the retail segment may be the ability and ability to provide DFS. At the same time, the literature review showed that different authors interpret digital channels in different ways. That is why we proceed from the fact that there are two main ways to use the bank's services: through branches and technology-enabled channels. At the same time, it is rather challenging to determine technology-enabled channels because both the tools (mobile phones or laptops) and the existing networks through which data is transferred (the Internet) are mixed. It should also be noted that countries can have significant differences in the use of banking services (for example, countries in Africa, where mobile banking has an entirely different meaning than in Europe). Therefore, further proposals for constructing the index are more related to European practice.

In general, if you imagine how you can show digital retail banking, then the visualization can be demonstrated in Figure 1.

Figure 1 Taxonomy of digital retail banking



Source: Authors based on Mergel et al. (2019)

During the Covid pandemic, there was an acceleration in the digitalization of banks, despite regulatory problems, in particular in the area of customer identification. As a result, more and more banks, in our opinion, will undergo digital transformation. In our article, we want to achieve a triad of goals: firstly, creating a digitalization index of the banking system. Secondly, outlining the criteria for determining digital banks from traditional banks, using traditional quantitative indicators and digital footprint methods. Thirdly, showing whether the digital banks have managed to bypass the banking sector as a whole.

Covid influence investigation on digital financial services growth was conducted in five countries in Central and Eastern Europe, which had a steady upward digitalization trend and had a significant impact of Covid-19 in 2020 according to Our World in Data, 2022.

It was used two main metrics: the Digital Economy and Society Index (DESI), suitable only for EU countries; the author's proposed Digitalization index, suitable for all the chosen countries. The proposed digitalization index is based on World Bank data and represents digitalization adoption and banking readiness (see Table 1).

Table 1 Components of digitalization index of banking system

Components		Blocks	Scale		Calculation
			Min	Max	
Core indicators					
World Bank Data	Tier 1	Automated teller machines (ATMs) (per 100 000 adults)	0	50	$q = \sqrt[n-1]{\frac{b_n}{b_1}}$ $b_1, b_1 q, b_1 q^2, b_1 q^3 \ldots b_1 q^{n-1}$ Python script for scoring
		Commercial bank branches (per 100 000 adults)	0	50	
	Tier 2	Population density (people per sq. km of land area)	0	50	
		Employment to population ratio, 15+, total (%)	0	50	
	Tier 3	Secure Internet servers (per 1 million people)	0	50	
		Fixed broadband subscriptions (per 100 people)	0	50	
		Fixed telephone subscriptions (per 100 people)	0	50	
		Individuals using the Internet (% of population)	0	50	
		Mobile cellular subscriptions (per 100 people)	0	50	
Encouraged indicators					
Bank Level Data	Tier 4	Number of payment cards issued	1	n	Python script for scoring
		Volume of online payments	1	n	
		Number of online payments	1	n	
		Number of POS terminals	1	n	
	Tier 5	Identification without physical office visiting (deposits)	0	1	Boolean
		Paperless workflow (deposits)	0	1	
	Tier 6	Ability to borrow money without physical office visiting	0	1	
		Identification without physical office visiting (loans)	0	1	
		Paperless workflow (loans)	0	1	
	Tier 7	Number of app downloads	1	m	Python script for scoring
		Number of unique websites visitors	1	m	

Source: Authors

A respective methodology was proposed to estimate the level of banking sector digitalization. It has two main dimensions: Core Indicators and Encouraged Indicators, describing overall banking sector digitalization level and particular banking entity digitalization level, respectively.

Core indicators are the crucial part of estimation because the high digitalization level of a separate bank can be strangled by a system that is not ready for such options in banking and vice versa. Core indicators are subdivided into three tiers, containing scored raw indicators obtained from World Bank statistics. Tier 1 is aimed to estimate overall banking readiness and is represented by two indicators: automated teller machines and commercial bank branches, both per 100 000 adults. These indicators should reveal society and the banking sector readiness to reduce cash and physical banking operations and increase digitally-driven analogs. Both are inversed, so the higher are raw values the lower are scores for them. This tier is partially limited and influenced by historical and geographical issues. Still, it could be enhanced or normalized by adding some more specific indicators, but in prejudice of raw data availability.

Tier 2 represents the main potential consumers' digitally-driven banking operations. It is subdivided into two scored raw indices like employment to population ratio and population density. Population density is inversed in terms of scoring because the denser is population, the less stimulus is to reduce cash operations and banking entities visiting due to their high availability. The employment to population ratio is straight in scoring and represents potential customers with available resources and efforts for digitalization shifts in behavior, so the higher raw values are, the higher score will be. This set of indicators can also be considered limited due to external and internal influences, but raw data availability is not affected while representing potential banking digitalization adopters. Tier 2 can also be extended and normalized via additional raw indices describing some specific issues or overall situations but in terms of data availability.

Tier 3 is a set of indicators representing infrastructure readiness in terms of both stable and high-speed Internet access availability and Internet security orientation. It is based on five scored raw indicators such as share of Internet users in overall population; amount of secure Internet servers per 1 million people; mobile cellular, fixed telephone and fixed broadband subscriptions per 100 people. There is one inverted indicator, fixed telephone subscriptions. Even in terms of utilizing fixed telephone subscriber line as a media for DSL connection, it is out of date and slackens overall infrastructure readiness. Other indicators are calculated simply, higher is better. Proposed indicators are not the only to describe current and historical situation but is quite sufficient for estimation. Also set could be expanded to normalize or to deepen scoring, but it could affect data availability.

Every country-level raw value forming core indicators is available in the World Bank statistics. For estimation purposes, all raw values are normalized by cutting the first and last 5% interval and then distributed by geometric sequence with precalculated ratio and scored with Python script from 0 to 50. Normalizing with cut intervals is used to level out significant shifts of worst and best raw values that can affect distribution, which is also enhanced with geometric sequence to normalize gaps in some raw indicators. Also, there are some evolutional and historical shifts in raw data, making raw scoring inefficient. Although scoring can be made with Excel or Python script, the final decision will depend on the scoring interval. If the scoring interval is tightened to 1–10, the overall amount of calculation will drop significantly and can be done manually. In the case of widening the scoring interval to 1–100, there will be a drastic increase in the calculation, and scripting is a better instrument.

As it was stated, encouraged indicators are aimed to estimate bank-level data. The encouraged indicators reflect four tiers, from Tier 4 to Tier 6. These tiers are based on different data sources and can be excluded or modified to improve overall scoring efficiency.

The first set of indicators could be calculated in two dimensions concerning available data. This set is Tier 4, subdivided into four indicators: number of payment cards issued, volume of online payments, number of online payments, and number of POS terminals. Depending on available data, these indicators

can represent the overall banking industry, in case of only country-level data availability or banking entity data. Scoring will be done in respect of obtained data e.g., country-level data should be scored via Python script, in case of banking entity data – simple ascending scoring of raw values, in case of both available country and entity data available – share of particular entity should be used and then scored in ascending order from less to higher values. So, the scoring interval could be 1-n, where n is a number of entities or countries in comparison.

Tier 5 is aimed to evaluate deposit-taking activities' readiness to be digitally transformed. It consists of two Boolean indicators: identification without physical office visiting and paperless workflow. These indicators are mostly developed to be used with a single bank. Still, in the case of banking industry evaluation, it could be transformed to calculating the share of individual banks that match requirements as a fraction of 1 or, in case of data availability issues, it can be a simple Boolean where 0 is the legislative prohibition of such activities and 1 – no prohibitions.

Tier 6 is used to estimate the digital readiness in lending. It is three dimensional Boolean based on identification without physical office visiting and paperless workflow, and the ability to borrow money without physical office visiting. Tier 6 is also developed to be used with a single bank. Still, in the case of banking industry evaluation, it could be transformed to calculating the share of individual banks that match requirements as a fraction of 1 or, again, in case of data availability issues, it can be a simple Boolean where 0 is the legislative prohibition of such activities and 1 – no prohibitions.

Tier 5 and Tier 6 are expert-based estimations according to legislation aspects and data provided by banks within the comparison. In case there are no legislative limitations, only the bank's data is analyzed. In contrast, if there are limitations, it should be analyzed both to exclude law violating and cheating subjects from scoring. In case of insufficient data provided by the bank (uncertain norms), it is assumed that the bank is not providing such an opportunity.

Tier 7 is a set of specific indicators representing a particular banking entity's readiness to convert physical operations to digital through a website or mobile app. This Tier is optional in the banking industry examination but is helpful for particular entities comparison. To estimate Tier 7, app downloads and unique website visitors' numbers are used. Depending on available data, app downloads can be calculated in different ways. The simplest way is to use raw downloads number from any mobile application stores, e.g., Play Market; a more accurate way is to calculate the sum of downloads in AppStore and Play Market. The most accurate and challenging course is calculating year-to-year growth, but this calculation is highly limited by data availability. Website visitor's indicator also has different estimation variations. Depending on used metrics, annual, monthly, quarterly, or even instant data can be used. The best way to score obtained data is to compare raw values and set scores from 1 to m, where m is the number of banks in the comparison group, in ascending order, so the higher indicator value is better. Overall Tier 7 score will be a sum of individual indicators divided by 2.

After scoring, each individual index overall index is calculated as a sum of Tier 1–Tier 3 indexes, divided by the number of underlying indexes, plus Tier 4, Tier 5, Tier 6, and Tier 7 indexes. The final index value will be in (0 – 50) + (0 – 5) + (1 – n) + (1 – m) interval, 0 – 50 for Tier 1–Tier 3 and 0 – 5 for Tier 5–Tier 6, 1 – n for Tier 4 and 1 – m for Tier 7.

Digitalization index allows to get a stable year to year evaluation because it is not sensitive to insignificant changes in one of the underlying indices and is not prone to significant random fluctuations, allowing to analyze its dynamics qualitatively: assess the general vector and speed of development, possible cycles, and the impact of short-term factors.

These two metrics are aimed to show the trend of digital adoption at both the country-level and banking system-level, in particular, to compare the difference on different system levels.

To understand the differences in digital banks' behavior on the market, each country's three top digital banks were chosen according to website and application usage and visitors metrics (see Table 2).

Table 2 TOP-3 digital banks according to Similarweb

Country	Digital bank	App	Place in Play Store	Place in App Store	Total visits to the site of digital bank, mln	Usage rank by SimilarWeb
Poland	PKO Bank Polski SA	IKO	1	1	6.86	1
	Santander Bank Polska S.A.	Santander mobile	2	6	7.39	5
	Bank Millennium SA	Bank Millennium	3	7	7.29	6
Hungary	OTP Bank Nyrt.	OTP Bank HU	1	1	6.42	14
	K&H Bank Zrt.	K&H mobilbank	11	10	1.37	2
	Erste	George Magyarország	14	11	1.42	5
Czech Republic	Československá obchodní banka, a.s.	ČSOB Smart	2	6	5.19	85
	Česká spořitelna, a.s.	George Česká spořitelna	4	3	5.55	1
	MONETA Money Bank	Smart Banka	8	11	2.07	6
Slovak Republic	Slovenská sporiteľňa, a.s.	George Slovakia (George Slovensko)	5	2	2.85	1
	VÚB, a.s.	VÚB Mobile Banking	6	12	1.51	3
	365.bank, a.s.	365.bank	7	11	0.08	12
Ukraine	Universal bank | Monobank	monobank	2	2	0.70	2
	JSC CB PrivatBank	Privat24	3	1	11.73	1
	Oschadbank	Oschad 24/7	4	3	3.36	3

Source: Authors, based on <*similarweb.com*> (21.12.2021)

The determining criterion for choosing digital banks was the place of the bank`s application in the Play Store, since as of December 2021, according to StatCounter, the share of Android users is 70.01%, and IOS – 29.24%. Also, the place of the bank's application in the App Store, Total Visits to the site of the digital bank, and Usage Rank by SimilarWeb was taken into account. Finally, it is worth noting that another criterion that can be used to analyze the growth in popularity of the digital banks' applications can be Google Trends. Still, to use it, you need to know the nuances of the search (language, abbreviations, different bank and application names, etc.) in a particular country.

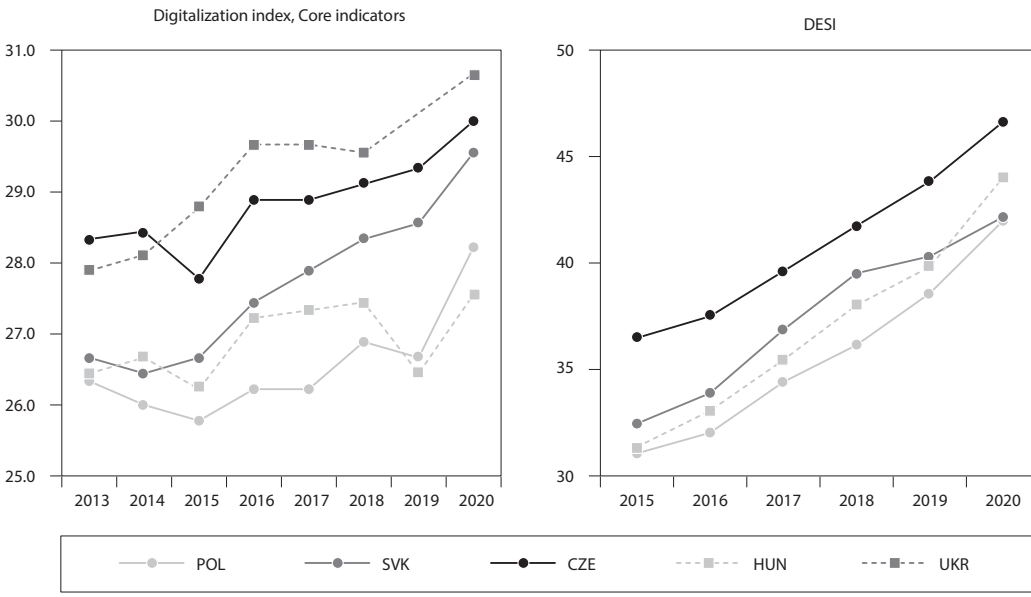
3 RESULTS AND DISCUSSION

First, the digitalization of banks in itself is meaningless without necessary prerequisites (Internet connection, etc.). That is why it is so critical to consider such factors. It is possible either through the definition of the DESI or through the Digitalization Index, particularly the core indicators (Tier 1–Tier3). Secondly, it is vital to determine which banks are digital. That is why we are introducing encouraged indicators. At the same time, there are severe issues with data collection, which banks do not always disclose. Thirdly, to understand whether digital banks really stand out among other banks, we demonstrate in the example of Ukraine a comparison for three groups of banks depending on the form of ownership: state, foreign and local private.

According to the DESI comparison chosen set of countries showed mostly similar uprising trends. The Czech Republic showed the best results in every year of the observed period. Poland showed lower results. The proposed methodology also highlights the uprising trend in the Digitalization index. Ukraine got the dominant position in digitalization adoption, overcoming the Czech Republic, whose dominant position among European countries is undoubted. Different approaches can explain this situation

in the EU and Ukraine because Ukraine is trying to withstand internal and external crises while heading for integration processes. Some significant changes in the Digitalization index, for example, in Hungary, can be explained by Covid shifts in raw data (see Figure 2).

Figure 2 Digitalization Index and DESI comparison



Note: Digitalization Index is calculated based on Tier 1–Tier 3 Data.
Source: DESI, World Bank Data

All countries have had a gradual upward trend in the digitalization level during 2013–2019. There was a relatively uniform growth rate 0.1–0.4 points per year (see Table 3), with the exception of Hungary, where the weakest growth trend was accompanied by cyclical fluctuations.

Table 3 Digitalization index growth comparison

Countries	Expected digitalization index (forecast)	Expected growth in 2020, p.p.	Average annual growth in 2013–2019 (fact), p.p.	Actual index value	Actual growth in 2020, p.p.
Poland	26.8	0.1	0.1	28.2	1.6
Czech Republic	29.5	0.1	0.2	30.0	0.7
Slovak Republic	29.0	0.4	0.3	29.6	1.0
Hungary	27.2	0.8	0.0	27.6	1.1
Ukraine	30.9	0.8	0.4	30.7	0.6

Source: Authors, based on Digitalization Index Data

The SARS-CoV-2 virus has significantly increased the 2020 digitalization level in all countries. However, in order to objectively assess the impact of a set of financial factors due to the emergence of SARS-CoV-2

virus, it is necessary to compare the actual estimate of 2020 with the expected estimations (projected) 2020 digitalization level that countries should achieve in the absence of SARS-CoV-2.

The digitalization level is determined by 9 components (underlying indices). Their analysis allows to identify significant growth factors, assess the stabilization process of each country, characterize the specifics and assess the development potential (see Figure 3).

Figure 3 Digitalization Index underlying indices comparison

		Automated teller machines (ATMs) (per 100 000 adults)	Commercial bank branches (per 100 000 adults)	Employment to population ratio, 15+, total (%)	Fixed broadband subscriptions (per 100 people)	Fixed telephone subscriptions (per 100 people)	Individuals using the Internet (% of population)	Mobile cellular subscriptions (per 100 people)	Population ages 15–64 (% of total population)	Population density (people per sq. km of land area)	Secure Internet servers (per 1 million people)
Poland	2013	9	6	19	45	9	44	47	47	19	39
	2014	8	5	20	45	7	44	47	46	19	39
	2015	7	4	21	44	7	44	46	44	19	40
	2016	8	4	22	44	8	45	45	43	19	41
	2017	8	5	23	44	9	44	43	42	19	41
	2018	9	5	25	44	10	44	41	40	20	44
	2019	10	6	25	44	9	44	39	38	20	43
	2020	13	8	27	44	10	47	41	36	20	44
Slovak Republic	2013	9	9	19	46	11	47	40	49	20	39
	2014	9	7	20	46	11	47	39	48	20	39
	2015	9	5	22	46	11	46	40	47	20	41
	2016	10	6	23	46	11	47	42	46	20	42
	2017	10	7	25	46	12	46	43	45	20	42
	2018	11	8	26	46	12	45	43	43	21	43
	2019	12	9	26	47	12	45	42	41	21	43
	2020	14	10	28	47	13	47	42	40	21	44
Czech Republic	2013	11	10	25	48	10	47	43	42	18	43
	2014	11	10	26	48	10	46	43	41	18	44
	2015	11	8	27	47	10	46	39	39	18	44
	2016	12	10	28	47	11	45	39	37	18	50
	2017	11	11	29	47	11	45	39	36	18	49
	2018	12	11	30	47	12	45	38	34	18	49
	2019	13	13	30	48	11	45	37	32	19	48
	2020	15	14	31	48	12	45	37	31	19	49

Figure 3 (continuation)

		Automated teller machines (ATMs) (per 100 000 adults)	Commercial bank branches (per 100 000 adults)	Employment to population ratio, 15+, total (%)	Fixed broadband subscriptions (per 100 people)	Fixed telephone subscriptions (per 100 people)	Individuals using the Internet (% of population)	Mobile cellular subscriptions (per 100 people)	Population ages 15–64 (% of total population)	Population density (people per sq. km of land area)	Secure Internet servers (per 1 million people)
Hungary	2013	9	17	13	48	6	46	40	43	20	39
	2014	9	17	17	47	5	46	40	42	20	39
	2015	10	16	19	47	5	45	33	41	21	40
	2016	10	17	22	47	4	46	34	40	21	44
	2017	10	17	23	47	4	45	34	39	21	45
	2018	11	18	24	48	4	44	32	37	21	45
	2019	12	10	24	48	4	44	31	35	21	44
	2020	15	10	27	48	4	46	31	34	22	45
Ukraine	2013	2	50	20	40	7	37	45	45	24	26
	2014	4	50	17	39	8	38	46	45	24	27
	2015	5	50	17	41	9	38	46	44	24	29
	2016	5	50	17	41	9	38	43	43	24	40
	2017	4	50	17	41	10	39	43	42	24	39
	2018	4	50	17	41	11	38	41	40	25	39
	2019	5	50	17	42	14	40	40	39	25	38
	2020	7	50	17	43	16	40	40	38	25	38

Source: Authors' calculation based on World Bank Data

Worst  Best

The highest growth rates of the Digitalization level were observed in 2020 in Poland: 5.8% growth, which is due to most components of the index, which reflects the integrated development of digitalization processes. The most significant contribution to the Digitalization level growth in the indices structure was made by: Automated teller machines (ATMs) (per 100 000 adults) and Individuals using the Internet (% of population) together these factors account for 44% growth; less contribution was made by: Commercial bank branches (per 100 000 adults), Employment to population ratio, 15+, total (%), Mobile cellular subscriptions (per 100 people): 14% each; the lowest weight was: Fixed telephone subscriptions (per 100 people), Secure Internet servers (per 1 million people): 7%.

However, such rapid growth is taking place at a relatively low Digitalization level, which was formed in the previous years, and, therefore, despite the rapid growth in 2020, Poland failed to catch up with the leading countries.

Hungary occupied a second place in terms of the digitalization level growth in 2020: 4.2%, but, like Poland, with a low baseline digitalization level, could not significantly improve its results. Among the growth factors of digitalization, there is no comprehensive growth in all components, for example, Automated teller machines (ATMs) (per 100 000 adults) and Employment to population ratio, 15+,

total (%) – a total of 60% growth; on Individuals using the Internet (% of population): 20%; and 20% – the remaining factors.

Hungary showed a significant increase in the level of digitalization in 2020, but an analysis of the components of the index suggests that the country has not created the preconditions for the development of all areas of digitalization.

Slovakia has shown a 3.5% increase in digitalization level. The structure analysis of the influence factors allowed us to conclude the comprehensiveness of growth, which is formed by: Commercial bank branches (per 100 000 adults) and Fixed telephone subscriptions (per 100 people) – a total of 58%; Automated teller machines (ATMs) (per 100 000 adults), Population density (people per sq. km of land area) and Fixed broadband subscriptions (per 100 people): 14% each. However, even in the evidentiary period, the country showed a rapid growth in the digitalization level in 3–4 underlying indices, which indicates a sustainable, integrated development of digitalization processes.

A significant increase of digitalization level is observed in the Czech Republic: 2.3%. In 2020, the growth is driven by five factors: Automated teller machines (ATMs) (per 100 000 adults): 33%; Commercial bank branches (per 100 000 adults), employment to population ratio, 15+, total (%), fixed telephone subscriptions (per 100 people), secure Internet servers (per 1 million people): 16–17%. However, in previous years the country has shown a growth trend due to the influence of 3–4 components of the index, which is higher than in other countries. The country has less digitalization level growth. compared to Poland and Slovakia, but high growth in 2020 and the existed pre-pandemic growth trend (for 2013–2019) can make the Czech Republic leader.

In 2020, Ukraine showed the lowest digitalization level growth: 1.8%. The low growth estimation is due to the lowest number of growth factors among the analyzed countries, which indicates the worst preconditions for the growth of digitalization of the country, or the lack of such necessity. In 2020, fixed telephone subscriptions (per 100 people) and automated teller machines (ATMs) (per 100 000 adults) provided a total of 80% growth in digitalization level, and fixed broadband subscriptions (per 100 people) another 20%.

Despite the low growth level in 2020, Ukraine has a high digitalization level, partly due to significant achievements of previous years, such as the digital reform of 2019 – `State in a smartphone`. The reform allowed citizens to obtain many public services, for example, the introduction of Ukrainian citizens electronic documents, which can be applied online to: organize their own business, interact with financial institutions, pay taxes, obtain various certificates from government agencies, obtain licenses, ensure intellectual property and solve many other social issues using only one digital portal or a smartphone application. According to the results of 2019, the level of digitalization increased by 1.9% (in Slovakia and the Czech Republic the growth was 0.8%), and the most important factors were: fixed broadband subscriptions (per 100 people), fixed telephone subscriptions (per 100 people), individuals using the Internet (% of population). However, the completion of reforms in 2019, although it gave a significant impetus, still exhausted the preconditions for further rapid growth.

The growth rate of assets of digital banks significantly outpaces the growth of assets of all other commercial banks (excluding assets of digital banks). Of course, the increase in assets occurs for various reasons, but 2020 shows how the development of digital banks has accelerated the overall market due to their willingness to quickly respond to changes caused by the Covid-19 pandemic. For example, in Poland, the average increase in assets of the three most popular digital banks for the period 2015–2020 amounted to 8.7%, while the growth of assets of other commercial banks averaged 4.4%. A similar situation was observed in Slovakia, where the TOP-3 digital banks increased their assets by 6.1% over the analyzed period, while the rest of the banking sector showed an increase of only 5.4% over the same period. In the Czech Republic, the situation in the banking sector developed in the same way – the growth of assets of the monitored digital banks was almost 2 times faster than the growth of the rest of the banking sector

(growth rates of 13.6% and 6.6%, respectively). An even greater gap between the development of the TOP-3 digital banks and the rest of the market was observed in Hungary – the difference was 11.1 p.p. (the average increase in assets of the first over 5 years amounted to 14.1%, and the rest of commercial banks: 3.0%). The biggest gap between the growth rate of digital banking leaders and the rest of the banking sector was observed in Ukraine, where the asset growth rates were 22.6% and 7.2%, respectively. The growth rate of the TOP-3 digital banks in Ukraine outpaced the average market growth of the banking sector by more than 3 times for the period 2015–2020. From the observed trend, it follows that the digital banks are a powerful locomotive of the banking sector and serve as a catalyst for its development based on market changes (see Table 4).

Table 4 Assets growth in digital banks

Country	Bank	2016	2017	2018	2019	2020
Poland	PKO Bank Polski SA	7.0%	4.0%	9.2%	7.3%	7.7%
	Santander Bank Polska S.A.	7.4%	4.7%	31.5%	1.4%	9.5%
	Bank Millennium	3.9%	3.4%	13.1%	20.2%	0.3%
	Total assets. other commercial banks	4.5%	2.8%	–0.1%	4.1%	10.7%
Hungary	OTP	5.5%	16.6%	10.6%	37.9%	16.0%
	K&H	9.5%	6.5%	6.2%	10.1%	24.3%
	ERSTE Bank	5.2%	6.9%	12.8%	16.4%	26.5%
	Total assets. other commercial banks	–0.5%	–4.7%	5.0%	–13.3%	28.6%
Czech Republic	ČSOB	13.5%	21.2%	4.7%	18.4%	7.7%
	Česká spořitelna	11.1%	24.6%	7.3%	2.3%	5.4%
	MONETA Money Bank	6.7%	33.7%	3.6%	5.9%	37.4%
	Total assets. other commercial banks	8.1%	13.7%	2.1%	0.3%	8.8%
Slovak Republic	Slovenská spořitelna. a.s.	6.0%	10.2%	6.7%	6.7%	11.2%
	VÚB. a.s.	11.2%	6.7%	11.3%	5.9%	9.0%
	365.bank. a.s.	2.0%	1.3%	–1.0%	2.4%	1.5%
	Total assets. other commercial banks	4.5%	5.1%	4.1%	5.6%	7.7%
Ukraine	Universal bank | Monobank	–12.2%	22.4%	35.3%	110.9%	90.6%
	JSC CB PrivatBank	–20.7%	24.2%	9.1%	11.4%	23.5%
	Oschadbank	32.4%	11.0%	–6.9%	14.4%	–6.3%
	Total assets. other commercial banks	1.2%	0.1%	0.9%	6.4%	27.2%

Source: Annual reports – IMF, The Hungarian National Bank (21.2.2022)

Analysis of the effectiveness of the banking sector also shows the leadership of digital banks. The ratio of net profit to the bank's assets (ROA) had different dynamics for the studied six years in each country. Still, nevertheless, in almost every period, the return on assets of the TOP-3 digital bank was higher than the market average. In most countries, the gap in indicators was up to 1.7 percentage points between digital banks and the whole bank sector. The maximum gap was demonstrated by the Ukrainian Oschadbank, which managed to reach 0.3% against the backdrop of the banking sector ROA minus 12.5% in 2016 (see Table 5). In 2014, a deep financial crisis began in Ukraine, the components of which were the currency and banking crisis. The trigger for the banking crisis was the military aggression of the Russian Federation,

followed by the annexation of Crimea and the occupation of parts of the Donetsk and Luhansk regions of Ukraine. The result of military aggression was an economic recession and depreciation of the national currency, which had an extremely negative impact on the activities of banks. A high level of dollarization of banking activities (dollarization of liabilities and assets was more than 50%) against the backdrop of currency depreciation from 2014 to 2017 was one of the reasons for unprecedented losses of banks. As a result of the crisis, more than 50% of the total quantity of the banks were liquidated; namely, 33 banks in 2014, 33 banks in 2015, and 21 banks in 2016.

The situation is similar when we mention the profitability of capital: on average, the gap was up to 3 p.p. in favour of digital banks.

Indicators of Ukrainian banks in the period 2015–2016 should be viewed through the prism of the outbreak of war in the east of the country and the annexation of Crimea, which significantly affected the loss rate of banks, especially non-state ones (Universal bank | Monobank and JSC CB PrivatBank). State bank Oschadbank also suffered a loss in 2015 but to a lesser extent thanks to massive government support. Later JSC CB PrivatBank was nationalized in 2016.

Table 5 Trends in interest income margin of digital banks

Country	ROA/ ROE	Bank	2015	2016	2017	2018	2019	2020
Poland	ROA	Other commercial banks	0.8%	0.8%	0.8%	0.7%	0.7%	0.0%
		PKO Bank Polski SA	1.0%	1.1%	1.0%	1.2%	1.2%	–0.9%
		Santander Bank Polska S.A.	1.4%	1.6%	1.5%	1.4%	1.2%	0.4%
		Bank Millennium	1.3%	1.0%	0.9%	1.0%	0.7%	0.0%
	ROE	Other commercial banks	9.1%	9.2%	8.2%	7.5%	7.8%	–0.3%
		PKO Bank Polski SA	8.9%	9.2%	8.1%	9.0%	9.7%	–7.5%
		Santander Bank Polska S.A.	9.8%	11.0%	9.6%	9.7%	8.8%	3.0%
		Bank Millennium	14.0%	10.0%	9.1%	9.2%	7.1%	0.2%
Hungary	ROA	Other commercial banks	0.2%	1.6%	1.9%	1.9%	2.0%	0.8%
		OTP	0.6%	1.8%	2.3%	2.3%	2.4%	1.2%
		K&H	1.3%	1.4%	1.4%	1.8%	1.5%	0.8%
		ERSTE Bank	–1.0%	1.8%	2.5%	2.5%	2.1%	0.6%
	ROE	Other commercial banks	1.9%	16.7%	19.7%	19.4%	19.5%	7.5%
		OTP	5.1%	15.3%	18.4%	18.4%	20.3%	10.5%
		K&H	16.4%	16.7%	15.9%	20.1%	15.3%	8.7%
		ERSTE Bank	–11.8%	15.9%	17.3%	17.1%	15.1%	4.8%
Czech Republic	ROA	Other commercial banks	1.5%	1.5%	1.4%	1.3%	1.4%	0.7%
		ČSOB	1.5%	1.5%	1.5%	1.2%	1.3%	0.5%
		Česká spořitelna	1.5%	1.5%	1.2%	1.1%	1.2%	0.7%
		MONETA Money Bank	3.2%	2.8%	2.2%	2.1%	1.9%	1.0%
	ROE	Other commercial banks	12.2%	12.6%	13.7%	14.2%	15.0%	7.3%
		ČSOB	15.9%	16.9%	19.2%	17.0%	20.7%	8.3%
		Česká spořitelna	12.6%	12.8%	12.0%	12.6%	13.7%	7.0%
		MONETA Money Bank	12.8%	14.7%	14.8%	16.5%	16.2%	10.1%

Table 5 (continuation)

Country	ROA/ ROE	Bank	2015	2016	2017	2018	2019	2020
Slovak Republic	ROA	Other commercial banks	1.3%	1.4%	1.1%	1.1%	1.0%	0.7%
		Slovenská spořitelna. a.s.	1.4%	1.5%	1.1%	1.1%	1.0%	0.5%
		VÚB. a.s.	1.3%	1.2%	1.2%	1.0%	0.7%	0.4%
		365.bank. a.s.	1.3%	1.2%	1.1%	1.2%	1.1%	1.0%
	ROE	Other commercial banks	8.4%	10.0%	7.7%	7.8%	7.5%	5.1%
		Slovenská spořitelna. a.s.	13.0%	13.8%	10.6%	12.0%	11.5%	6.3%
		VÚB. a.s.	11.2%	10.4%	11.2%	9.9%	7.4%	5.0%
		365.bank. a.s.	9.3%	8.2%	7.7%	7.9%	7.2%	6.7%
Ukraine	ROA	Other commercial banks	–5.6%	–12.5%	–1.8%	1.6%	4.7%	2.8%
		Universal bank | Monobank	–32.8%	1.4%	1.8%	0.7%	4.1%	2.6%
		JSC CB PrivatBank	0.1%	–	0.2%	4.8%	11.1%	7.0%
		Oschadbank	–8.7%	0.3%	0.3%	0.1%	0.1%	1.2%
	ROE	Other commercial banks	–65.1%	–121.9%	–17.6%	11.3%	34.4%	20.0%
		Universal bank | Monobank	–276.9%	10.4%	11.3%	6.1%	41.9%	31.4%
		JSC CB PrivatBank	0.9%	–	–27.7%	46.4%	75.8%	45.3%
		Oschadbank	–92.3%	4.0%	2.4%	0.6%	1.3%	13.4%

Source: Authors' calculations based on annual reports of banks, the IMF database on Financial Soundnes Indicators, the Hungarian National Bank (21.02.2022)

CONCLUSIONS

Digital transformation is increasing its role in many fields of economic activity, involving banking as one of the most active spheres. Different legislative aspects left their footprint on traditional and digital banks, changing their presence and role in the market. However, digital banks were more demanded by households and individuals during Covid than traditional ones (assets and profitability growth are stated herein).

The digitalization index allows us to consider in a more structured way the factors that influence its changes. It also reflects changes in the level of digitalization of the banking system more precisely than the DESI approach. Analysis of underlying indices dynamics by country allowed us to conclude:

- Automated teller machines (ATMs) (per 100 000 adults): 14–40% of the contribution, depending on the country. The descending trend of ATMs in all countries is considered as a positive trend due to decreasing cash payments and an increase of electronic ones, which are the main part of digital interaction.

- Less critical but common to almost all countries are the components: Commercial bank branches (per 100 000 adults), Employment to population ratio, 15+, total (%), Fixed telephone subscriptions (per 100 people), Secure Internet servers (per 1 million people) these indices explain some infrastructure changes and the number of possible digitalization adopters.

- Factors of specific nature of influence should include those that have less weight of power, manifested in countries with different trends and rates of change: mobile cellular subscriptions (per 100 people), population density (people per sq. km of land area), individuals using the Internet (% of the population), and fixed broadband subscriptions (per 100 people). This may be caused by the achievement of a sufficient level of influence on the digitalization level, the change of which becomes possible only due to the influence of random/atypical events.

ACKNOWLEDGMENT

The authors would like to thank the participants of the Scientia Iuventa 2021 and International Scientific Conference "Contemporary Issues on Business, Management and Economics Engineering" 2021 for their valuable comments on the research.

References

AGUR, I., PERIA, S. M., ROCHON, C. (2020). *Digital financial services and the pandemic: Opportunities and risks for emerging and developing economies* [online]. International Monetary Fund Special Series on Covid-19. <https://www.imf.org/-/media/Files/Publications/covid19-special-notes/en-special-series-on-covid-19-digital-financial-services-and-the-pandemic.ashx>.

BANNA, H., ALAM, M. R. (2021). Is digital financial inclusion good for bank stability and sustainable economic development? Evidence from Emerging Asia [online]. *ADBI Working Paper Series.* <https://www.adb.org/sites/default/files/publication/692471/adbi-wp1242.pdf>.

CARLETTI, E., CLAESSENS, S., FATAS, A., VIVES, X. (2020). *The Bank Business Model in the Post Covid-19 World* [online]. IESE Business School, Centre for Economic Policy Research. <https://media.iese.edu/research/pdfs/ST-0549-E.pdf >.

CHIEW, C. J. (2020). Applied Statistical Learning in Python [online]. In: CELI, L., MAJUMDER, M., ORDÓÑEZ, P., OSORIO, J., PAIK, K., SOMAI, M. (eds.) *Leveraging Data Sci Global Health, Springer, Cham.* <https://doi.org/10.1007/978-3-030-47994-7_8>.

CODEBASE TECHNOLOGIES. (2021). *Digital Payments Report 2021* [online]. Demystifying Digital Financial Services. <https://www.codebtech.com/digital-payments-report-2021/>.

DELOITTE DIGITAL. (2020). *Digital Banking Maturity 2020. How banks are responding to digital (r)evolution?* [online]. <https://www2.deloitte.com/si/en/pages/finance/articles/digital-banking-maturity-2020.html>.

DESI. (2020). *Digital Scoreboard. Data & Indicators* [online]. <https://digital-agenda-data.eu/datasets/desi>.

DIIA. (2021). *Government services online* [online]. <https://diia.gov.ua/>.

DZIAWGO, T. (2021). Wealth Tech Impact on Wealth Management Sector. *European Research Studies*, 24(3B): 141–151.

ECB PAYMENTS STATISTICS. (2021). [online]. <https://sdw.ecb.europa.eu/reports.do?node=1000004051>.

EHRENTRAUD, J., OCAMPO, D. G., VEGA, C. Q. (2020). Regulating fintech financing: digital banks and fintech platforms [online]. *FSI Insights on policy implementation.* <https://www.bis.org/fsi/publ/insights27.pdf>.

HACHIMI, C. E., BELAQZIZ, S., KHABBA, S.,CHEHBOUNI, A. (2022). Data Science Toolkit: an all-in-one python library to help researchers and practitioners in implementing data science-related algorithms with less effort [online]. *Software Impacts*, 12(2): 100240. <https://doi.org/10.1016/j.simpa.2022.100240>.

JOHANSSON, R. (2015). *Numerical Python* [online]. Apress Berkeley, CA. <https://doi.org/10.1007/978-1-4842-0553-2>.

KITSIOS, F., GIATSIDIS, I., KAMARIOTOU, M. (2021). Digital Transformation and Strategy in the Banking Sector: Evaluating the Acceptance Rate of E-Services. *Journal of Open Innovation: Technology, Market, Complexity*, 7(3): 1–14.

LISOWSKI, J., CHOJAN, A. (2021). InsurTech in CEE Region – Where Are We? [online]. In: HOROBET, A., BELASCU, L., POLYCHRONIDOU, P., KARASAVVOGLOU, A. (eds.) *Global, Regional and Local Perspectives on the Economies of Southeastern Europe*, Springer Proceedings in Business and Economics, Springer, Cham. <https://doi.org/10.1007/978-3-030-57953-1_11>.

LUMPKIN, S., SCHICH, S. (2020). Banks, Digital Banking Initiatives and the Financial Safety Net: Theory and Analytical Framework [online]. *Journal of Economic Science Research*, 3: 24–46. <http://doi.org/10.30564/jesr.v3i1.1113>.

LYONS, A., KASS-HANNA, J. (2021). The Evolution of Financial Services in the Digital Age [online]. In: GABLE, J., CHATTERJEE, S. (eds.) *Handbook of Personal Finance*, Berlin, Germany: De Gruyter. <http://dx.doi.org/10.2139/ssrn.3873370>.

MAHESHWARI, A. (2019). *Digital Transformation: Building Intelligent Enterprises.* Hoboken, New Jersey: John Wiley & Sons.

MERGEL, I., EDELMANN, N., HAUG, N. (2019). Defining digital transformation: Results from expert interviews [online]. *Government Information Quarterly*, 36(4): 101385. <https://doi.org/10.1016/j.giq.2019.06.002>.

NATIONAL BANK OF UKRAINE. (2021). *Special Project* [online]. <https://badbanks.bank.gov.ua/#timeline>.

NATIONAL BANK OF UKRAINE. (2021). *Supervisory Data* [online]. <https://bank.gov.ua/en/statistic/supervision-statist>.

NIEMAND, T. et al. (2021). Digitalization in the financial industry: a contingency approach of entrepreneurial orientation and strategic vision on digitalization [online]. *European Management Journal*, 39(3): 317–326. <https://doi.org/10.1016/j.emj.2020.04.008>.

OUR WORLD IN DATA. (2021). *Number of COVID-19 patients in ICU per million* [online]. <https://ourworldindata.org/grapher/covid-icu-patients-per-million>.

PAZARBASIOGLU, C., MORA, A. G., UTTAMCHANDANI, M., NATARAJAN, H., FEYEN, E., SAAL, M. (2020). *Digital Financial Services.* World Bank Group [online]. <https://pubdocs.worldbank.org/en/230281588169110691/Digital-Financial-Services.pdf>.

RILEY, P., ROMORINI, S., GOLUB, E., STOKES, M. (2020). *Digital Financial Services in the MENA Region* [online]. Rockville, MD: Sustaining Health Outcomes through the Private Sector Plus Project, Abt Associates Inc. [online]. <https://shopsplusproject.org/sites/default/files/resources/ Digital%20Financial%20Services%20in%20the%20MENA%20Region.pdf>.

SHARMA, A., KANSAL, A. (2014). Technological Infovations in Banking Sector: Impact, Behaviour and Services. *International Journal of Information & Computation Technology*, 4(9): 885–890.

STATCOUNTER. (2021). *Mobile Operating System Market Share Worldwide* [online]. <https://gs.statcounter.com/os-market-share/mobile/worldwide>.

THE HUNGARIAN NATIONAL BANK. (2021). *Monetary and other balance sheet statistics* [online]. <https://www.mnb.hu/en/statistics/statistical-data-and-information/statistical-time-series/x-monetary-and-other-balance-sheet-statistics>.

VERHOEF, P.C., BROEKHUIZEN, T., BART, Y., QI DONG, J., FABIAN, N., HAENLEIN, M. (2021). Digital transformation: a multidisciplinary reflection and research agenda [online]. *Journal of Business Research*, 122: 889–901. <https://doi.org/10.1016/j.jbusres.2019.09.022>.

VERSAL, N., ERASTOV, V., BALYTSKA, M. (2021). Digital transformation in banks as a helping hand in withstanding the COVID-19 [online]. *Scientia Iuventa, Book of extended abstracts from the international scientific conference of doctoral students and young scientists*, Matej Bel University in Banská Bystrica, 41–49. <http://si.umb.sk/wp-content/uploads/2021/08/Book-of-Extended-Abstracts-SI_2021.pdf>.

VERSAL, N., ERASTOV, V., BALYTSKA, M. (2021). Is digital 'new normal' or 'challenge' for banks under COVID-19? [online]. *International Scientific Conference Contemporary Issues on Business, Management and Economics Engineering* May 13–14, Business Management Faculty, Vilnius Tech University, Vilnius. <http://cibmee.vgtu.lt/index.php/verslas/2021/paper/viewFile/608/259>.

VIAL, G. (2019). Understanding digital transformation: a review and a research agenda [online]. *The Journal of Strategic Information Systems*, 28(2): 118–144. <https://doi.org/10.1016/j.jsis.2019.01.003>.

WEWEGE, L. THOMSETT, M. C. (2019). *The Digital Banking Revolution: How Fintech Companies are Transforming the Retail Banking Industry Through Disruptive Financial Innovation* [online]. Berlin, Boston: De Gruyter. <https://doi.org/10.1515/9781547401598>.

WORLD BANK AND CAMBRIDGE CENTRE FOR ALTERNATIVE FINANCE. (2020). *The Global COVID-19 FinTech Regulatory Rapid Assessment Study* [online]. World Bank Group, Washington, DC and Cambridge, UK. <https://www.jbs.cam.ac.uk/wp-content/uploads/2020/10/2020-ccaf-report-fintech-regulatory-rapid-assessment.pdf >.

WORLD BANK. (2021). *World Bank Open Data* [online]. <https://data.worldbank.org>.

ZAOUI, F., SOUISSI, N. (2020). Roadmap for digital transformation: a literature review. *Procedia Computer Science*, 17: 621–628.

Churn Prediction for High-Value Players in Freemium Mobile Games: Using Random Under-Sampling

Guan-Yuan Wang[1] | *Vilnius University, Vilnius, Lithuania*

Received 7.4.2022, **Accepted** (reviewed) 8.6.2022, **Published** 16.12.2022

Abstract

Many game development companies use game data analysis for mining insights about users' behaviour and possible product growth. One of the most important analysis tasks for game development is user churn prediction. Effective churn prediction can help hold users in the game by initiating additional actions for their engagement. We focused on high-value user churn prediction as it is of particular interest for any business to keep paying customers satisfied and engaged. We consider the churn prediction problem as a classification problem and conduct the random under-sampling approach to address imbalanced class distribution between churners and active users. Based on our real-life data from a freemium casual mobile game, although the best model was chosen as the final classification algorithm for extracted data, we can definitely say there is no general solution to the stated problem. Model performance highly depends on the churn definition, user segmentation and feature engineering, it is therefore necessary to have a custom approach to churn analysis in each specific case.

Keywords	***DOI***	***JEL code***
Churn prediction, mobile games, classification models, resamlpling methods, imbalanced class distribution, machine learning	*https://doi.org/10.54694/stat.2022.18*	*C10, C53, M20, M30*

INTRODUCTION

In the mobile gaming industry, data mining and machine learning methods are widely applied to solve various business problems. The two primary reasons for this are the availability of vast amounts of records on players' in-game behaviour (essentially, every action players make is being recorded) and a limited amount of direct communication with players, which requires understanding their needs by other methods. The majority of mobile games nowadays operate under the so-called "freemium" business model, which implies that players can download the game and play completely free of charge but are offered a range

[1] Faculty of Mathematics and Informatics, Vilnius University, Universiteto g. 3, Vilnius 01513, Lithuania. E-mail: guan-yuan.wang@tprs.stud.vu.lt, phone: (+44)7766387130.

of available in-game purchases that are supposed to enhance gaming experiences. In such a context, player retention is one of the core business problems for game developers, since abandoning the game is costless and effortless for a player, while typically it is more costly for the company to acquire new players than to maintain the current ones. Therefore, one of the popular applications of machine learning in the industry is player churn prediction, which can be generalized as predicting which players, when and with what probability will abandon the game. On a general level, such predictions help a company to develop better business plans and forecasts and grasp a better understanding of its business, while a more specific application would be to directly address the potential churners identified early on and try to retain them in the game.

In this paper, we use a dataset from a casual freemium mobile game to develop a churn probability prediction model for high-value players.

1 LITERATURE REVIEW

1.1 Churn definition

The first issue that is crucial for churn detection in a freemium environment is how churn is actually determined. The churn prediction problem heavily depends on the definition of churn, so different labelling approaches for churn yield different results. Since in the games operating under the freemium business model there is no formal event of discontinuing participation, there is no clear signal that a player has churned. Therefore, a typical approach is to consider a player churned if he does not login into the game for a prespecified period of time. The periods considered for identifying churners range vastly among studies. While some researchers focused on predicting player retention on a specific day (Drachen et al., 2016; Fu et al., 2017), others considered periods of inactivity from 9 days (Bertens et al., 2017) through 14 days (Kristensen and Burelli, 2019) to 13 weeks (Lee et al., 2020) as churn signals. There seems to be a trade-off between the desire to identify churners as early as possible and the risk of misclassification. Hence, several approaches were suggested for empirical identification of the appropriate period taking this trade-off into account.

Bertens et al. (2017) and Guitart et al. (2019) tuned the churn signal period by calculating the percentages of false churners and missed sales (revenue from false churners) for different periods and choosing a period which resulted in less than 5% false churners and 1% of missed sales. A simpler approach was suggested by Runge et al. (2014), who computed a frequency distribution of gaps between logins and chose a period corresponding to the 98% percentile of this distribution. Lee et al. (2020) used the game development cycle (26 weeks) as a starting point to calculate the period for when a user is considered to have churned, arguing that a user who does not respond to the major updates should be considered lost. The final churn period (13 weeks) was based on a compromise between certainty and profit; the compromise was that among the players defined as churners, 25% are not going to churn according to the definition of the game development cycle (26 weeks). Rothmeier et al. (2020) discussed four disparate approaches: naive approach, sliding windows approach, quartile approach and trend over varying dates approach on churn detection issues.

1.2 Segmentation

Another issue highlighted by several researchers is that the player base usually is not homogeneous, so segmentation might be needed to achieve better business outcomes and prediction accuracy. From a business perspective, several researchers highlight the fact that not all customers are equally valuable to the company and hence, it is reasonable to focus the retention management and churn prediction modelling effort specifically on the high-value customer segment (Liu et al., 2019; Runge et al., 2014; Lee et al., 2020, e.g.). Typically, high-value customers are defined as those bringing most of the revenue (Runge et al., 2014); however, an important point was made by Liu et al. (2019) that in the social gaming

context customer's experience with the game and his/her social influence in it might serve as additional measures of customer's worth that should not be neglected. Moreover, segmenting customers based on their in-game behaviour and modelling churn for different groups separately might improve the accuracy of the predictions (Liu et al., 2019; Fu et al., 2017).

1.3 Feature selection

Careful feature variable selection is another way considered by researchers to improve the effect of customer churn prediction. The more powerful feature variables are, the better the customer churn prediction effect is. Ascarza et al. (2018) provided an overview of customer retention management research and identified usage behaviour, user characteristics, satisfaction and social connectivity as the most common predictors in churn modelling. However, since there is often little information available about the players and direct marketing surveys, which could provide data about players' satisfaction are not widespread, in the gaming industry churn modelling mostly has to rely on user behaviour and social connectivity data alone. The user behaviour domain can be considered as consisting of two primary subcategories: in-game activity data (what players do in the game) and engagement data (how frequently and intensively players do). Thus, most of the research we considered used a combination of in-game activity data, engagement data and in-game social interaction data.

In-game activity (also called user performance (Fu et al., 2017) was operationalized with such variables as rounds played, in-game currency balance (Runge et al., 2014), level, number of quests completed, coins collected (Lee et al., 2020), experience points gained, number of quests, number of characters controlled, average character level, number of levels advanced etc. (Borbora and Srivastava, 2012).

User engagement was operationalized as days in the game, time series of logins, last purchase, days since last purchase (Runge et al., 2014), total inter-session length (Borbora and Srivastava, 2012), login frequency, length of login time and average playtime (Fu et al., 2017). In regards to time series data of user activity, Migueis et al. (2012) measured the similarity of the sequence of customers' first purchases to model customer churn, exploring the predictive power of the likelihood of the first product category purchase sequence made by a new customer to identify whether they are churners or not. Although this was done in a retail setting, a similar approach can be introduced in the gaming industry as well to specifically predict the churn of new users.

The last group of features is in-game social interaction features, which include, but are not limited to such variables as invites sent (Runge et al., 2014), friend quitting a game, reduction of the legion (Lee et al., 2020), rate of group interactions, number of churners interacted with (Borbora and Srivastava, 2012), number of in-game friends, whether or not player joined a guild and guild role (Fu et al., 2017). Considering social interaction features, it is worth mentioning that besides directly using players' social interaction data as predictors, it can be used to identify interconnected players and model the cross-effects they might have. For instance, Liu et al. (2019) used players' in-game interaction to identify the most influential social neighbours for each player and showed that introducing the effect of neighbours' features into the player's churn prediction model might improve the model accuracy in some cases.

The last group of features is in-game social interaction features, which include, but are not limited to such variables as invites sent (Runge et al., 2014), friends quitting a game, reduction of the legion (Lee et al., 2020), rate of group interactions, number of churners interacted with (Borbora and Srivastava, 2012), number of in-game friends, whether or not player joined a guild and guild role (Fu et al., 2017). Considering social interaction features, it is worth mentioning that besides directly using players' social interaction data as predictors, it can be used to identify interconnected players and model the cross-effects they might have. For instance, Liu et al. (2019) used players' in-game interaction to identify the most influential social neighbours for each player and showed that introducing the effect of neighbours' features into the player churn prediction model might improve the model accuracy in some cases.

Another group of predictors that emerges from the research is RFM-based (Yeh et al., 2009; Liu et al., 2019; Rahim et al., 2021). This group of predictors is somewhat similar to user engagement as it consists of login and purchase data: recency and frequency of the usage and monetary value spent in the game. Besides just selecting the features that seem most meaningful out of the data available, some researchers engage in designing complex predictors of their own out of the initial data. An example of such an approach is provided by Liu et al. (2019), who designed a complex single measure of what they called "Activity Energy" of the player out of the data available and used the trend in it as the main predictor in their models. Interestingly, Borbora and Srivastava (2012) compared data-driven and theory-driven models with different approaches to variable selection and suggested that even though theory-driven models may be less accurate, they can be better interpretable and, therefore, more preferable over complex data-driven models.

1.4 Churn modelling

Up to this point, we outlined several important issues that might be useful to consider prior to churn modelling identified within the prior research. Next, we briefly consider modelling approaches that are being used for churn prediction. The majority of studies treat churn prediction as a binary classification problem (Hadiji et al., 2014; Runge et al., 2014; Xie et al., 2015; Fathian et al., 2016; Kim et al., 2017; Liu et al., 2019; Lee et al., 2020). In other words, their goal was to predict whether a player will churn at some particular point in time or, equivalently, to identify those players who will churn on a specific day. A wide variety of classification models and algorithms were considered for this purpose, including KNN, ADTreesLogit, Random Forest, Naive Bayes, Neural Networks, Logistic Regression, Decision Trees and Support Vector Machines. Binary classification output makes it easy to compare performances of the models using such metrics as Area under the ROC Curve (AUC) or F-score, so typically several models are compared. However, for business applications, it might be more useful to predict specific churn probabilities in order to distinguish 'on-the-edge' churners who might be responsive to additional marketing efforts from the definite ones (Ascarza et al., 2018).

Despite its popularity, binary classification is not the only way to frame the churn prediction problem. As an alternative, it might be approached with time-series modelling methods (del Río et al., 2021) or survival analysis techniques (Bertens et al., 2017; Guitart et al., 2018). Time-series modelling provides a more general take on the issue compared to the classification and survival analysis approaches. Del Río et al. (2021) used time-series state-space models (ARIMA and Unobserved Components) to model the transition and retention time series for churned, paying, and non-paying user groups. Such an approach allows predicting how many users are expected to churn in a given period of time but does not provide any insights on which users it might be. On the opposite side, survival analysis approach might be seen as the most detailed out of three, since it models a survival curve for each player using such methods as survival ensembles, Cox regression or Kaplan-Meier Model (Guitart et al., 2018). Another aspect worth mentioning is that very few studies considered ensemble learning methods. Fathian et al. (2016) found that classifier ensembles perform better than single classifiers, and boosting methods are superior to bagging, while Guitart et al. (2018) showed that survival ensembles outperform single Cox regression in a survival analysis context.

2 DATA

In this research, we are modelling player churn in a mobile farm game context. The game is available on Android and iOS platforms and has more than 10 000 000+ installs on Android alone. It operates under the freemium business model, which implies that the game is technically free, but users have an opportunity to purchase certain things within the game that are supposed to enhance the gaming experience.

2.1 Sample selection and main definitions

Since we have access to all the user's log data collected in the game, our first goal is to collect a dataset for modelling. The goal of this research is to predict churn of the high-value users, so these concepts have to be operationalized. High-value users were defined as top-paying users, who cumulatively contributed 50% of total revenue in the 90 days from the observation date. Setting the observation date on August 30, 2021, we identified 34 769 high-value users. In order to identify an appropriate period of inactivity to be used as a signal of churn, the distributions of gaps between logins for these high-value users were investigated. Two particular distributions were considered: the distribution of all the gaps between logins for these users and the distribution of maximum gaps between logins. The upper quantiles are presented in Table 1.

Table 1 Distribution of gaps between logins

Quantile	50%	61%	90%	95%	98%	99.5%
All gaps (days)	0	0	1	2	4	4
Max gaps (days)	8	14	71	116	215	468

Source: Own construction

These two approaches demonstrate rather different pictures. Overall, the results suggest that high-value users typically do not take considerable breaks from the game: when they play it, they play it almost every day, as demonstrated by all gaps distribution. On the other hand, some high-value users seem to abandon the game for a long time but come back afterwards. For example, the 90% quantile of max gaps of 71 days suggests that 10% of users under consideration returned to the game after taking a break for more than two months. These results highlight an interesting dilemma for the churn definition in freemium mobile games. On one hand, the interval of inactivity used for churn labelling should be highly unlikely in order to keep the number of misclassified users low. On the other hand, taking a period of months would be rather meaningless from both business and modelling perspectives, since for such a period, it first would be too long to act upon and for the second, we would end up with too few potential churners labelled. Therefore, we chose a 14-day period corresponding to the 99.5% quantile of all gaps and 61% quantile of max gaps as a churn labelling inactivity period. In other words, a user who does not log into the game for 14 consecutive days is considered to be churned. While 39% of users would actually log into the game at least one more time after being labelled churned according to such a definition, the probability of such a gap is only 0.5% as suggested by all gaps distribution. These statements may appear contradictory at first, but this contradiction might be resolved by accepting a less strict churn interpretation and the concept of "returnees" or the users, who churn at one point, but come back into the game later on. Following this definition of churn, we included into the final sample 22 547 high-value users who would be considered active on the observation date and labelled 2 676 (11.87%) of those who churned right afterwards as churned.

2.2 Features extraction

For the selected sample of users, we extract the set of predictors to be used for modelling from the raw players' log data. In order to capture the trend in the user's behaviour, the features are constructed for weekly time intervals 12 weeks before the observation date. In other words, a single feature, for example, the total number of logins is split into 12 predictors: number of logins in the week prior to the observation date, two weeks prior to the observation date, etc. The idea behind such an approach is that the dynamics of a user's past behaviour, rather than the behaviour itself, is expected to hold most of the necessary signals to predict player churn and we would like to capture it in our models. The final set of features is extracted including information about:

- User's logins in the previous 12 weeks: total number of logins, average session duration, total time spent in the game.
- User's payments in the previous 12 weeks: sum of payments, number of payments.
- User's in-game activity in the previous 12 weeks: number of tasks completed, number of achievements received, average time spent on completing a task.
- User's progress in the game in the previous 12 weeks: maximum level achieved, additional levels achieved
- User's profile data: platform (device OS), days since installed, source of installation.

Overall, 141 predictors were extracted.

3 METHODOLOGY

In order to evaluate the performance of the models trained, the dataset is split into training and test sets by allocating 70% of observations to the training set and the rest for testing. Since the prevalence of churners is relatively low, we ensure that they are equally represented in both sets. Training dataset consist of 15 782 observations, including 1 872 (11.86%) churners and 13 910 active users, while the test dataset consist of 6 765 observations: 804 (11.88%) churners and 5 961 active users. It bespeaks that there is an imbalanced class distribution in our dataset. Specifically, the sample size of churners s far smaller than that of non-churners, which can cause high overall classification accuracy but low classification accuracy of churners. In practice, this misclassification of churners usually causes heavier economic losses, especially in our case study of high-value players.

In order to address the imbalanced data issue (11.87% of users are labelled as churners), we employ the Random Under-Sampling (RUS) approach. By undersampling without replacement of the majority class, we end up with a sample of size 3 733 (1 872 churners and 1 861 active users). Undersampling is selected due to higher prediction accuracy and sensitivity compared to oversampling. As a secondary benefit, undersampling of the training set allows us to lower the computational resources needed for training models. Van Hulse et al. (2007), Seiffert et al. (2014), Bauder et al. (2018) and Xiao et al. (2021) show that random undersampling significantly leads to improving the models' performance for datasets with the issue of severe imbalanced class.

3.1 Modelling

3.1.1 Regularized Discriminant Analysis (RDA)

Regularized Discriminant Analysis is a flexible classification technique serving as a bridge between the stricter Linear and Quadratic Discriminant Analyses. However, it imposes certain requirements on the initial data for the model to fit. One such requirement is the invertibility of the covariance matrices for each class of the response variable. In our case, the initial dataset violated this requirement, however, the issue was solved by Principal Component Analysis (PCA) pre-processing of the scaled and centred data. Given that RDA performs the implicit feature selection and PCA is a dimension reduction technique, in this case, the two-step predictor selection was performed. The two RDA parameters: lambda and gamma were tuned over the range between the extreme values of 0 and 1 with step 0.1. The optimal parameters based on the AUC were chosen to be lambda = 1 (essentially assuming the QDA covariance matrices) and gamma = 0.1. Monte-Carlo cross-validation with 25 groups was used to ensure the model's reliability.

3.1.2 Neural Networks (NN)

Neural Networks are powerful and rather universal models; however, they are also computationally demanding to train. In this research, the final NN model was selected in two steps. First, the multilayer structure was investigated by comparing the models with 1 to 3 layers, 1 to 5 nodes in the first layer and 0 to 5 in the 2nd and 3rd and 0 to 2 decay parameters. Since many tuning parameters were considered

in this step, a simple 3-fold cross-validation was used in this step in order to decrease the computational burden and the models were trained on the under-sampled training set. The best model was chosen based on the AUC performance metric and turned out to be a single-layer network with a single node and 0 decay. Therefore, in the second step, just the single-layer models were evaluated over the same ranges of the number of nodes (1 to 5) and decay (0 to 2), but with much more rigorous Monte-Carlo 25 group cross-validation. The final model was chosen based on the AUC and similarly to the previous step included just one node and small decay of 0.1. In both steps, the models were trained on the scaled, centred and spatial sign-transformed data.

3.1.3 K-Nearest Neighbours (KNN)

KNN is one of the simplest classification models, so it is often used as a benchmark for other models to be compared against. The KNN was trained on the centred and scaled under-sampled training set using Monte-Carlo Cross-Validation with 25 groups. The tuning parameter K (number of neighbours) was tuned based on the ROC curve as a performance metric and was set to 43 in the final model.

3.1.4 Logistic regression

Logistic regression is a simple statistical model that uses the logistic function to model a binary dependent variable, although more complex extensions exist. The logarithm of the odds for positive class (churn in this case) is a linear combination of one or more categorical or continuous variables. Here we trained the model on the centred and scaled under-sampled training set.

3.1.5 Random forest

Random forest is a machine learning algorithm that combines multiple decision trees to reach a single result. The basic idea of the algorithm is to gather information from multiple decision trees and select variables and threshold values that maximize classification accuracy. Decision trees, on the other hand, are built by recursively evaluating different features and at each node using the feature that best splits the data. Random forest is a tree-based model and hence does not require feature scaling (contrary to distance-based models). In addition, it automatically detects variables that are important and ignores less important ones; therefore, all predictors were used for modelling. The model was trained on the under-sampled training set. The final model had 500 trees and used 11 predictors for splitting at each node.

3.1.6 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a linear model for classification and regression problems. The idea of SVM is simple: the algorithm creates a line or a hyperplane which separates the data into classes - class boundaries. In this work, we used non-linear SVM. Boundaries in this kind of SVM don't have to be a straight line. It allows capturing more complex relationships between data points, but training time for such a kind of SVM is longer as it's much more computationally intensive. SVM has two main parameters: C (regularization parameter) and Gamma. C parameter controls the tradeoff between smooth decision boundary and classifying training points correctly. Gamma, in its turn, defines how far the influence of a single training example reaches. If it has a low value it means that every point has a far reach and conversely high value of gamma means that every point has a close reach. In the final model, C = 1 and Gamma is inversely proportional to the number of features.

3.1.7 XGBoost (XGB)

XGBoost (Extreme Gradient Boosting) is a decision-tree-based ensemble algorithm that uses gradient boosting. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. The training

of the model proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It uses a gradient descent algorithm to minimize the loss when adding new models. In the model were used 300 trees (n estimators = 300) with maximum depth equals 6 (max depth = 6).

3.1.8 Classifier Ensemble

Besides training single learning algorithms, we considered the Ensemble model which consists of KNN, RDA, NN, Random Forest and logistic regression models. Three types of ensembles are used: majority voting, probability average and weighted probability average. For the weighted probability average, since RDA with the highest sensitivity and random forest with the highest AUC score, churn probabilities for these two models were multiplied by two times compared to the other models, which is by 0.286 and the rest by 0.143.

4 RESULTS

Before analysing the results, let us know how to evaluate and choose our model. In this study, we present five criteria: AUC, accuracy, sensitivity, specificity, and F1-score. Accuracy measures how many observations, both positive and negative, are correctly classified. AUC means the area under the ROC, which is defined as a trade-off between the true positive rate and the false positive rate, characterized as a curve corresponding to each threshold. Sensitivity is the ability of a test to correctly identify an observation with a positive result (in our case, churner). Specificity is the ability of a test to correctly identify people with a negative result (non-churner). F1-score combines sensitivity and specificity into one metric by calculating the harmonic mean between those two criteria. AUC and F1-score both are robust evaluation metrics but researchers are more commonly using AUC as the criterion. With that said, we also presented the F1-score into our list for interested readers. While the AUC is often used as a single comparison measure, it does not account for the cost of any particular type of misclassification. At the same time, in the context of churn modelling, the cost of the missed churner is often much higher than the cost of incorrectly treating an active user as a potential churner. Therefore, the second metric we should pay attention to is sensitivity, or a model's ability to correctly classify the churners. Lastly, we

Table 2 Model evaluation metrics

Model	AUC	Accuracy	Sensitivity	Specificity	F1-score
KNN	0.844	0.6922	0.8756	0.6675	0.4034
RDA	0.895	0.7418	0.8955	0.7210	0.4518
NN	0.906	0.8306	0.8545	0.8274	0.5452
Random forest	0.930	0.8661	0.8632	0.8665	0.6051
Logistic regression	0.882	0.7938	0.8408	0.7874	0.4921
SVM	0.912	0.8017	0.8462	0.4181	0.7315
XGB	0.915	0.9178	0.6405	0.6585	0.6494
Ensemble (majority voting)		0.8085	0.8939	0.7969	0.5259
Ensemble (probability average)	0.917	0.8142	0.8914	0.8038	0.5328
Ensemble (weighted probability average)	0.920	0.8194	0.8989	0.8087	0.5420
Random forest (imbalanced data)	0.929	0.9172	0.6019	0.9597	0.6335

Source: Own construction

should not use accuracy on imbalanced problems. It is easy to obtain a high accuracy score by simply classifying all observations as the majority class, which is shown in our result.

Table 2 presents the comparison of the main classifiers' performance metrics for the trained models. The sensitivity and specificity are computed for the standard threshold of 50% and treating predicted churn as a positive case. In addition, ROC relies on the concept of the adjustable threshold in order to adjust the trade-off between sensitivity and specificity. Majority voting is just used for classification by considering all models which are already constructed. There is no parameter we can adjust; therefore, we cannot calculate the AUC for this method. Considering both metrics, AUC and sensitivity, the top-performing models are Random Forest, Neural Network and the Classifier Ensembles, especially the weighted probability average one. The Classifier Ensembles managed to improve slightly in sensitivity over the Random Forest, but at the cost of specificity. Comparing the classifiers' performance in the context of highly unbalanced classes might not be exactly straightforward. We showed that if we classify the imbalanced data, we will obtain relatively low sensitivity even using the best model, Random Forest. As we discussed earlier, that would bring inevitably massive economic losses, but the random under-sampling approach indeed helps us address the imbalanced class problem.

CONCLUSION

In this paper, we considered a problem of user churn prediction in the mobile freemium game context, provided an example of such predictive modelling and compared the performance of different classification models. While in general, the concept of user churn is relatively straightforward, its operationalization becomes rather challenging and nuanced in a highly unstructured environment of modern mobile gaming. The issue lying at the core of the problem is when the user should be considered churned in the first place. The approach utilized in the current paper is based on the analysis of the distribution of the intervals between logins of the target users provides an objective foundation for churn definition; however, it still does not result in a single correct answer and the modelling results might differ substantially based on the definition chosen. Thus, we conclude that in the practical application a variety of churn definitions should be investigated and the one which produces the best business results should be used.

Another important point that should be highlighted is that it is often reasonable to model churn separately for different segments of users. The first part of this idea is that from the business perspective not all users are equally important, so it is reasonable to focus on the high-value ones and tailor the models towards the specifics of the behaviour of this particular group as it was done in the current research. The second part worth mentioning although it was not directly utilized in the practical part of this research is that further user segmentation based on some user's behavioural profiles (for instance, user's level of social connectedness in the game) might also improve the churn prediction results as it was highlighted in the literature review.

From the modelling side, several conclusions can be drawn from the current research. First of all, we would like to stress that class imbalance is an inherent issue in churn modelling because at any point in time only a small portion of users are expected to churn. At the same time, from the business perspective, the sensitivity of the model or, in other words, its ability to correctly classify soon churners, is more important than general model accuracy, since it is typically more expensive for the company to miss the churner than to spend some extra resources on encouraging an active user. Thus, resampling the training set might be utilized to balance the classes. In the case of current research, undersampling of the majority class (active users) produced better results than oversampling of the underrepresented class of churners, while also lowering the computational burden by decreasing the size of the training set, which might be desirable in some cases.

As for the models' performance, the Random Forest algorithm yielded the best results in the current research, while the simplest KNN expectedly showed the worst performance. The combination of different

models into the classifier ensemble did not yield significant performance improvement over the single Random Forest as was expected. Perhaps, the reason for that was that to achieve performance improvement by combining several different classifiers, they should perform approximately equally in the first place.

Finally, this work certainly has lots of limitations and is not a final say in the churn modelling. It provides an overview of the key issues one should consider for the practical churn modelling in the mobile game environment, but the practical results will highly depend on the specific choices of the churn definition, user segmentation approach, features extracted, algorithms used, etc. Thus, it is a careful and consistent choice and adjustment of all the factors discussed that shall lead to the practically applicable model.

References

ASCARZA, E., NESLIN, S. A., NETZER, O., ANDERSON, Z., FADER, P. S., GUPTA, S., HARDIE, B. G. S., LEMMENS, A., LIBAI, B., NEAL, D., PROVOST, F., SCHRIFT, R. (2018). In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions [online]. *Customer Needs and Solutions*, 5(1–2): 6–81. <http://doi.org/10.1007/s40547-017-0080-0>.

BAUDER, R. A., KHOSHGOFTAAR, T. M., HASANIN, T. (2018). Data sampling approaches with severely imbalanced big data for medicare fraud detection [online]. *2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI)*, 137–142. <http://doi.org/10.1109/ICTAI.2018.00030>.

BERTENS, P., GUITART, A., PERIANEZ, A. (2017). Games and big data: a scalable multidimensional churn prediction model [online]. *2017 IEEE Conference on Computational Intelligence and Games, CIG*, 33–36. <http://doi.org/10.1109/CIG.2017.8080412>.

BORBORA, Z. H., SRIVASTAVA, J. (2012). User Behavior Modelling Approach for Churn Prediction in Online Games [online]. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 51–60. <http://doi.org/10.1109/SocialCom-PASSAT.2012.84>.

DEL RÍO, A. F., GUITART, A., PERIANEZ, A. (2021). A time series approach to player churn and conversion in videogames [online]. *Intelligent Data Analysis*, 25(1): 177–203. <http://doi.org/10.3233/IDA-194940>.

DRACHEN, A., LUNDQUIST, E. T., KUNG, Y., RAO, P., SIFA, R., RUNGE, J., KLABJAN, D. (2016). Rapid prediction of player retention in free-to-play mobile games. *Twelfth artificial intelligence and interactive digital entertainment conference.*

FATHIAN, M., HOSEINPOOR, Y., MINAEI-BIDGOLI, B. (2016). Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods [online]. *Kybernetes*, 45(5): 732–743. <http://doi.org/10.1108/K-07-2015-0172>.

FU, X., CHEN, X., SHI, Y. T., BOSE, I., CAI, S. (2017). User segmentation for retention management in online social games [online]. *Decision Support Systems*, 101: 51–68.<http://doi.org/10.1016/j.dss.2017.05.015>.

GUITART, A., CHEN, P., PERIANEZ, A. (2018). The Winning Solution to the IEEE CIG 2017 Game Data Mining Competition [online]. *Machine Learning and Knowledge Extraction*, 1(1): 252–264. <http://doi.org/10.3390/make1010016>.

GUITART, A., DEL RIO, A. F., PERIANEZ, A. (2019). Understanding player engagement and in-game purchasing behavior with ensemble learning [online]. *20th International Conference on Intelligent Games and Simulation, GAME-ON 2019*, 1: 78–85. <https://doi.org/10.48550/arxiv.1907.03947>.

HADIJI, F., SIFA, R., DRACHEN, A., THURAU, C., KERSTING, K., BAUCKHAGE, C. (2014). Predicting player churn in the wild [online]. *2014 IEEE Conference on Computational Intelligence and Games*, 1–8. <https://doi.org/10.1109/CIG.2014.6932876>.

KIM, S., CHOI, D., LEE, E., RHEE, W. (2017). Churn prediction of mobile and online casual games using play log data [online]. *PLoS ONE*, 12(7): 1–20. <http://doi.org/10.1371/journal.pone.0180735>.

KRISTENSEN, J. T., BURELLI, P. (2019). Combining sequential and aggregated data for churn prediction in casual freemium games [online]. *2019 IEEE Conference on Games (CoG)*, 1–8. <http://doi.org/10.1109/CIG.2019.8848106>.

LEE, E., KIM, B., KANG, S., KANG, B., JANG, Y., KIM, H. K. (2020). Profit optimizing churn prediction for long-term loyal customers in online games [online]. *IEEE Transactions on Games*, 12(1): 41–53. <http://doi.org/10.1109/TG.2018.2871215>.

LIU, D. R., LIAO, H. Y., CHEN, K. Y., CHIU, Y. L. (2019). Churn prediction and social neighbour influences for different types of user groups in virtual worlds [online]. *Expert Systems*, 36(3): 1–20. <http://doi.org/10.1111/exsy.12384>.

MIGUEIS, V. L., POEL, D. V., CAMANHO, A. S., CUNHA, J. F. (2012). Predicting partial customer churn using markov for discrimination for modeling first purchase sequences [online]. *Advances in Data Analysis and Classification*, 6(4): 337–353. <https://doi.org/10.1007/s11634-012-0121-3>.

RAHIM, M. A., MUSHAFIQ, M., KHAN, S., ARAIN, Z. A. (2021). RFM-based repurchase behavior for customer classification and segmentation [online]. *Journal of Retailing and Consumer Services*, 61(C). <http://doi.org/10.1016/j.jretconser.2021.102566>.

ROTHMEIER, K., PFLANZL, N., HÜLLMANN, J. A., PREUSS, M. (2020). Prediction of player churn and disengagement based on user activity data of a freemium online strategy game [online]. *IEEE Transactions on Games*, 13(1): 78–88. <http://doi.org/10.1109/TG.2020.2992282>.

RUNGE, J., GAO, P., GARCIN, F., FALTINGS, B. (2014). Churn prediction for high-value players in casual social games [online]. *IEEE Conference on Computatonal Intelligence and Games, CIG.* <http://doi.org/10.1109/CIG.2014.6932875>.

SEIFFERT, C., KHOSHGOFTAAR, T. M., VAN HULSE, J., FOLLECO, A. (2014). An empirical study of the classification performance of learners on imbalanced and noisy software quality data [online]. *Information Sciences*, 259: 571–595. <https://doi.org/10.1016/j.ins.2010.12.016>.

VAN HULSE, J., KHOSHGOFTAAR, T. M., NAPOLITANO, A. (2007). Experimental perspectives on learning from imbalanced data [online]. *Proceedings of the 24th international conference on Machine learning*, 935–942. <https://doi.org/10.1145/1273496.1273614>.

XIAO, J., WANG, Y., CHEN, J., XIE, L., HUANG, J. (2021). Impact of resampling methods and classification models on the imbalanced credit scoring problems [online]. *Information Sciences*, 569: 508–526. <https://doi.org/10.1016/j.ins.2021.05.029>.

XIE, H., DEVLIN, S., KUDENKO, D., COWLING, P. (2015). Predicting player disengagement and first purchase with event-frequency based data representation [online]. *2015 IEEE Conference on Computational Intelligence and Games (CIG)*, 230–237 <http://doi.org/10.1016/10.1109/CIG.2015.7317919>.

YEH, I. C., YANG, K. J., TING, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence [online]. *Expert Systems with Applications*, 36(3 PART 2): 5866–5871. <http://doi.org/10.1016/j.eswa.2008.07.018>.

A New Viterbi-Based Decoding Strategy for Market Risk Tracking: an Application to the Tunisian Foreign Debt Portfolio During 2010–2012

Mohamed Saidane[1] | *Qassim University, Buraidah, Kingdom of Saudi Arabia*

Received 4.4.2022 (revision received 6.6.2022), **Accepted** (reviewed) 8.6.2022, **Published** 16.12.2022

Abstract

In this paper, a novel market risk tracking and prediction strategy is introduced. Our approach takes volatility clustering into account and allows for the possibility of regime shifts in the intra-portfolio's latent correlation structure. The proposed specification combines hidden Markov models (HMM) with latent factor models that takes into account the presence of both the conditional skewness and leverage effects in stock returns.

A computationally efficient expectation-maximization (EM) algorithm based on the Viterbi decoder is developed to estimate the model parameters. Using daily exchange rate data of the Tunisian dinar versus the currencies of the main Tunisian government's creditors, during the 2011 revolution period, the model parameters are estimated. Then, the suitable model is used in conjunction with a Monte Carlo simulation strategy to predict the Value-at-Risk (VaR) of the Tunisian government's foreign debt portfolio. The backtesting results indicate that the new approach appears to give a good fit to the data and can improve the VaR predictions, particularly during financial instability periods.

Keywords	***DOI***	***JEL code***
Factor analysis, volatility clustering, hidden Markov models, Viterbi-EM algorithm, portfolio's Value-at-Risk	*https://doi.org/10.54694/stat.2022.17*	*C38, C53, G17, G32*

INTRODUCTION

According to Saidane (2017) and Mosbahi et al. (2017), the understanding of co-movements among asset returns is a central element in the portfolio risk management process. The authors advocate the use of a mixture of probabilistic factor analyzers and the conditionally heteroskedastic latent factor model to handle co-movements, heterogeneity and time-varying volatility embedded in financial data. They

[1] Department of Management Information Systems and Production Management, College of Business and Economics, Qassim University, P.O. Box 6666, Buraidah 51452, Kingdom of Saudi Arabia. E-mail: M.Saidane@qu.edu.sa.

demonstrate how their proposed strategies can be applied to the estimation of the portfolio's Value-at-Risk (VaR). However, an assumption of these models is that the correlation structure of the portfolio is assumed to be constant over time, but recent empirical works (e.g. Saidane, 2019; Tsang and Chen, 2018; Hamilton, 2016; Ang and Timmermann, 2012) have shown that this assumption of structural stability is invalid for financial returns, especially during crisis periods. For example, when the economy is hit by a permanent or temporary exogenous unpredictable shock, the cross-correlation behavior among several financial assets and the inter-relationship between volatilities can be expected to shift simultaneously. In light of this, we propose a novel market risk prediction strategy considering the possibility of regime switching in the interrelationships among several asset classes.

The new approach presented in this paper allows for the possibility of regime shifts in the intra-portfolio's latent correlation structure and takes volatility clustering into account. The proposed specification combines latent factor models that takes into account the presence of both the conditional skewness and leverage effects with hidden Markov models (HMM). To capture the volatility clustering and the leverage effect patterns of the return series, we assume that the common variances are modeled separately using quadratic generalized autoregressive conditionally heteroskedastic (GQARCH) processes. This provides a more tractable way to handle the time-varying volatility, co-movements and the latent heterogeneity in financial data.

For the maximum likelihood estimation we proceed in two steps. In the first step, we use the Viterbi decoding algorithm to find the most probable path through the HMM, given the observed data, which we take as an estimate of the true path. In the second step, we implement the Expectation-Maximization (EM) algorithm introduced by Dempster et al. (1977), to estimate the model parameters. Our proposed estimation strategy overcomes the complexity and limitations of the exact learning algorithm, especially when the number of hidden states and the length of the time sequence become larger.

The remainder of this paper is organized as follows. In Section 1, we provide further background on the factorial hidden Markov volatility model. In section 2, we discuss the inference procedure for the latent factors structure. We then present our iterative maximum-likelihood expectation-maximization (EM) algorithm in Section 3. We describe the portfolio's VaR simulation-based Viterbi tracking strategy in Section 4 and report on the backtesting results in Section 5. In this paper, the currency risk of the Tunisian government's foreign debt portfolio during the revolution period of 14 January 2011 is considered as the basis for an application to our novel prediction strategy. Our portfolio includes the main debt currencies against the Tunisian dinar, such as the European euro, the American dollar, the Japanese yen, the Swiss franc and the British pound. Finally, we conclude the paper by summarizing our contributions and discussing the future research directions.

1 THE FACTORIAL HIDDEN MARKOV VOLATILITY MODEL

Throughout this paper, we consider a multivariate discrete-time model. The closing price of the k-th asset in the portfolio at the t-th trading day is denoted by $p_{k,t}$, and the opening price at the first trading day by $p_{k,0}$.

For each $t \geq 1$, let $r_{k,t} = \log(p_{k,t} / p_{k,t-1})$ be the log-return of the k-th asset. Our model assumes a Markov switching relationship between the observed variables (the log-returns) and a set of q latent factors, which depend on the market regime. This new framework, called factorial hidden Markov volatility model (FHMV), is defined by:

$$\boldsymbol{r}_t = \boldsymbol{\Phi}_j \boldsymbol{z}_t + \epsilon_t, \tag{1}$$

where: $\forall\ t = 1, \ldots, T$, $\boldsymbol{r}_t$ is a $(p \times 1)$ vector of log-returns.

The transition probabilities of the first order homogenous hidden Markov process from state i to state j ($\forall\ i, j = 1, \ldots, n$) are represented by $p(S_t = j \mid S_{t-1} = i)$, where j is the actual market regime at time t, given the previous regime i at time $t - 1$. In a specified regime $S_t = j$, $\boldsymbol{\Phi}_j$ is the $(p \times q)$ factor loadings matrix.

The common latent factors $\boldsymbol{z}_t$ are generated from the multivariate normal distributions:

$$\boldsymbol{z}_t \sim N(\mathbf{0}, \boldsymbol{\Omega}_j) , \tag{2}$$

where: $\mathbf{0}$ and $\boldsymbol{\Omega}_j$ denote, respectively, the $(q \times 1)$ mean vectors and $(q \times q)$ diagonal covariance matrices of the latent vectors $\boldsymbol{z}_t$.

The diagonal elements of $\boldsymbol{\Omega}_j$ (common variances) are described by switching univariate quadratic GARCH(1,1) processes. Under a particular regime $S_t = j$ since $S_{t-1} = i$, the l-th common factor variance is given by:

$$\omega_{l,t}^{j} = \beta_{0l}^{j} + \beta_{1l}^{j} z_{l,t-1}^{i} + \beta_{2l}^{j} z_{l,t-1}^{i2} + \beta_{3l}^{j} \omega_{l,t-1}^{i} . \tag{3}$$

Assuming that $\beta_{1l}^{j}, \beta_{2l}^{j} > 0$, if $z_{l,t-1} < 0$, its impact on the variance $\omega_{l,t}$ is lower than in the case where $z_{l,t-1} > 0$.

Finally, the $(p \times 1)$ vector of specific factors can be written as follows:

$$\epsilon_t \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j) , \tag{4}$$

where: $\boldsymbol{\mu}_j$ and $\boldsymbol{\Lambda}_j$ are, respectively, the $(p \times 1)$ mean vectors and $(p \times p)$ diagonal covariance matrices of the specific factors.

Assumption 1: In order to insure the positivity of the common variances and the stationarity of the covariance structure of the studied series, we introduce some constraints on the parameters of the quadratic GARCH specification, such as: $\beta_{2l}^{j} + \beta_{3l}^{j} < 1$, $\beta_{0l}^{j}, \beta_{2l}^{j}, \beta_{3l}^{j} > 0$ and $\beta_{1l}^{j2} \leq 4\beta_{0l}^{j}\beta_{2l}^{j}$, $\forall_j = 1, \ldots, n$, $l = 1, \ldots, q$.

Assumption 2: To guarantee the model identification in (1), we assume that $\forall_j$, rank $(\boldsymbol{\Phi}_j) = q$ and $p \geq q$. The factors $\boldsymbol{z}_t$ and ϵ_t are also assumed to be uncorrelated and mutually independent. For more detailed discussions of the identification problem, the reader can refer to Saidane and Lavergne, (2011), and Carnero (2004).

2 INFERENCE OF THE LATENT FACTORS STRUCTURES

Our model can be expressed as a switching state-space system with a measurement equation:

$$\boldsymbol{r}_t = \boldsymbol{\mu}_j + \boldsymbol{\Phi}_j \boldsymbol{z}_t + \epsilon_t , \tag{5}$$

and a transition equation:

$$\boldsymbol{z}_t = \mathbf{0} \cdot \boldsymbol{z}_{t-1} + \boldsymbol{z}_t , \tag{6}$$

In order to find the optimal sequences of hidden states S_t and latent factors $\boldsymbol{z}_t$, we can use the Viterbi decoding algorithm based on the minimization of the Hamiltonian cost function given by the following equation:

$$\begin{aligned} H(r_{1:T}, Z_{1:T}, S_{1:T}) \simeq c + \boldsymbol{S}_1^{'}(-\log\boldsymbol{\pi}) + \sum_{t=2}^{T} \boldsymbol{S}_1^{'}(-\log\boldsymbol{P})\, \boldsymbol{S}_{t-1} + \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{n} [\log \mid \boldsymbol{\Lambda}_j \mid + (\boldsymbol{r}_t - \boldsymbol{\Phi}_j \boldsymbol{z}_t - \boldsymbol{\mu}_j)^{'} \boldsymbol{\Lambda}_j^{-1} \\ (\boldsymbol{r}_t - \boldsymbol{\Phi}_j \boldsymbol{z}_t - \boldsymbol{\mu}_j)]\, S_t(\mathrm{j}) + \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{n} [\log \mid \boldsymbol{\Omega}_j \mid + \boldsymbol{z}_t^{'} \boldsymbol{\Omega}_j^{-1} \boldsymbol{z}_t]\, S_t(\mathrm{j}), \end{aligned} \tag{7}$$

where: $r_{1:\tau} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_\tau\}$, $Z_{1:\tau} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_\tau\}$, $S_{1:\tau} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \ldots, \boldsymbol{S}_\tau\}$ are, respectively, the sequences of observed returns, latent common factors and HMM states up to time τ; $\boldsymbol{\pi}$ the vector of initial state probabilities, of length n-states and summing to 1; $\boldsymbol{P}$ the matrix of transition probabilities of the hidden Markov chain,

the sum of all elements in the i-th row $[p_{i1} \dots p_{in}]$ is 1, $\forall\ i = 1, \dots, n$ and $\boldsymbol{S}_t = [S_t(1), \dots, St(n)]'$, where $S_t(j) = 1$, if $S_t = j$ and 0 otherwise.

If we denote by $S^*_{1:T}$ the optimal sequence of HMM states, the posterior distribution $p(Z_{1:T}, S_{1:T} \mid r_{1:T})$ can be approximated as:

$$p(Z_{1:T}, S_{1:T} \mid r_{1:T}) \simeq \eta(S_{1:T} - S^*_{1:T})\, p(Z_{1:T} \mid S_{1:T}, r_{1:T}), \tag{8}$$

i.e. the posterior distribution of the HMM state sequence $p(S_{1:T} \mid r_{1:T})$ is approached by its mode, where $\eta(y) = 1$ for $y = \phi$ and zero otherwise. The optimal sequence of HMM states can formally be obtained by solving the dynamic optimization program: $S^*_{1:T} = \arg\max_{S_{1:T}} p(S_{1:T} \mid r_{1:T})$. In this case, an almost optimal solution can be reached by maximizing recursively the probability of the best HMM sequence up to time t:

$$\delta_{t,j} = \max_{S_{1:t-1}} p(S_{1:t-1}, S_t = j, r_{1:t}) \simeq \max_i \{p(\boldsymbol{r}_t \mid S_t = j, S_{t-1} = i, S^*_{1:t-2}(i), r_{1:t-1})\, p(S_t = j \mid S_{t-1} = i) \times \max_{S_{1:t-2}} p(S_{1:t-2}), S_{t-1} = \mathrm{i}, r_{1:t-1})\}, \tag{9}$$

where: $S^*_{1:t-2}(i) = \arg\max_{S_{1:t-2}} \delta_{t-1,i}$ is the "optimal" HMM sequence up to time $t - 1$ when the market state is in regime i at time $t - 1$.

Firstly we define the "optimal" partial Hamiltonian cost up to time t of the observed log-return sequence $r_{1:t}$ when the market state is in regime j at time t:

$$\delta_{t,j} = \min_{S_{1:t-1}, Z_{1:t}} H(Z_{1:t}, \{S_{1:t-1}, S_t = j\}, r_{1:t}). \tag{10}$$

To calculate this cost correctly, we need the optimal filtered estimates of the common latent factors $\boldsymbol{z}^j_{t|t} = \mathbb{E}[\boldsymbol{z}_t | r_{1:t}, S_t = j]$, the one-step ahead predictions of the common latent factors $\boldsymbol{z}^{i(j)}_{t|t-1} = \mathbb{E}[\boldsymbol{z}_t | r_{1:t-1}, S_t = j, S_{t-1} = i]$ and their optimal filtered estimates, $\boldsymbol{z}^{i(j)}_{t|t} = \mathbb{E}[\boldsymbol{z}_t | r_{1:t}, S_t = j, S_{t-1} = i]$. We need also the predicted and filtered common variances:

$$\boldsymbol{\Omega}^{i(j)}_{t|t-1} = \mathbb{E}[(\boldsymbol{z}_t - \boldsymbol{z}^{i(j)}_{t|t-1})\, (\boldsymbol{z}_t - \boldsymbol{z}^{i(j)}_{t|t-1})' \mid r_{1:t-1}, S_t = j, S_{t-1} = i], \tag{11}$$

$$\boldsymbol{\Omega}^{ij}_{t|t} = \mathbb{E}[(\boldsymbol{z}_t - \boldsymbol{z}^j_{t|t})\, (\boldsymbol{z}_t - \boldsymbol{z}^j_{t|t})' \mid r_{1:t}, S_t = j, S_t = j], \tag{12}$$

and the covariance matrices:

$$\boldsymbol{\Omega}^{i(j)}_{t|t} = \mathbb{E}[(\boldsymbol{z}_t - \boldsymbol{z}^j_{t|t})\, (\boldsymbol{z}_t - \boldsymbol{z}^j_{t|t})' \mid r_{1:t}, S_t = j, S_{t-1} = i], \tag{13}$$

$$\boldsymbol{\Omega}^{(j)k}_{t|\tau} = \mathbb{E}[(\boldsymbol{z}_t - \boldsymbol{z}^{j(k)}_{t|\tau})\, (\boldsymbol{z}_t - \boldsymbol{z}^{j(k)}_{t|\tau})' \mid r_{1:\tau}, S_t = j, S_{t+1} = k], \tag{14}$$

where: $\boldsymbol{z}^{j(k)}_{t|\tau} = \mathbb{E}[\boldsymbol{z}_t \mid r_{1:\tau}, S_t = \mathrm{j}, S_{t+1} = k]$. From the prediction step of the switching Kalman filter (Saidane and Lavergne, 2007) we obtain the time updating formula for the common latent factors, $\boldsymbol{z}^{i(j)}_{t|t-1} = \boldsymbol{0}$, $\forall\ i, j = 1, \dots, n$ and their corresponding covariance matrices, $\boldsymbol{\Omega}^{i(j)}_{t|t-1} = diag[\omega^{i(j)}_{t|t-1}]$, where $\omega^{i(j)}_{l,t|t-1} = \beta^j_{0l} + \beta^j_{1l}\, z^i_{l,t-1|t-1} + \beta^j_{2l}[z^{i2}_{l,t-1|t-1} + \omega^i_{l,t-1|t-1}] + \beta^j_{3l}\, \omega^i_{l,t-1|t-2}$, for $l = 1, \dots, q$. Given the information set $D_{1:t-1} = \{r_{1:t-1}, Z_{1:t-1}, S_{1:t-1}\}$, the predicted variances are calculated as the conditional expectations of the predicted volatilities, $\mathbb{E}(\omega_{l,t} \mid D_{1:t-1})$, and from the total variance formula $\mathbb{E}(z^{i2}_{l,t-1} \mid D_{1:t-1}) = Var(z^i_{l,t-1} \mid D_{1:t-1}) + \mathbb{E}(z^i_{l,t-1} \mid D_{1:t-1})^2 = \omega^i_{l,t-1|t-1} + z^{i2}_{l,t-1|t-1}$, we obtain the filtered variances $\omega^i_{l,t-1|t-1}$. When a novel observation r_t becomes available, all the prediction estimates can be updated recursively via the Kalman filtering equations:

$$\boldsymbol{z}^{i(j)}_{t|t} = \boldsymbol{z}^{i(j)}_{t|t-1} + \boldsymbol{K}_t(i,j)[\boldsymbol{r}_t - \boldsymbol{\mu}_j - \boldsymbol{\Phi}_j \boldsymbol{z}^{i(j)}_{t|t-1}], \tag{15}$$

$$\boldsymbol{\Omega}^{i(j)}_{t|t} = [\boldsymbol{I}_k - \boldsymbol{K}_t(i,j)\boldsymbol{\Phi}_j]\, \boldsymbol{\Omega}^{i(j)}_{t|t-1} = \boldsymbol{\Omega}^{i(j)}_{t|t-1} - \boldsymbol{K}_t(i,j)\boldsymbol{\Gamma}^{i(j)}_{t|t-1}\boldsymbol{K}_t(i,j)^{'}, \tag{16}$$

with $\boldsymbol{\Gamma}^{i(j)}_{t|t-1} = \boldsymbol{\Lambda}_j + \boldsymbol{\Phi}_j\, \boldsymbol{\Omega}^{i(j)}_{t|t-1}\, \boldsymbol{\Phi}_j^{'}$ and $\boldsymbol{K}_t(i,j) = \boldsymbol{\Omega}^{i(j)}_{t|t-1}\, \boldsymbol{\Phi}_j^{'}\, \boldsymbol{\Gamma}^{i(j)^{-1}}_{t|t-1}$. The innovation cost $\delta_{t,t-1,i,j}$ related to each transition from state i to state j, is given by:

$$\delta_{t,t-1,i,j} = \frac{1}{2}\log|\boldsymbol{\Gamma}^{i(j)}_{t|t-1}| + \frac{1}{2}\,[\boldsymbol{r}_t - \boldsymbol{\mu}_j - \boldsymbol{\Phi}_j \boldsymbol{z}^{i(j)}_{t|t-1}]^{'}\, \boldsymbol{\Gamma}^{i(j)^{-1}}_{t|t-1}[\boldsymbol{r}_t - \boldsymbol{\mu}_j - \boldsymbol{\Phi}_j \boldsymbol{z}^{i(j)}_{t|t-1}] - \log p_{ij}. \tag{17}$$

A substantial part of this cost is exclusively due to the transition of the latent factors, as illustrated by the innovation component in Formula (17). The remaining part ($-\log p_{ij}$) reflects the transition of the market state from regime i to regime j. In this case, the minimization of the global cost at time t requires the selection of the optimal previous market state i: $\delta_{t,j} = \min_i\{\delta_{t,t-1,i,j} + \delta_{t-1,i}\}$. The resulting index is then recorded in the regime switching record, $\lambda_{t-1,j} = \arg\min_i\{\delta_{t,t-1,i,j} + \delta_{t-1,i})\ \}$. As a result, we obtain for each time t the "optimal" filtered latent factors $\boldsymbol{z}^{j}_{t|t} = \boldsymbol{z}^{\lambda t-1,j(j)}_{t|t}$ and their corresponding variances $\boldsymbol{\Omega}^{j}_{t|t} = \boldsymbol{\Omega}^{\lambda t-1,j(j)}_{t|t} = diag[\omega^{\lambda t-1,j(j)}_{l,t|t-1}]$.

When all the log-returns $r_{1:T}$ become available, we obtain the optimal global cost $\delta_T^{\,*} = \min_j\{\delta_{T,j}\}$. Then, we use the index of the optimal final state in order to decode the optimal sequence of HMM states: $j_T^{\,*} = \arg\min_j\{\delta_{T,j})\ \}$. To get the best regime for all time steps, we trace back through the market regime switching record: $j_T^{\,*} = \lambda_{t,j^*_{t+1}}$.

We note here that the smoothing gain matrix $\boldsymbol{L}_t^{(j)k} = \boldsymbol{\Omega}^{j}_{t|t}\, \boldsymbol{0}_k^{'} \boldsymbol{\Omega}^{(j)k-1}_{t+1|t} = \boldsymbol{0}$ and the smoothing equations are simply given by:

$$\boldsymbol{z}^{(j)k}_{t|T} = \boldsymbol{z}^{j}_{t|t} + \boldsymbol{L}_t^{(j)k}[\boldsymbol{z}^{k}_{t+1|T} - \boldsymbol{z}^{j(k)}_{t+1|t}] = \boldsymbol{z}^{j}_{t|t}\,, \tag{18}$$

$$\boldsymbol{\Omega}^{(j)k}_{t|T} = \boldsymbol{\Omega}^{j}_{t|t} + \boldsymbol{L}_t^{(j)k}[\boldsymbol{\Omega}^{k}_{t+1|T} - \boldsymbol{\Omega}^{j(k)}_{t+1|t}]\, \boldsymbol{L}_t^{(j)k'} = \boldsymbol{\Omega}^{j}_{t|t}\,. \tag{19}$$

Following the smoothing procedure developed by Saidane and Lavergne (2008), the sufficient statistics for our estimation problem will be given by: $\mathbb{E}(S_t|\cdot) = S_t\ (j^*)$, $\mathbb{E}(S_t\, S_{t-1}^{'}|\cdot) = S_t(j^*)\ S_{t-1}(j^*)^{'}$ and $\mathbb{E}(\boldsymbol{z}_t\, S_t(j)|\cdot) = \boldsymbol{z}^{j^*_t}_{t|T}$), if $j = j_t^{\,*}$ and $\boldsymbol{0}$ otherwise. In this case, the operator $\mathbb{E}(|\cdot)$ denotes the expectation with respect to the distribution $p(Z, S|r)$.

3 MAXIMUM LIKELIHOOD ESTIMATION

We propose a two-step learning algorithm combining the expectation maximization (EM) algorithm (Dempster et al., 1977) and the Viterbi decoding algorithm in order to estimate the parameters Θ of our model. The E-step subsists in calculating the expected value of the complete data log-likelihood function with respect to the conditional distribution of the unobserved variables (Z, S) given the observed returns r and $\Theta^{(e)}$, the value of the parameter at the current iteration (e). The conditional expectation is then, maximized with respect to Θ at the M-step. In this case, the auxiliary function that will be maximized can be approximated as follows:

$$\begin{aligned} Q(\Theta,\Theta^{(e)} \simeq\ & \sum_{j=1}^{n} S_1(j)\, \log p(S_1) - \sum_{t=2}^{T}\sum_{i=1}^{n}\sum_{j=1}^{n} S_t(j)\ S_{t-1}(i)\, \log p_{ij} - \frac{1}{2}\ \sum_{i=1}^{n}\sum_{j=1}^{n} S_t(j)[\log|\boldsymbol{\Lambda}_j| + \mathbb{E}\{(\boldsymbol{r}_t - \boldsymbol{\mu}_j - \boldsymbol{\Phi}_j\, \boldsymbol{z}^{j}_t)^{'} \\ & \boldsymbol{\Lambda}_j^{-1}(\boldsymbol{r}_t - \boldsymbol{\mu}_j - \boldsymbol{\Phi}_j\, \boldsymbol{z}^{j}_t)|r_{1:T}), \Theta^{(e)}\}] - \frac{1}{2}\sum_{j=1}^{n}\sum_{l=2}^{q}\sum_{t=1}^{T} S_t(j)\ \mathbb{E}[\log(\omega^{j}_{l,t}) + \frac{z^2_{l,t}}{\omega^{j}_{l,t}}|r_{1:T}, \Theta^{(e)}], \end{aligned} \tag{20}$$

and the conditional expectations can be derived using the sufficient statistics obtained by the Viterbi algorithm in Section 3.

The basic idea behind our algorithm is summarized as follows: At the end of each iteration (e) we find $\Theta^{(e+1)}$, the optimal value of the parameter Θ that maximizes the function in equation (20) over all possible values of Θ. Then $\Theta^{(e+1)}$ replaces $\Theta^{(e)}$ in the E-step and $\Theta^{(e+2)}$ is chosen to maximize $Q(\Theta,\Theta^{(e+1)})$, and so on until convergence. However, given the nonlinear dependency of the common variance parameters in the last summation of Formula (20), we can maximize in a first time this function with respect to the probabilities of the initial state π_j, the transition probabilities p_{ij}, the specific means $\boldsymbol{\mu}_j$, the factor loadings $\boldsymbol{\Phi}_j$ and the specific variances $\boldsymbol{\Lambda}_j$. In a second time, the parameters of the common variances can be determined numerically.

For the initial state probabilities π_j, we use the Lagrange multipliers approach subject to the condition that the sum $\sum_{j=1}^{n} \pi_j = 1$, and we obtain the updated estimation:

$$\hat{\pi}_j = \frac{S_1(j)}{\sum_{j=1}^{n} S_1(i)}. \tag{21}$$

We use also the Lagrange formalism, subject to the unity constraint $\sum_{j=1}^{n} p_{ij} = 1$, to obtain the updated transition probabilities:

$$\hat{p}_{ij} = \frac{\sum_{t=2}^{T} S_t(j) S_{t-1}(i)}{\sum_{t=2}^{T} S_{t-1}(i)}. \tag{22}$$

The maximization of the auxiliary function with respect to the specific means yields the updated estimates:

$$\hat{\boldsymbol{\mu}}_j = \frac{1}{\sum_{t=1}^{T} S(j)} \sum_{t=1}^{T} S_t(j)(\boldsymbol{r}_t - \boldsymbol{\Phi}_j \boldsymbol{z}_{t|T}^{j}). \tag{23}$$

The updated l-th row of the factor loadings matrix $\hat{\boldsymbol{\Phi}}_j$ can be expressed as follows:

$$\hat{\boldsymbol{\phi}}_{l,j} = \left[\sum_{t=1}^{T} S_t(j)(r_{l,t} - \mu_{l,j})\boldsymbol{z}_{t|T}^{j}\right]' \left[\sum_{t=1}^{T} S_t(j)[\boldsymbol{\Omega}_{t|T}^{j} + \boldsymbol{z}_{t|T}^{j}\boldsymbol{z}_{t|T}^{j'}]\right]^{-1}, \tag{24}$$

where: $\mu_{l,j}$ is the specific mean of the l-th asset return $r_{l,t}$ under the market regime j. Then, given these updated parameters, we can update the specific variances according to the following rule:

$$\hat{\boldsymbol{\Lambda}}_j = \frac{1}{\sum_{t=1}^{T} S(j)} \sum_{t=1}^{T} S_t(j) diag\,[\boldsymbol{\Phi}_j \boldsymbol{\Omega}_{t|T}^{j} \boldsymbol{\Phi}_j' + (\boldsymbol{r}_t - \boldsymbol{\mu}_j - \boldsymbol{\Phi}_j \boldsymbol{z}_{t|T}^{j})(\boldsymbol{r}_t - \boldsymbol{\mu}_j - \boldsymbol{\Phi}_j \boldsymbol{z}_{t|T}^{j})]'. \tag{25}$$

In a second time, given the new values of π_j, p_{ij}, $\boldsymbol{\mu}_j$, $\boldsymbol{\Phi}_j$ and $\boldsymbol{\Lambda}_j$, we can approximate the conditional distribution of the log-returns by the normal distribution: $\boldsymbol{r}_t \,|r_{1:t-1}, S_t = j, S_{1:t-1} \sim N[\boldsymbol{\mu}_j, \boldsymbol{\Gamma}_{t|t-1}^{j}]$ (e.g. Harvey et al., 1992). In this case, $\boldsymbol{\Gamma}_{t|t-1}^{j} = \boldsymbol{\Lambda}_j + \boldsymbol{\Phi}_j \boldsymbol{\Omega}_{t|t-1}^{j} \boldsymbol{\Phi}_j'$ and $\boldsymbol{\Omega}_{t|t-1}^{j} = diag[\omega_{t|t-1}^{\lambda_{t-1,j}(j)}]$ is the conditional expectation of $\boldsymbol{\Omega}_t$, given the sequences $r_{1:t-1}$ and $S_{1:t-1}$, obtained via the modified Kalman filter approach based on the Viterbi decoder developed in Section 3.

Using these approximations and ignoring the initial conditions, we obtain the following pseudo log-likelihood function:

$$L^* = c - \frac{1}{2}\sum_{t=1}^{T}\sum_{j=1}^{n} S_t(j)[\log|\mathbf{\Gamma}^{j-1}_{t|t-1}| + (\boldsymbol{r}_t - \boldsymbol{\mu}_j)'\mathbf{\Gamma}^{j-1}_{t|t-1}(\boldsymbol{r}_t - \boldsymbol{\mu}_j)] \ . \tag{26}$$

In a first stage, we ignore the elements in the last summation of Formula (20) and then we maximize the remaining terms with respect to (π_j, p_{ij}, $\boldsymbol{\mu}_j$, $\mathbf{\Phi}_j$ and $\mathbf{\Lambda}_j$) using the EM algorithm. During this step, the parameters of the quadratic GARCH processes $\beta = \{\beta_0, \beta_1, \beta_2, \beta_3\}$ are kept fixed to their values obtained in the previous iteration. In a second stage, we optimize the pseudo log-likelihood in Formula (26) with respect to β, using the values of π_j, p_{ij}, $\boldsymbol{\mu}_j$, $\mathbf{\Phi}_j$ and $\mathbf{\Lambda}_j$ found in the first step. The R package NlcOptim, developed by Chen and Yin (2019), can be used in this step to find quickly and most accurately the parameters of the conditionally heteroskedastic component β.

4 THE FHMV APPROACH FOR VALUE-AT-RISK PREDCTION

Formally put, Value-at-Risk is a financial metric that measures the worst expected loss that could happen in an investment portfolio over a given horizon for a given confidence level. In this section a general Monte Carlo simulation FHMV-based framework for value-at-risk prediction, under regime switching dynamics, will be proposed. This approach will then be used, in Section 5, for the evaluation of the currency risk associated with the Tunisian government's foreign debt portfolio during the revolution period of 14 January 2011.

4.1 Forecasting future market regime changes

Given the information set available at time t, $D_{1:t}$, and the actual market regime i, the conditional mean of the multivariate predictive distribution given by our FHMV model is as follows:

$$\mathbb{E}(\boldsymbol{r}_{t+1}|D_{1:t}) = \boldsymbol{\mu}_j, \tag{27}$$

and the conditional variance-covariance matrix is given by:

$$\mathbf{\Gamma}^j_{t+1|t} = \mathbf{\Lambda}_j + \mathbf{\Phi}_j\mathbf{\Omega}^j_{t+1|t}\mathbf{\Phi}_j' \ . \tag{28}$$

Within this framework, the forecasts of the future market regime jumps and the model parameters updating process are implemented simultaneously. Thus, by the end of each transaction day the closing prices will be included in the database. Thereafter, the parameters of our model will be updated using the newer information set available at this point in time, and the updated one-step-ahead forecasts of the common latent factor variances will be derived via the relation: $\tilde{\omega}^j_{l,t+1|t} = \beta^j_{0l} + \beta^j_{1l}z^j_{l,t|t} + \beta^j_2 z^{i2}_{l,t|t} + \beta^j_3\omega^i_{l,t|t}$. Then, the future market regime $S_{t+1} = j$ can be obtained as a solution of the optimization problem: $\hat{S}_{t+1|t} = \arg\max_j p(S_{t+1} = j \mid S_t = i_t^*)$, where i_t^* is the optimal market regime at time t obtained by the Viterbi algorithm, through the state transition record $\lambda_{t-1,i}$ at each time step. The FHMV model with the optimal future hidden state $\hat{S}_{t+1|t}$, will be used in the simulation procedure as the data generating process, to calculate the VaR of our portfolio.

4.2 The simulation strategy

Our simulation strategy consists of the following steps:

1. Firstly, we define the coverage rate α of the VaR.
2. Then, taking into account the presence of leverage effects and conditional skewness in financial time series, we simulate different return scenarios from the conditional distribution of the common latent factors $\boldsymbol{z}^{\hat{S}}_{t+1|t}$, using the optimal specification obtained by the Viterbi algorithm at time t (Section 5.1).
   a. We use in a first time the normal distribution $N(\mathbf{0},\boldsymbol{I}_q)$ to generate the standardized factors $\boldsymbol{z}^*_{t+1}$.

b. Then, we compute the lower triangular Cholesky factor $\boldsymbol{\Omega}^{j*}_{t+1|t}$ of the variance-covariance matrix $\boldsymbol{\Omega}^{j}_{t+1|t}$, and we obtain: $\boldsymbol{z}^{s}_{t+1|t} = \boldsymbol{\Omega}^{j*}_{t+1|t}\, \boldsymbol{z}^{*}_{t+1}$.

3. After that, we simulate different return scenarios from the conditional distribution of the specific factors $\epsilon^{s}_{t+1|t}$, using also the optimal specification obtained by the Viterbi algorithm at time t.
   a. We generate in a first time from the normal distribution $N(\mathbf{0},\boldsymbol{I}_p)$ the standardized specificities $\epsilon^{*}_{t+1|t}$.
   b. Then, we compute the lower triangular Cholesky factor $\boldsymbol{\Lambda}_j{}^{*}$ of the variance-covariance matrix $\boldsymbol{\Lambda}_j$, and we obtain: $\epsilon^{s}_{t+1|t} = \boldsymbol{\Lambda}_j{}^{*}\, \epsilon^{*}_{t+1}$.
4. In the fourth step, we compute m different portfolio's returns for the period $t + 1$ as, $R_{s,t+1|t} = \gamma_1\, r^{s}_{1,t+1|t} + \gamma_2\, r^{s}_{2,t+1|t} + \cdots + \gamma_p\, r^{s}_{p,t+1|t}$, where $\gamma_1, \gamma_2, \ldots, \gamma_p$ denote the portfolio weights of the p risk factors and $\boldsymbol{r}^{s}_{t+1|t} = \boldsymbol{\mu}_j + \boldsymbol{\Phi}_j\, \boldsymbol{z}^{s}_{t+1|t} + \epsilon^{s}_{t+1|t}$ $(\forall s = 1, \ldots, m)$.
5. Finally, to compute the portfolio's VaR for the period $t + 1$, we sort the simulated values in ascending order and we exclude the α% lowest returns $R_{s,t+1|t}$. In this case, the predicted VaR is the minimum of the remaining returns.

5 NUMERICAL EXPERIMENTS USING EXCHANGE RATE DATA

In this empirical experiment, we use the FHMV model to analyze the dynamic latent correlation structure of the five dominant currencies in the Tunisian government's foreign debt portfolio. The optimal specification obtained with the Viterbi algorithm will then be used to evaluate, through a backtesting exercise, the performance of the new methodology in detecting the foreign exchange risk associated with this portfolio. All the numerical results and the graphs in this section are obtained using the R statistical freeware, version 4.1.

5.1 Data presentation and summary

We focus in this section on the main five currencies forming the basis for the Tunisian government's foreign debt portfolio, namely the European euro (EUR), the American dollar (USD), the Japanese yen (JPY), the Swiss franc (CHF) and the British pound (GBP). Our dataset, downloaded from the Yahoo

Table 1 Basic descriptive characteristics of the daily exchange rate returns from 2/1/2010 to 30/12/2012

Statistic	EUR/TND	USD/TND	JPY/TND	CHF/TND	GBP/TND
Mean	0.000327	0.000381	0.000179	0.000274	0.000363
Max	0.0439	0.0423	0.0514	0.0424	0.1468
Median	0.000198	0.000255	0.000163	0.000308	0.000212
Min	–0.0538	–0.0619	–0.0718	–0.0653	–0.0083
Std. Dev.	0.00425	0.00510	0.00674	0.00583	0.00620
D'Agost. test	7.49623	2.72961	4.14286	–6.21753	9.57321
	(0.0000)	(0.0047)	(0.0000)	(0.0081)	(0.0000)
A-Glyn. test	18.751	16.173	15.927	16.369	21.916
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
LB. test	117.67	108.21	41.962	62.114	123.813
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
J-Bera. test	65.321	29.655	14.533	26.123	34.259
	(0.0000)	(0.0000)	(0.0074)	(0.0000)	(0.0000)

Note: The values into brackets represent the p-values of the corresponding tests.
Source: Own construction

Finance website, spread over the period between 2/1/2007 and 30/12/2012, consists of 1 500 daily exchange rates for the different currencies expressed in terms of Tunisian dinar (TND). This dataset includes the period of social mobilization and political change in Tunisia (the revolution of 14 January 2011). In this case, taking into account the period of social instability we will be permitted to investigate the efficiency of our Jump-VaR methodology during crisis times.

In Table 1, we give a variety of descriptive statistics to study the distributional characteristics of the data and to test the empirical skewness and Kurtosis against the values of normal distributions (e.g. D'Agostino, 1970; Anscombe-Glynn, 1983). We implemented also the normality test (Jarque-Bera, 1980). From these results, we note that all the log-returns are non-normally distributed, they are still skewed (positive for EUR, USD, JPY and GBP and negative for CHF). We note also a positive excess kurtosis for all the currencies. The results of the Ljung-Box (1978) statistic show the presence of volatility clustering. This imply that we have a non-constant conditional volatility, and the use of a Markov-switching specification with a time-varying co-movement structure for the log-return series, is more realistic in this situation.

5.2 A preliminary latent structure analysis of the data

In order to select the most appropriate model that fits better our dataset, we used the Akaike (AIC) and the Bayesian (BIC) information criteria. To this end, we trained standard and conditionally heteroskedastic models using one or two common factors and a number of hidden states varying between one and three, on the period from 2/1/2010 to 30/12/2012. Then, we used the selection criteria to identify the best model with the minimum AIC and BIC values.

Table 2 AIC and BIC values for the different specifications over the period 2/1/2010–30/12/2012 (750 observations)

The number of common factors	Criterion	Number of hidden states		
		1	2	3
1	AIC	2 436.2	1 973.6	2 199.2
		(2 826.9)	(2 610.5)	(2 585.9)
	BIC	2 518.3	2 424.4	2 591.0
		(3 097.6)	(2 731.8)	(2 644.3)
2		2 394.8	1 958.3	2 123.8
		(3 016.4)	(2 415.2)	(2 275.5)
		2 509.6	2 151.8	2 269.9
		(3 211.1)	(2 597.8)	(2 403.7)

Note: The selection criteria values for the standard models are given into brackets.
Source: Own construction

The results reported in Table 2 show that the FHMV model with two common factors and two HMM states is the best one fitting our dataset. For this optimal specification, the initial state probability vector and the transition probability matrix are as follows:

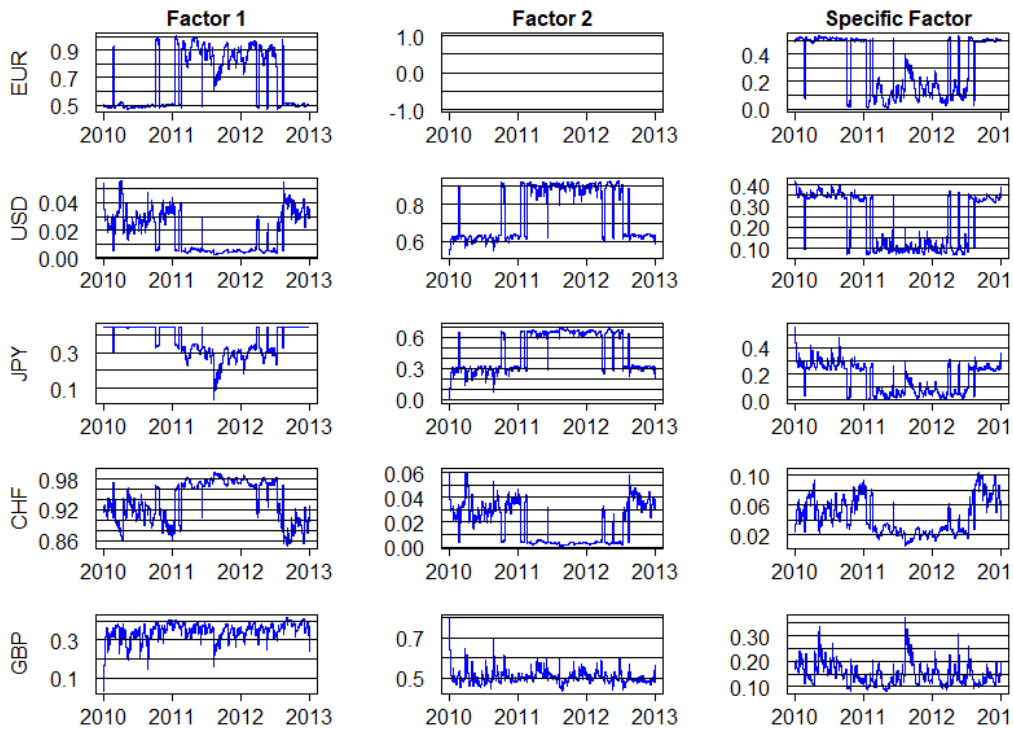
$$\pi = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \text{ and } \boldsymbol{P} = \begin{bmatrix} 0.9491 & 0.0509 \\ 0.2311 & 0.7689 \end{bmatrix}.$$

In Figure 1, we depict the percentage of the variability of the different currencies expressed in terms of specific and common factors. During the crisis period, we can see that, on average, 90% and 95% of the variances of the EUR and the CHF are explained by the first common factor. During the normal

period (before and after the 2011–2012), we can see that the first common factor explains on average 50% and 90% of the variability of the EUR and CHF.

The contribution of the second common factor to the variability of the EUR and CHF, during the instability period, is almost insignificant. During the revolution period, this factor explains more than 85% and 65% of the variability of the USD and JPY. However, the contribution of the first factor to the JPY variability is around 45% during the normal period. For the GBP, the contribution of this factor is around 35% over the whole period.

Figure 1 The percentage of variability of the different log-returns associated to the common and specific factors over the period 2/1/2010–30/12/2012



Source: Own construction

From these results, we can conclude that the first common factor is associated with the volatility dynamics of the European currencies. During the social mobilization period, the second factor is associated with the volatility dynamics of the American and Japanese currencies. We can conclude also that the first common latent factor expresses the relative value of the TND against the major trading partner's currencies (the European community countries). The second factor reproduces the relative value of the TND against a basket of global currencies in which the American and Japanese currencies are dominant.

From the estimation results presented in Table 3, we note that the first common factor can be regarded as a European factor: it represents a basket of currencies, where the EUR dominates with relatively high loadings (50% in the first regime and 76% in the second regime). The weight of the GBP is relatively reduced in this basket. We note also that the second common factor represents a basket

of currencies, where the USD dominates with relatively high loadings (52% in the first regime and nearly 60% in the second regime). In order to satisfy the identification constraints (e.g. Saidane and Lavergne 2011), we have taken $\phi_{1,2,j} = 0$, $\forall_j = 1, 2$, which imply that the European currency EUR is entirely absent from the second factor. The relative weight of the CHF is also reduced in this factor. Hence, we can consider the second common factor as an American factor.

Table 3 Estimation results of the optimal FHMV model during the period 2/1/2010–30/12/2012

Model parameters (1e-04)	Currencies				
	EUR	USD	JPY	CHF	GBP
$\hat{\mu}_1$	1.6104	1.5216	1.1622	–1.1131	1.3109
$\hat{\mu}_2$	2.1724	3.0221	1.7447	0.0156	1.7932
$\hat{\Phi}_1$	7.7028	1.9780	1.2636	2.6804	1.7611
	0.0000	5.1303	2.3204	0.3217	2.1032
$\hat{\Phi}_2$	9.1526	0.7315	0.3949	1.4710	0.2389
	0.0000	8.9482	2.5801	0.6206	2.8312
$\hat{\Lambda}_1$	0.6405	1.0623	1.5049	1.1458	2.1004
$\hat{\Lambda}_2$	0.2230	0.9071	1.2047	1.0167	2.7103

Source: Own construction

Table 4 Estimated parameters of the conditionally heteroskedastic components of the optimal FHMV model during the period 2/1/2010–30/12/2012

Common Variance parameters	β_0	β_1	β_2	β_3
	Factor 1			
Regime 1	0.1224	0.0009	0.2411	0.3648
Regime 2	0.1521	0.1382	0.2363	0.7601
	Factor 2			
Regime 1	0.0843	0.1774	0.1892	0.4667
Regime 2	0.1102	0.1430	0.2711	0.7092

Source: Own construction

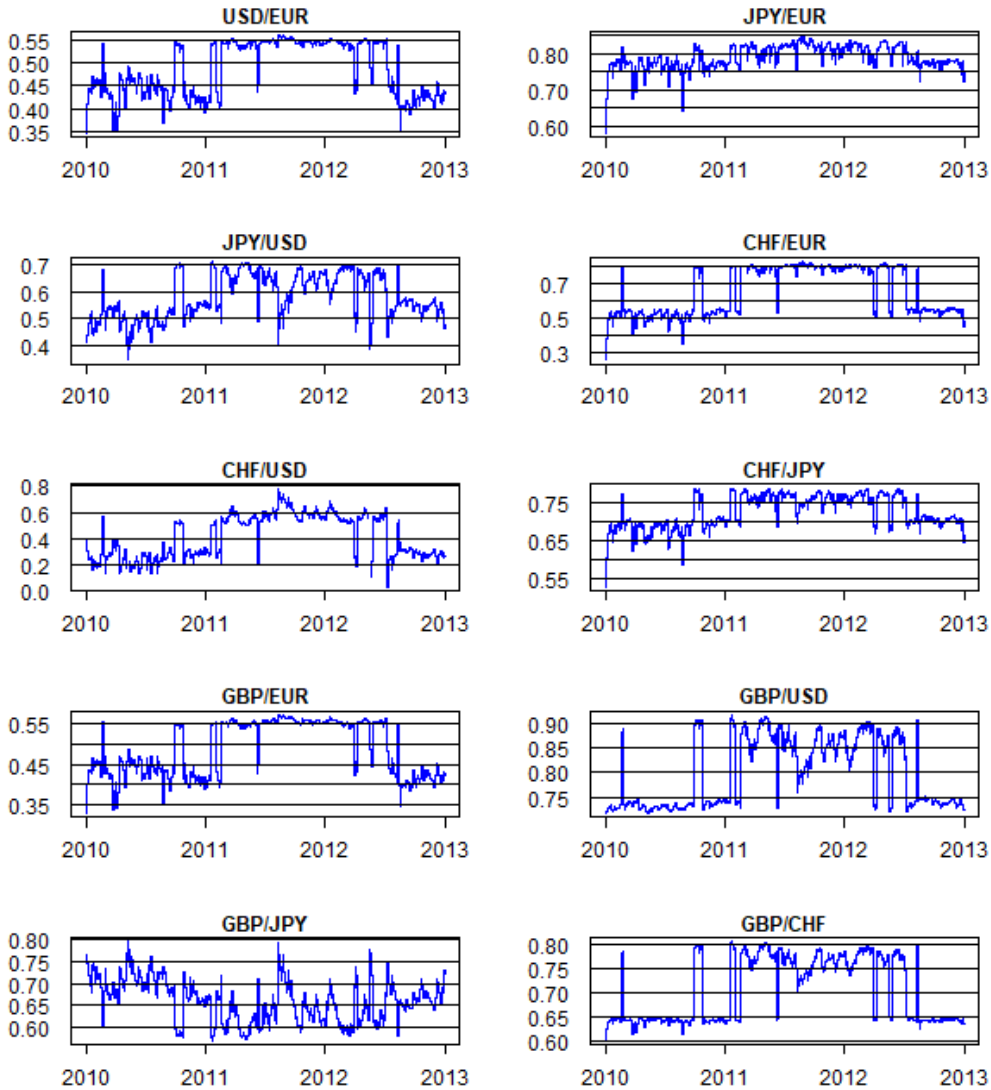
From Table 4, it appears that the excess of volatility during the political instability period (the second hidden regime) is relatively due to the significant increase in volatility persistence (e.g. Klaassen, 2002). We observe from this table that the sum of the volatility parameters, β_2 and β_3, of the two common factors in the second regime is nearly close to 1.

All the previous conclusions are strongly confirmed by the estimated values of the specific variances, given in Table 3. Hence, during the social mobilization period, the specific variance of the British pound is relatively high, which indicates its aberration from its latent factorial class. On the other hand, the specific variances of the European euro and the American dollar are the smallest ones, which indicate their determinant role in their latent factorial class.

Finally, in Figure 2, we depict the correlation structure of the different log-returns during the period 2/1/2010 to 30/12/2012. The graph picks up co-movement increases between all the log-returns from

the beginning of 2011 until near the end of the study period. This result confirms the financial contagion that affected the Tunisian economy during the revolution period.

Figure 2 Plot of the estimated correlations from the best FHMV model, during the period 2/1/2010–30/12/2012



Source: Own construction

5.3 Selection of the most appropriate VaR model

In order to assess the currency risk associated with the Tunisian government's foreign debt portfolio during the social mobilization period of 14 January 2011, we divided in a first time our dataset into calibration set and test set. The calibration, called also training, set contains the log-returns of the different exchange rates during the period 2/1/2007–30/12/2009 (750 observations). The test, called also

backtesting, set contains the remaining 750 observations covering the period 2/1/2010–30/12/2012. Then, we used the Monte Carlo simulation strategy (Section 4.2) to evaluate the VaR of our portfolio. For each coverage rate α, we used the portfolio weights given in Table 5. Here, the weight of each exchange rate $\gamma_{\_k}$ is determined by the relative share of currency k in the payment of the total foreign debt. For example, in 2010 Tunisia settled 61.3% of its foreign loans in Euro, 14.3% in American dollar, 16.1% in Japanese yen, 2.4% in Swiss franc and 5.9% in British pounds. Hence, in 2010, $\gamma_1 = 0.613$, $\gamma_2 = 0.143$, $\gamma_3 = 0.161$, $\gamma_4 = 0.024$ and $\gamma_5 = 0.059$. For 2011 and 2012, the weights are determined in the same way.

Table 5 Structure of the Tunisian foreign debt by settlement currency for the period 2010–2012

Date	Indicators				
	EUR	USD	JPY	GBP	CHF
31/12/2010	61.3	14.3	16.1	5.9	2.4
31/12/2011	56.8	20.1	15.3	5.6	2.2
31/12/2012	59.6	18.9	13.8	5.1	2.6

Source: Monetary and financial statistics of the Tunisian central bank

In a second time, the effectiveness of our methodology is justified by some experiments, using unconditional (Kupiec, 1995) and conditional (Christoffersen, 2012) tests and the rolling sample method based on a one-day moving window scheme with the coverage rates from the level of 0.005 to 0.1 by 0.005. All these calculations have been carried out by simulations from our FHMV model, the mixed factorial hidden Markov model (MFHMM) by Saidane (2019), the latent factor model with time varying volatility (FM) by Saidane (2017) and the classical Monte Carlo simulation method (CMC) by Mosbahi et al. (2017).

In order to compromise between precision and efficiency, we generated $m = 25\,000$ scenarios from each competing model (e.g. Saidane, 2022; Lu et al., 2014; Bastianin, 2009; Fantazzini, 2008). Then, we calculated the VaR, the exception rates and the likelihood ratios for the proportion of failure test (LR-*pof*), the independence test (LR-*ind*) and the conditional coverage test (LR-*cc*).

All the results of the backtesting experiments are given in Table 6 and Figures 3–4. The Kupiec and Christoffersen backtesting results show that the optimal FHMV model, with 2 latent factors and 2 hidden states, provides good results and gives exception rates very close to the target (the true coverage rates α). The likelihood ratios associated with the unconditional and independence tests, for our proposed model, are always lower than the critical values, which imply a significant conditional coverage tests for all the confidence levels.[2]

Table 6 Backtesting results for the FHMV model with two hidden states and two common factors

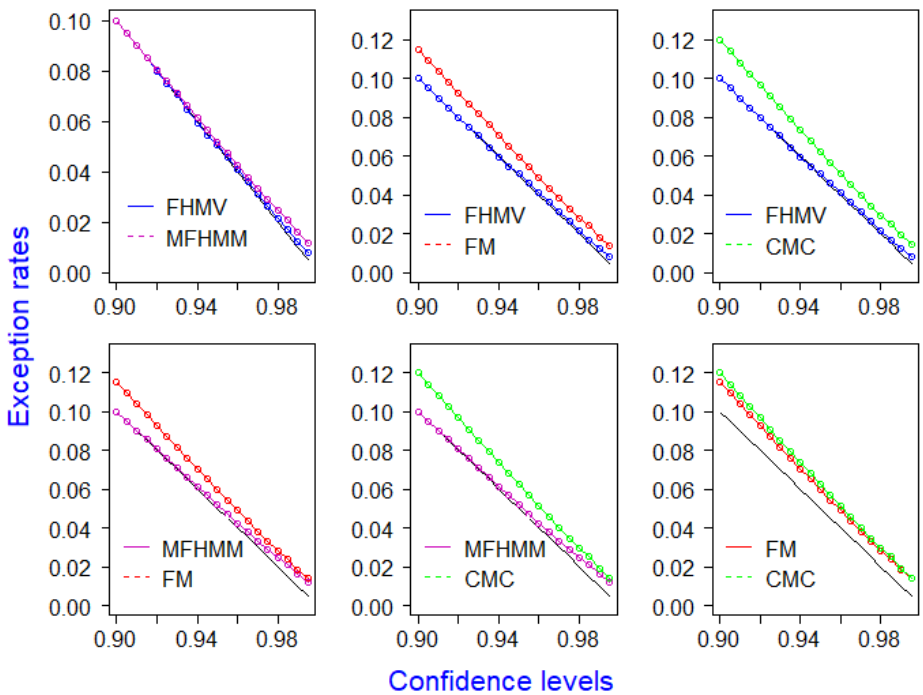
Confidence level	Exception rate	1st exception	LR-pof	LR-ind	LR-cc
0.950	0.050	18	1.7934	1.1152	2.9086
0.960	0.040	22	1.8372	1.1256	2.9628
0.970	0.029	38	2.0839	1.1398	3.2237
0.980	0.022	38	2.1856	1.1458	3.3314
0.990	0.013	56	2.3122	1.1593	3.4715

Source: Own construction

[2] The critical values for the Kupiec and Christoffersen tests are, respectively, χ^2 (1) = 3.8414 and $\chi^2(2) = 5.9915$ for 95% VaR.

For the coverage rates from 0.5% to 2%, Figure 3 shows promising results for the optimal FHMV model compared to those given by the best MFHMM (with 2 mixture components and 2 latent factors). For the significance level 2%, our FHMV model gives, for example, an exception rate equal to 2.17%, versus 2.45% obtained by the MFHMM. From this figure, we can see also that the optimal MFHMM looks better than the FM and CMC, especially at low confidence levels. Hence, we can argue that the FHMV model is the more precise and yields higher-quality predictions, as compared to the other competing models.

Figure 3 Exception rates for various confidence levels from the rolling window experiments

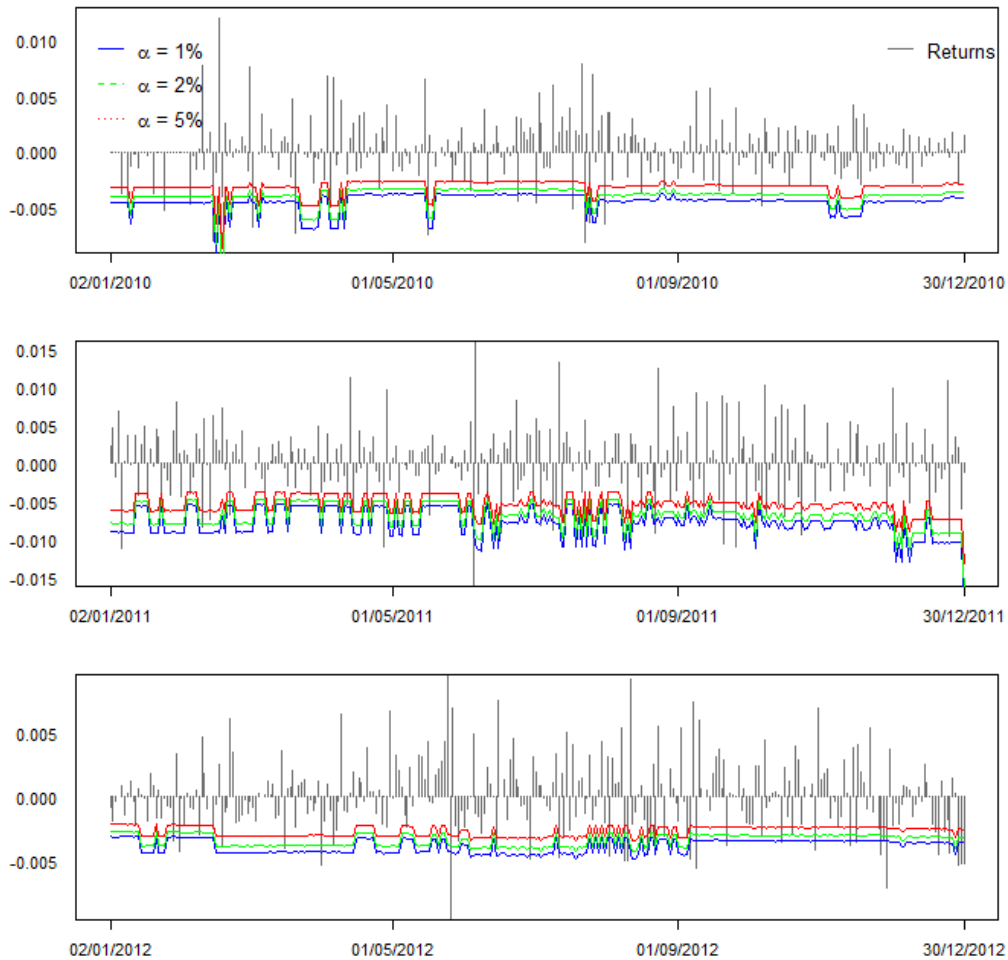


Source: Own construction

In order to compare the results given by the different models, we used the squared relative prediction error criteria $S = \sum_{i=1}^{20}[(E_i - \alpha_i)/\alpha_i]^2$, where E_i are the estimated exception rates obtained with the different specifications, and α_i the coverage rates. It appears from the results that the most adequate model to evaluate the VaR of our portfolio, during this period, is the FHMV framework. This specification gives the estimated exception rates closest to all the true significance levels with $S = 0.4561$. The second ranked model is the MFHMM ($S = 2.5824$), the third is the FM with ($S = 4.9783$), and the CMC is the worst one with $S = 6.0673$.

Finally, we can see from Figure 4 the significant effect of the volatility shocks on the predicted VaR given by the optimal FHMV model. Hence, we can argue that the major reason for the bad results given by the CMC, FM, and to a lesser degree the MFHMM, is that they do not take into account the abnormal switching behaviors, which can affect the volatility and the co-movement dynamics in financial markets during crisis periods.

Figure 4 The VaR exceptions obtained by the optimal FHMV model for different coverage rates and different portfolio weights



Source: Own construction

CONCLUSION

This paper develops a new multivariate approach for Value-at-Risk (VaR) prediction. Our strategy considers the possibility of regime jumps in the intra-portfolio's latent correlation structure and allows for time-varying volatility in the factor variances. The proposed framework combines factor analysis models with GQARCH processes and hidden Markov models. During financial crisis periods, this specification provides a more tractable way to capture simultaneously the switching interrelations between assets and the time-varying volatility of each individual asset.

The accuracy of the new prediction approach in comparison with other existing models (such as the mixed factorial hidden Markov model, the latent factor model with time varying volatility and the classical Monte Carlo method) is evaluated through a real dataset example from the Tunisian foreign exchange market for the period 2/1/2010 to 30/12/2012. Our strategy aims to select the best model that

could predict the VaR of the Tunisian government's foreign debt portfolio during the social mobilization period of 14 January 2011. In that period the Tunisian economy has experienced the longest, deepest and most broad-based recession in its history since the 1978. The main results of the empirical example and the backtesting experiments, based on the rolling sample method, show that the new approach appears to give a good fit to the data, allows to more close forecasts to the market changes and can improve the VaR predictions and offer more accurate VaR estimates than the other competing models for all coverage rates from 0.5% to 10%.

We conclude that our Viterbi-based decoding strategy using the factorial hidden Markov volatility model seems to be a useful tool for portfolio risk management and control, especially during periods of financial market stress. These results support our argument for integrating time-varying volatility and regime jumps into the risk measurement framework. In the forthcoming works, we intend to reflect the interaction between the common latent factors with a dynamic structure for the idiosyncratic variances. We will address also nonlinear behaviors, non-homogeneous transition probabilities and other areas of application, like options, or credit derivatives.

ACKNOWLEDGMENT

I would like to thank the editor and the reviewers for their constructive, careful and helpful feedback.

References

ANG, A., TIMMERMANN, A. (2012). Regime Changes and Financial Markets [online]. *Annual Review of Financial Economics*, 4(1): 313–337. <https://doi.org/10.1146/annurev-financial-110311-101808>.

ANSCOMBE, F. J., GLYNN, W. J. (1983). Distribution of Kurtosis Statistic for Normal Statistics [online]. *Biometrika*, 70(1): 227–234. <https://doi.org/10.1093/biomet/70.1.227>.

BASTIANIN, A. (2009). Modelling Asymmetric Dependence Using Copula Functions: an application to Value-at-Risk in the Energy Sector [online]. *FEEM Working Paper,* No 24/2009, Center for European Studies, Milan – Italy. <http://dx.doi.org/10.2139/ssrn.1425548>.

CARNERO, M. A. (2004). Persistence and Kurtosis in GARCH and Stochastic Volatility Models [online]. *Journal of Financial Econometrics*, 2(2): 319–342. <https://doi.org/10.1093/jjfinec/nbh012>.

CHEN, X., YIN, X. (2019). *Solve Nonlinear Optimization with Nonlinear Constraints, Version 0.6* [online]. CRAN. <https://cran.r-project.org/web/packages/NlcOptim/NlcOptim.pdf>.

CHRISTOFFERSEN, P. F. (2012). *Elements of Financial Risk Management.* 2nd Ed. Academic Press, Imprint of Elsevier Science, USA.

D'AGOSTINO, R. B. (1970). Transformation to Normality of the Null Distribution of G1 [online]. *Biometrika*, 57(3): 679–681. <https://doi.org/10.1093/biomet/57.3.679>.

DEMPSTER, A. P., LAIRD, N. M., RUBIN, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm [online]. *Journal of the Royal Statistical Society – Series B*, 39(1): 1–38. <https://www.jstor.org/stable/2984875>.

FANTAZZINI, D. (2008). Dynamic Copula Modelling for Value at Risk [online]. *Frontiers in Finance and Economics*, 5(2): 72–108. <https://ssrn.com/abstract=1542608>.

HAMILTON, J. D. (2016). Macroeconomic Regimes and Regime Shifts [online]. In: *Handbook of Macroeconomics*, Chapter 3, Elsevier, The Netherlands, 2: 163–201. <https://doi.org/10.1016/bs.hesmac.2016.03.004>.

HARVEY, A., RUIZ, E., SENTANA, E. (1992). Unobserved Component Time Series Models with ARCH Disturbances [online]. *Journal of Econometrics*, 52(1–2): 129–157. <https://doi.org/10.1016/0304-4076(92)90068-3>.

JARQUE, C. M., BERA, A. K. (1980). Efficient Test for Normality, Homoscedasticity and Serial Independence of Residuals [online]. *Economic Letters*, 6(3): 255–259. <https://doi.org/10.1016/0165-1765(80)90024-5>.

KLAASSEN, F. (2002). Improving GARCH Volatility Forecasts with Regime-Switching GARCH [online]. *Empirical Economics*, 27(2): 363–394. <https://doi.org/10.1007/s001810100100>.

KUPIEC, P. (1995). Techniques for Verifying the Accuracy of Risk Measurement Models [online]. *Journal of Derivative*, 3(2): 73–84. <https://doi.org/10.3905/jod.1995.407942>.

LJUNG, G. M., BOX, G. E. P. (1978). On a Measure of Lack of Fit in Time Series Models [online]. *Biometrika*, 65(2): 297–303. <https://doi.org/10.1093/biomet/65.2.297>.

LU, X. F., LAI, K. K., LIANG, L. (2014). Portfolio Value-at-Risk Estimation in Energy Futures Markets with Time-Varying Copula-GARCH Model [online]. *Annals of Operation Research*, 219(1): 333–357. <https://doi.org/10.1007/s10479-011-0900-9>.

MOSBAHI, M. N., SAIDANE, M., MESSABEB, S. (2017). Mixture of Probabilistic Factor Analyzers for Market Risk Measurement: Empirical Evidence from the Tunisian Foreign Exchange Market [online]. *Risk governance & control: financial markets & institutions*, 7(2): 158–169. <https://doi.org/10.22495/rgcv7i2c1p4>.

SAIDANE, M. (2022). Switching latent factor value-at-risk models for conditionally heteroskedastic portfolios: a comparative approach [online]. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 8(2): 282–307. <https://doi.org/10.1080/23737484.2022.2031346>.

SAIDANE, M. (2019). Forecasting Portfolio-Value-at-Risk with Mixed Factorial Hidden Markov Models [online]. *Croatian Operational Research Review*, 10(2): 241–255. <https://doi.org/10.17535/crorr.2019.0021>.

SAIDANE, M. (2017). A Monte-Carlo-based Latent Factor Modeling Approach with Time-Varying Volatility for Value-at-Risk Estimation: Case of the Tunisian Foreign Exchange Market [online]. *Industrial Engineering & Management Systems*, 16(3): 400–414. <https://doi.org/10.7232/iems.2017.16.3.271>.

SAIDANE, M., LAVERGNE, C. (2011). Can the GQARCH Latent Factor Model Improve the Prediction Performance of Multivariate Financial Time Series? [online]. *American Journal of Mathematical and Management Sciences*, 31(1–2): 73–116. <https://doi.org/10.1080/01966324.2011.10737801>.

SAIDANE, M., LAVERGNE, C. (2009). Optimal Prediction with Conditionally Heteroskedastic Factor Analysed Hidden Markov Models [online]. *Computational Economics*, 34(4): 323–364. <https://doi.org/10.1007/s10614-009-9181-7>.

SAIDANE, M., LAVERGNE, C. (2008). An EM-Based Viterbi Approximation Algorithm for Mixed-State Latent Factor Models [online]. *Communications in Statistics – Theory and Methods*, 37(17): 2795–2814. <https://doi.org/10.1080/03610920802040415>.

SAIDANE, M., LAVERGNE, C. (2007). Conditionally heteroscedastic factorial HMMs for Time Series in Finance [online]. *Applied Stochastic Models in Business and Industry*, 23(6): 503–529. <https://doi.org/10.1002/asmb.687>.

TSANG, E. P. K., CHEN, J. (2018). Regime Change Detection Using Directional Change Indicators in the Foreign Exchange Market to Chart Brexit [online]. *IEEE Transactions in Emerging Technology in Computational Intelligence*, 2(3): 185–193. <https://doi.org/10.1109/TETCI.2017.2775235>.

ROBUST 2022 (Volyně) 22nd Event of International Statistical Conference

Ondřej Vozár[1] | *Prague University of Economics and Business, Prague, Czech Republic*

The 22nd event of the well-established biennial statistical conference ROBUST 2022 took place in Volyně nearby Strakonice (Czech Republic) during 12–17 June 2022.[2] It was organised by joint effort of the Expert Group of Computational Statistics of the Czech Mathematical Society (Section of the Union of Czech Mathematicians and Physicists), Department of Probability and Mathematical Statistics of the Faculty of Mathematics and Physics, Charles University, Prague and the Czech Statistical Society. The conference took place in the buildings of the Technical College Volyně specialised in wood processing and wooden buildings. It is important to note that 21st ROBUST scheduled in Bardějov (Slovakia) was postponed several times because of Covid-19 pandemics and finally due to relative proximity of the battlefields of the conflict in Ukraine. Persistent chief ROBUST organisers decided to postpone it to the June of 2024.

In total, 78 participants from 5 countries presented and discussed contributions covering a broad spectrum ranging from theoretical statistics, probability and stochastic analysis, machine learning and computer science to applied statistics in several fields, including forestry, insurance and finance mathematics, medicine, health and epidemiology, metrology, traffic safety, online advertisement, and official statistics. Foreign participants came from Slovakia (12), Switzerland (2), the USA (1), and Nigeria (1). The idea behind the ROBUST conferences has always been to bring together statisticians of all generations and all fields from different Czech and Slovak institutions, Czech and Slovak experts living abroad to enable the exchange of ideas and to provide them with interdisciplinary insight into the research in statistics.

Five invited lectures were delivered. The leading expert in supercomputing, prof. George Ostrouchov (Oak Ridge National Laboratory and University of Tennessee, currently Fulbright Scholar at Faculty of Mathematics and Physics, Charles University, Prague) delivered a four-hour tutorial on high performance computing with R, utilising his deep knowledge of both theory and practical implementation (including long-term engagement in Ostrava supercomputer project IT4).

Dr. Pavel Charamza presented a broad range of real life implementation of statistical methods in credit risk, public opinion polls, and new media. He discussed the application of commonly recommended statistical methods and use of machine learning techniques.

Doc. Arnošt Komárek (Department of Probability and Mathematical Statistics, Faculty of Mathematics and Physics, Charles University, Prague) reflected his experience as a publicly speaking statistician during the Covid-19 pandemics. Main idea of the talk that statistician stood quiet and let other specialists (like data scientists, mathematical biologists) to ”steal” statistics provoked a very lively discussion in the audience.

---

[1] Department of Statistics and Probability, Faculty of Informatics and Statistics, Prague University of Economics and Business, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: vozo01@vse.cz. Also the Czech Statistical Office, Na padesátém 81, Unit for Mathematical and Statistical Methods and Statistical Quality, 100 82 Prague 10, Czech Republic. E-mail: ondrej.vozar@czso.cz.

[2] More at: <*https://robust.nipax.cz*>.

Doc. Jan Koláček (Department of Statistics and Mathematics, Faculty of Science, Masaryk University in Brno) delivered a functional data analysis. It was focused on an important modification for the real life setting - irregularly observed data.

Dr. Lukáš Lafférs (Department of Mathematics of the Faculty of Natural Sciences at the Matej Bel University in Banská Bystrica, Slovakia) provided an interesting presentation of a causal machine learning methodology. The first part was focused on presentation of this novel method. The second part was focused on applications of an economic policy assessment. The most interesting case study dealt with an assessment of the effects of interventions (for example, retraining courses provided by an employment department) for case of high dimensional data.

In this Conference, there is a long-term tradition of participation of doctoral and master degree students in the dedicated section, who presented 17 posters. Prizes were awarded in two categories: Master and First-Year Doctoral Students and Advanced Doctoral Students. The prizes were sponsored, as usual, by RSJ Securities, a.s., and RSJ Foundation. Conference fees for many participants (mostly master and first-year doctoral students) were sponsored, as in the past conferences, by RSJ and the Czech Statistical Society.

The Editor-in-Chief of the *Statistika: Statistics and Economy Journal* kindly invited participants to submit their papers on relevant topics to the Journal.

24th International Conference *Applications of Mathematics and Statistics in Economics (AMSE 2022)*

Stanislava Hronová[1] | *Prague University of Economics and Business, Prague, Czech Republic*

On 31 August to 4 September 2022 the 24th international conference *Applications of Mathematics and Statistics in Economy* took place in a beautiful landscape of the Jeseníky Mountains, in Velké Losiny town. This year's conference was organized by the Department of Statistics and Probability and Economic Statistics of the Faculty of Informatics and Statistics, Prague University of Economics and Business. The conference was attended by over 40 experts from the Czech Republic, Slovakia and Poland representing the Prague University of Economics and Business, Matej Bel University in Banská Bystrica, Wroclaw University of Economics and Business, University of Pardubice, University of Chemistry and Technology in Prague and the Czech Statistical Office.

The characteristic feature of this traditional trilateral conference is an exchange of knowledge and expertise, presentation of the latest results of the research and discussion about the new procedures and methods. This conference traditionally includes workshops of representatives of co-operating institutions as well as planning of the future trends of scientific and pedagogic co-operation. After the break caused by Covid pandemic when in 2020 the conference was cancelled and in 2021 was held as an on-line one-day event, this year's meeting was even more enjoyable and not only for regular conference participants but also for post graduate students having chance to meet their colleagues from co-operating institutions.

The programme of the conference was opened by the President of the Czech Statistical Office, Marek Rojíček, by his speech called *Official statistics between past and future*, where he summarized the principal tasks which the modern state statistics faces at present, i.e. how to react to changing needs of users, to new data sources while preserving the quality of the information provided. There is a permanently increasing demand to statistical recording of the phenomena such as social welfare, environmental indicators, financial transactions, global value chains, etc. In addition, appearing are new and rich data sources processed by retailers or telecommunication operators (Big data), which can be used by the state statistics. And, finally, it is absolutely necessary to optimize also the ways of communication of the statistical surveys results so that they are able to accommodate the needs of all users.

Other conference sessions were organised in six sections: Macroeconomic issues, Impact of Covid pandemic, Insurance and demography, Time series analysis methods, Statistical methods in Economy and History of statistics. It is very difficult to highlight the most interesting contributions; let me point

[1] Department of Economic Statistics, Faculty of Informatics and Statistics, Prague University of Economics and Business, Prague, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: hronova@vse.cz.

out just some of the presentations especially of young colleagues which I consider interesting and well prepared, when new methodologic approaches were applied.

In the section Macroeconomic issues the most interesting was the paper of J. Fischerová *National accounts and natural catastrophes – 20 years from devastating floods in the Czech Republic (case of insurance companies).* The author summarised the manner of recording of damage from disasters in the national accounts (from the aspect of national economy at one hand and of insurance companies on the other hand) and emphasized the changes in the ESA 2010 standard compared to the ESA 1995 standard in response to terrorist attracts and national disasters after 2000. She also showed the differences between earlier recording of claims in current transfers and present recording in other capital transfers and highlighted the values connected with natural disasters in time series 1993–2021 in the Czech Republic.

Presented papers in the section *Impact of Covid pandemic* were focused on topical items such as impact of the pandemic on various areas of the life of the whole society – on demography (*The identification of changes in the structure of causes of death in Poland, Czechia and Slovakia or Impact of Covid-19 pandemic on the population projections*), on insurance (*The influence of coronavirus pandemic on the widows' reverse annuity benefits*), on private businesses (*Models of self-employed persons termination in pre-Covid-19 and Covid-19 period*) or on health care (*Excess deaths during a pandemic – numbers, causes, recommendations*). The last of the above mentioned contributions (M. Biernacki and C. Kozyra) try to identify the main causes of the situation which can be distinguish from the hidden Covid-19 but also point out the poorer (e.g. delayed) access to medical service during pandemic restrictions. The effects of relationship between the health status, mortality and (non) administration of vaccines. The analysis was based of official data provided by Polish Institute of Public Health.

In section Insurance and demography it is necessary to highlight the contribution of J. Novák and J. Sixta called *Evaluation of synthetic microdata from population census 2011 through correspondence analysis.* Authors present an original method of synthetic simulation of microdata from the population census and their comparison to the original microdata A new (artificial) synthetic dataset of microdata was created while preserving the structure and relations between variables as in the original dataset. Synthetic microdata as a way of protecting microdata would enable the dissemination of microdata from the census via the SafeCentre of the Czech Statistical Office.

Contributions in section *Time series analysis* were focused mainly on the sphere of modelling of price and income development. Considering the importance and topicality of rapid price growth in the real estate market the paper *Model-based risk assessment of house-price developments* should be mentioned. The authors (M. Plašil and M. Andrle) presented a simple model-based approach to assess uncertainty associated with future house-price development conditioned by current state of the economy. The application on the Czech data proved usefulness, reliability and flexibility of this model approach.

In section *Statistical methods in economy* papers on different topics were presented focusing mainly on two areas – education (*Coefficient of economic demands: Czech and Slovak comparison of higher education funding* or *Can universities buy rank for money? World university rankings in the perspective of institutional budgets*) and water management (*Efficiency evaluation of water sector in the Czech Republic: network SBM approach of Efficiency evaluation for regulatory purposes – DEA and SFA-based case study of Czech water companies*). In academic milieu a broad discussion has arisen about the paper compiled by a young team of authors (Š. Stiburek, S. Kováč, S. Brožová and H. Flusková), analysing the relationship between the size of budget of universities and their international ranking while a special attention was paid to the results of universities in the Czech Republic, Slovakia and Poland. These universities show worse ranking than majority of schools from western and northern Europe having comparable funds in proportion to the number of students. However, the authors showed that analysis of the data for 2018 covering 246 universities from 20 European countries (ETER register and THE ranking) does not allow to confirm or refute the analysed causal relations. In other words, it is obvious that the size of budget

of a university does not guarantee a specific ranking of the respective university but it may determine the interval for ranking. Quality and relevance of different rankings and institutional background in which universities in individual countries act, i.e. aspects not included in analysis were widely discussed.

In respect of traditional section dealing with *History of statistics* we should mention the paper *Statistical methods in the publications of official doctors in the Bohemian Lands in the first half of the 19th century*, where authors (P. Závodský, and O. Šimpach) focused on formation of the state-run healthcare which includes also regional (or city) official doctors whose reports they analyse. A special attention was paid to important publications of F. A. Stelzig.

It has become a tradition to organize interesting trips for the Conference participants. A group of the most physically fit colleagues set for the highest hill of the Jeseníky Mountains called Praděd. However, majority of participants preferred to make excursion to Dlouhe Strane Pumped-Storage Hydro Power Station descending then to the village Kouty nad Desnou and admiring spectacular panorama of the Jeseníky Mountain. History lovers did not stay aside and visited chateau at Velké Losiny and other outstanding historical buildings nearby.

A complete AMSE 2022 programme including abstracts of presented papers see at: <*http://www.amse-conference.eu/*>. There you will also find the information about AMSE history and reference to previous years of this international event.[2]

Tradition of alternate organizing (Slovakia – Poland – Czech Republic) continues and the 25th AMSE conference will be prepared by the Department of Quantitative Methods of Matej Bel University, Banská Bystrica, at the end of August and beginning of September 2023 in Slovakia, at Rájecké Teplice.

[2] In this report on the Conference the texts of Book of abstracts <*www.amse-conference.eu*> were used.

International Conference *Interdisciplinary Information Management Talks (IDIMT 2022)*

Petr Doucek[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Lea Nedomová[2] | *Prague University of Economics and Business, Prague, Czech Republic*

The Interdisciplinary Information Management Talks (IDIMT)[3] conference is traditionally organized by the Department of Systems Analysis of the Faculty of Informatics and Statistics at the Prague University of Economics and Business, in co-operation with Johannes Kepler University Linz.

This year's conference session held from 7 to 9 September 2022 in Prague had a festive character, as it was the thirtieth time that the participants met. Especially for the older participants, the conference provided a venue to discuss the history of scientific work and teaching of computer science at both Johannes Kepler University Linz and Prague University Economics and Business. With the benefit of hindsight, it is very interesting to look back at what issues shook the world of business informatics in the past – from the conversion of text fields from one language to another, to the Year 2000 problem, to the current challenges of digitizing government processes, ensuring the security of information systems, privacy, ethical aspects of using information systems and technologies, and more.

The main topic of this year's event was "Digitization of Society, Business and Management in Pandemic". The conference attracted papers from a total of 149 authors, with 42 submitted papers being accepted together with 12 invited papers. The authors came from nine different countries: Australia, Austria, Czech Republic, Germany, Indonesia, Portugal, Russia, Slovakia and Slovenia. The conference was organized by the staff of the Faculty of Informatics and Statistics in the building of the Prague University of Economics and Business in Prague on its campus in Žižkov. After two years, we were able to return to the presentation form of the conference, where most of the international participants could come.

The plenary session of the conference was opened by Prof. Gerhard Chroust in his speech, where he recalled the humble beginnings of the cooperation between the two institutions in the 1990s, the founding members of the conference committee and emphasized the development of the whole issue of information management both in depth and breadth of the field. The traditional first highlight of the conference was the invited lecture Christian W. Loesch – IBM's former director for Central and Eastern Europe. He brought the conference from this general level down into individual topics, with a presentation entitled "Cornerstone Technology: Exposer, Risk and Future". His lectures have the ability to engage the audience both with such topics as the development of information technology and with the breadth of information content that Christin Loesch treats as the basis for his presentations. Another advantage of his presentations is the fact that he is able to grasp the topic of the development and future of information

[1] Faculty of Informatics and Statistics, Prague University of Economics and Business, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: doucek@vse.cz.
[2] Faculty of Informatics and Statistics, Prague University of Economics and Business, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: nedomova@vse.cz.
[3] More at: <*https://idimt.org*>.

technology not only from a technical point of view, but also deals with the economic dimension and, last but not least, the impact of the deployment of information technology in real business.

The eleven other sessions of the conference were divided into two parallel streams, which were started by the conference participants immediately after the plenary session.

The topics of the sessions covered a wide range of issues with which the discipline of information management is nowadays associated. The most attended session was the one on the topic of digitization. The theme is topical nowadays, as it reflects the European Union's efforts to transfer the processes of both business and state and public administration into cyberspace. This is reflected in the EU 2030 strategy and its elaboration on the conditions of the Czech Republic and other Member States. The issue of digitalization is also closely linked to the issue of security in two dimensions – the first is the protection of digital media content, which was addressed in the session "Cyber Security in a Digital World" and the second is the behaviour of systems (and not only digital) in the event of a crisis "Complex Digital Approaches for Crisis Management – Blackout in Pandemic Times." Both sessions, led by experienced colleagues from Austria, contained stimulating contributions that illustrated not only the benefits of digitalization, but also clearly highlighted its risks. This collection of sessions includes a session that addressed the issue of supplier-customer relations in the information society, "Smart Supply Chain."

The next group of papers that comprised the sections was devoted to innovation and change in everyday life as brought about by the Covid 19 epidemic in particular. This included the section "Innovations and Strategies in a Pandemic Era" and its de facto content-linked section "Virtual Collaboration & Exchange – Challenges and Emerging Approaches," which showed how some of the problems in human-to-human communication were addressed during the Covid 19 pandemic and where the benefits and risks lay. Among the invitational sketches, i.e. those that remained innovations, we can include the session "Smart Technologies for a Sustainable Green World", which addressed the pressing issues of the relationship between information technology and the environment. Traditional sessions include "Sustainability and Performance Management and Business Reporting" and "Social Media Authenticity and Transparency". The latter session dealt with media issues and activities in contemporary society, with an emphasis on social media and social networking. Here also appeared the nowadays much discussed issue of classification of information, its credibility, including the category of "fake news".

As the IDIMT conference also deals with the problems of informatics, one of its sessions was dedicated to the issue of application development. It was the session "Challenges and Trends in Software Development."

As a result of the conference, besides the presented results of the scientific work, the collaboration between Prague University of Economics and Business, Johannes Kepler University Linz and other universities, which were represented in the wide plenary of participants, was deepened.

This conference was partially co-funded through project IGA 409021 of the „Faculty of Informatics and Statistics, Prague University of Economics and Business, "Česká spořitelna, a.s." and Johannes Kepler University Linz, Austria.

Mathematical Methods in Economics (MME 2022) International Conference

Petra Zýková[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Josef Jablonský[2] | *Prague University of Economics and Business, Prague, Czech Republic*

The Mathematical Methods in Economics (MME) conference has a very long history and tradition. It is one of the most important scientific events organised in the Czech Republic in the field of operational research, econometrics, mathematical economics, and related research areas. In 2022, the 40th International Conference on Mathematical Methods in Economics was organised in the city of Jihlava from 7 to 9 September. In addition to the local organiser (the Department of Economic Studies, College of Polytechnics, Jihlava), leading organisers of the MME conference are the Czech Society for Operations Research (CSOR) and the Czech Econometric Society.[3]

The total number of participants in this year's MME conference was more than 80. Several participants chose virtual attendance, which was also possible this year. Participants came from the Czech Republic, Italy, China, Poland, Finland, Lithuania and Slovakia. The programme started with an opening ceremony, where the Chair of the Organising Committee, Martina Kuncová, introduced the main programme and all the facilities. After that, the plenary session started with two exciting lectures. The first one, titled *Labour Market Composite Indexes: Formulation, Estimation and Implications*, was presented by Doctor Jan Brůha from Czech National Bank. Professor Sebastiano Vitali from the University of Bergamo delivered the second plenary talk about *Stochastic optimization and stochastic dominance in Asset and Liability Management models.* After the plenary session, the conference was divided into three parallel sessions. The total number of presentations was more than 70. All accepted papers are published in the *Proceedings of the MME 2022.* As in previous years, they have been submitted for indexing in the Web of Science CPCI database.

It has been a long tradition for PhD students to compete for the best paper during MME conferences. The competition is organised and honoured by the CSOR. All submitted papers were peer-reviewed, and the programme committee further evaluated the papers with positive referee reports. Six best-selected papers were presented at the conference in a special session, and the evaluation committee decided on the winners. The five best papers were awarded after a conference dinner at Radniční restaurant. Lukáš Veverka (Prague University of Economics and Business, Prague, Czech Republic), with his paper 'The Optimal Settings of Genetic Algorithm for Variable Selection in a Non-linear Time Series Model', was the winner of the competition. Anna Selivanova (Czech University of Life Science, Prague, Czech Republic), with her paper 'Application of System Dynamics model of recovery', ranked second.

[1] Department of Econometrics, Faculty of Informatics and Statistics, Prague University of Economics and Business, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: petra.zykova@vse.cz.
[2] Department of Econometrics, Faculty of Informatics and Statistics, Prague University of Economics and Business, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: jablon@vse.cz.
[3] More at: <*https://mme2022.vspj.cz*>.

Petr Pokorný (Prague University of Economics and Business, Prague, Czech Republic) ranked third with his paper 'Reverse Channel Competition in Dual Sustainable Closed-Loop Supply Chain'. David Neděla (Technical University of Ostrava, Czech Republic) and Michal Tomíček (Technical University of Liberec, Czech Republic) won the remaining two awards.

The conference was organised at a great level. All sessions took place in the renovated buildings of the College of Polytechnics Jihlava. The welcome evening took place in the new hall of the College of Polytechnics Jihlava.

An essential part of all conferences is a social programme that offers many opportunities to discuss various problems in an informal environment. The organisers have prepared various combinations of several trips: a guided tour of the Zoo Jihlava, a guided tour of the church of St. Jacob, a guided tour of Jihlava Underground, Brewery, Gustav Mahler House or Gate of Holy Mother. The conference dinner took place at Radniční restaurant on Masaryk square.

This year's annual meeting of the CSOR decided that the 41st MME conference would be organised in the city of Prague by the Prague University of Economics and Business, the Faculty of Informatics and Statistics, the Department of Econometrics on 13–15 September 2023.

16th Year of the *International Days of Statistics and Economics (MSED 2022)*

Tomáš Löster[1] | *Prague University of Economics and Business, Prague, Czech Republic*
Jakub Danko[2] | *Prague University of Economics and Business, Prague, Czech Republic*

From 8th to 10th September 2022, a worldwide conference of the International Days of Statistics and Economics (MSED) took place at the Prague University of Economics and Business.[3] The conference belongs to traditional professional events; this year, the sixteenth year of this event was held. Prague University of Economics and Business (the Department of Statistics and Probability and the Department of Managerial Economics) was the main organizer, as usual; and was helped by the Faculty of Economics, the Technical University of Košice, and Ton Duc Thang University, as co-organizers. The conference ranks among important statistical and economic conferences, which can be proved by the fact that Online Conference Proceedings were included in the Conference Proceedings Citation Index (CPCI), which has been integrated within the Web of Science, Clarivate Analytics since 2011.

The traditional goal of this international scientific conference was a presentation of the contributions of individual authors and a discussion of current issues in the field of statistics, demography, economics, and management and their interconnection.

Due to the continuing global situation in connection with COVID 19, this year's conference and presentation were again in a hybrid form (online and real presentation at the university), which caused the participation of foreign nationals to be active.

The online implementation of the conference took place in individual channels of the conference teams in MS Teams (according to partial sections). The number of registered conference participants was a total of 156, of which 65 were foreign, e.g. Poland (26), Slovakia (13), etc. Among the conference participants were 19 doctoral students.

The received papers were first evaluated in terms of scientific content and suitability of the topic concerning the focus of the conference. After the exclusion of unsatisfactory abstracts, a double independent anonymous review procedure took place in the spring of this year.

We would also like to invite researchers, doctoral students, and the wide professional public to the seventeenth International Days of Statistics and Economics, which will take place at the Prague University of Economics and Business traditionally in early September 2023.

[1] Faculty of Informatics and Statistics, Prague University of Economics and Business, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: tomas.loster@vse.cz.
[2] Faculty of Informatics and Statistics, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: jakub.danko@vse.cz.
[3] More at: <*http://msed.vse.cz*>.

European Statistical System Peer Reviews

Eurostat and the national statistical authorities of all the EU and EFTA countries form a partnership called the European Statistical System (ESS). Together, they produce European statistics which respect a common quality framework. One instrument that ensures the implementation of the common quality framework and thus the quality of European statistics is the so-called ESS Peer Reviews.

Code of Practice

The common quality framework of the ESS is based on the *European Statistics Code of Practice*, a set of 16 principles covering the institutional environment, statistical processes, and statistical outputs.

The principles are complemented with a set of 84 indicators of best practices and standards to provide guidance and reference for reviewing the implementation of the Code (or CoP).

ESS Peer Reviews

Quality is the trademark of European statistics and makes them more trustworthy than other data that are readily available through many channels. To guarantee the quality of their statistics, the ESS created a common quality framework. The European Statistics Code of Practice is the cornerstone of this quality framework.

The objective of the peer reviews is to assess ESS members' the compliance with the principles and indicators of the Code. The subsequent recommendations should also help statistical authorities to further improve and develop their statistical systems.

All members of the ESS are reviewed, i.e. Eurostat and the national statistical authorities of the EU Member States and EFTA countries. Peer review expert teams are composed of four European experts in statistics, auditing and governance issues,

including an independent expert, to assess the national statistical authorities. An expert team from the European Statistical Governance Advisory Board (ESGAB) reviews Eurostat.

The peer reviews are carried out on a country-by-country basis according to these steps:

1. Each ESS national statistical authority assesses itself against the principles of the Code through a questionnaire and provides extensive documentation on its functioning;
2. These documents are checked and analysed by an expert team which subsequently carries out an in-country visit during which a further in-depth review is performed;
3. The expert team compiles a final report with recommendations for improvements;
4. This report is submitted to the national statistical authority for approval and the drafting of improvement actions.

The implementation of the improvement actions in the EU and EFTA countries is monitored on an annual basis by Eurostat. The implementation of the improvement actions for Eurostat is monitored by ESGAB.

Peer Review 2021–2023

Two previous rounds of Peer Reviews (2006–2008 and 2013–2015) were focused mainly on compliance with the European Statistics Code of Practice. Peer Review 2021–2023 (round III) will go further and help ESS partners to improve by making future-oriented recommendations that go beyond the current Code. In addition, future-oriented elements could help revise the Code to reflect new developments that experts will identify in this round.

Contents 2022 (No. 1–4, Vol. 102)

Statistika: Statistics and Economy Journal, Vol. 102 (1) 2022

ANALYSES

5 Stanislava Hronová, Luboš Marek, Richard Hindls
The Impact of Consumption Smoothing on the Development of the Czech Economy in the Most Recent 30 Years

20 Jaroslav Sixta, Karel Šafr
Productive Population and Czech Economy by 2060

35 Václav Rybáček
Deviations between Government Debt and Changes in Government Deficit, Why They Tend to Persist

46 KamilaTrzcińska
Income and Inequality Measures in Households in the Czech Republic and Poland based on Zenga Distribution

59 Veronika Jurčišinová, Ľubica Štiblárová
On What Really Matters: Evidence from Alternative Well-Being Indicator in EU-28 Countries

73 Ali Ari, Raif Cergibozan, Caner Demir
Did Covid-19 Precautions and Lockdowns Cause Better Air Quality? Empirical Findings from Turkish Provinces

84 Jan Kovanda
Material Flows Mobilized by Motor Vehicles and Transport Equipment Manufacturing and Use in the Czech Republic: an Application of Economy-Wide Material System Analysis

98 Ionut Pandelică, Cristina Popîrlan, Cristina Mihaela Barbu, Mihail Cristian Negulescu, Anca Madalina Bogdan, Simona Moise, Elena Bică, Mihaela Cocoșilă
A Demand-Supply Equilibrium Model – Study Case on the Electricity Market for Households from the Perspective of Prices Liberalization

INFORMATION

110 Conferences, Information

Statistika: Statistics and Economy Journal, Vol. 102 (2) 2022

ANALYSES

117 Jolana Gubalová, Petra Medveďová, Jana Špirková
The Global Pension Index of Slovakia

138 Peter Pisár, Alexandra Mertinková, Miroslav Šipikal, Mária Stachová
The Importance of Determinants of Transition from Unemployment to Self-Employment: Evidence from Slovak Micro-Data

153 Havanur Ergün Tatar, Gökhan Konat, Mehmet Temiz
The Relationship between Financial Development, Trade Openness and Economic Growth in Turkey: Evidence from Fourier Testst

168 Sonu Madan, Surender Mor
Is Gender Earnings Gap a Reality? Signals from Indian Labour Market

184 Fatih Chellai
Application of the Hybrid Forecasting Models to Road Traffic Accidents in Algeria

198 Juraj Medzihorský, Peter Krištofík
Can Individual Human Financial Behaviour Be Mathematically Modelled? A Case Study of Elon Musk's Dogecoin Tweetss

205 Jaromír Antoch, Francesco Mola, Ondřej Vozár
New Randomized Response Technique for Estimating the Population Total of a Quantitative Variable

INFORMATION

228 Conferences, Information

Statistika: Statistics and Economy Journal, Vol. 102 (3) 2022

ANALYSES

236 Joanna Dębicka, Edyta Mazurek, Katarzyna Ostasiewicz
Methodological Aspects of Measuring Preferences Using the Rank and Thurstone Scale

249 Jaroslav Horníček, Hana Řezanková
Missing Data Imputation for Categorical Variables

261 Joanna Dębicka, Stanisław Heilpern, Agnieszka Marciniuk
Modelling Marital Reverse Annuity Contract in a Stochastic Economic Environment

282 Piotr Sulewski, Jacek Białek
Probability Distribution Modeling of Scanner Prices and Relative Prices

299 Jiří Novák
Population Census Microdata Availability

331 Joanna Adrianowska
Selected Coefficients of Demographic Old Age in Traditional and Potential Terms on the Example of Poland and Czechia

CONSULTATION

347 Jakub Vincenc
Fisim Methodology and Options of Its Estimation: the Case of the Czech Republic in the Years 1993–2019

INFORMATION

360 Obituary Notice

362 Conferences

Statistika: Statistics and Economy Journal, Vol. 102 (4) 2022

ANALYSES

369 Boris Marton, Alena Mojsejová
Macroeconomic Indicators and Subjective Well-Being: Evidence from the European Union

382 Viera Pacáková, Ľubica Sipková, Petr Šild
Factors of Differences in the Highest Wages of Employees in the Slovak Republic (2020 vs. 2010)

396 Jana Cibulková, Barbora Kupková
Review of Visualization Methods for Categorical Data in Cluster Analysis

409 Vladimír Mucha, Ivana Faybíková, Ingrid Krčová
Use of Markov Chain Simulation in Long Term Care Insurance

426 Nataliia Versal, Vasyl Erastov, Mariia Balytska, Ihor Honchar
Digitalization Index: Case for Banking System

443 Guan-Yuan Wang
Churn Prediction for High-Value Players in Freemium Mobile Games: Using Random Under-Sampling

454 Mohamed Saidane
A New Viterbi-Based Decoding Strategy for Market Risk Tracking: an Application to the Tunisian Foreign Debt Portfolio During 2010–2012

INFORMATION

471 Ondřej Vozár
ROBUST 2022 (Volyně) 22nd Event of International Statistical Conference

473 Stanislava Hronová
24th International Conference *Applications of Mathematics and Statistics in Economics (AMSE 2022)*

476 Petr Doucek, Lea Nedomová
International Conference *Interdisciplinary Information Management Talks (IDIMT 2022)*

478 Petra Zýková, Josef Jablonský
Mathematical Methods in Economics (MME 2022) International Conference

480 Tomáš Löster, Jakub Danko
16th Year of the *International Days of Statistics and Economics (MSED 2022)*

481 European Statistical System Peer Reviews

483 Contents 2022 (Vol.102)

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 complex and advanced analyses based on the official statistics data focused on economic, environmental, social and other topics. Papers shall have up to 12 000 words or up to 20 1.5-spaced pages.

Discussion brings the opportunity to openly discuss the current or more general statistical or economic issues, in short what the authors would like to contribute to the scientific debate. Contribution shall have up to 6 000 words or up to 10 1.5-spaced pages.

In the ***Methodology*** section we publish articles dealing with possible approaches and methods of researching and exploring social, economic, environmental and other phenomena or indicators. Articles shall have up to 12 000 words or up to 20 1.5-spaced pages.

Consultation contains papers focused primarily on new perspectives or innovative approaches in statistics or economics about which the authors would like to inform the professional public. Consultation shall have up to 6 000 words or up to 10 1.5-spaced pages.

Book Review evaluates selected titles of recent books from the official statistics field (published in the Czech Republic or abroad). Reviews shall have up to 600 words or 1–2 1.5-spaced pages.

The ***Information*** section includes informative (descriptive) texts, information on latest publications (issued not only by the Czech Statistical Office), or recent and upcoming scientific conferences. Recommended 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 — Authors and Contacts — Abstract (max. 160 words) — Keywords (max. 6 words / phrases) — Introduction — 1 Literature survey — 2 Methods — 3 Results — 4 Discussion — Conclusion — Acknowledgments — References — Annex (Appendix) — Tables and Figures (for print at the end of the paper; for the review process shall be placed in the text).

Authors and Contacts
Rudolf Novak,1 Institution Name, City, Country
Jonathan Davis, Institution Name, City, Country
1 Address. Corresponding author: e-mail: rudolf.novak @domainname.cz, phone: (+420)111222333.

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 references in the text enclosing authors' names and the year of the reference, e.g. „…White (2009) points out that…", „…recent literature (Atkinson and Black, 2010a, 2010b, 2011; Chase et al., 2011: 12–14) conclude…". Note the use of alphabetical order. Between the names of two authors please insert „and", for more authors we recommend to put „et al.". Include page numbers if appropriate.

List of References
Arrange list of references alphabetically. Use the following reference styles: [book] HICKS, J. (1939). *Value and Capital: An Inquiry into Some Fundamental Principles of Economic Theory.* 1st Ed. Oxford: Clarendon Press. [chapter in an edited book] DASGUPTA, P. et al. (1999). 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. [on-line source] CZECH COAL. (2008). *Annual Report and Financial Statement 2007* [online]. Prague: Czech Coal. [cit. 20.9.2008]. <http://www.czechcoal.cz/cs/ur/zprava/ur2007cz.pdf>. [article in a journal] HRONOVÁ, S., HINDLS, R., ČABLA, A. (2011). Conjunctural Evolution of the Czech Economy. *Statistika: Statistics and Economy Journal*, 91(3): 4–17. [article in a journal with DOI]: Stewart, M. B. (2004). The Employment Effects of the National Minimum Wage [online]. *The Economic Journal*, 114(494): 110–116. <http://doi.org/10.1111/j.0013-0133.2003.0020.x>.
Please **add DOI numbers** to all articles where appropriate (prescribed format = link, see above).

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 and numbered.

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. Articles for the review process are accepted continuously and may contain tables and figures in the text (for final graphical typesetting must be supplied separately as specified in the instructions above). Please be informed about our Publication Ethics rules (i.e. authors responsibilities) published at: http://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 66 per copy + postage.
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
Czech Statistical Office | Information Services
Na padesátém 81 | 100 82 Prague 10 | Czech Republic
Phone: (+420) 274 052 733, (+420) 274 052 783
E-mail: objednavky@czso.cz

Design: Toman Design
Layout: Ondřej Pazdera
Typesetting: Družstvo TISKOGRAF, David Hošek
Print: Czech Statistical Office

102nd 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