STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **99** (4) 2019

100 CENTURY OF STATISTICS

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

EXECUTIVE BOARD

MANAGING EDITOR

Dear Readers,

In 2018, we celebrated the 100th anniversary of foundation of Czechoslovakia. This year, the Czech Statistical Office and the Statistical Office of the Slovak Republic celebrate the 100th anniversary of the state official statistics in our two countries (the State Statistical Office was founded in Czechoslovakia in 1919). *Statistika: Statistics and Economy Journal* will also celebrate its 100th anniversary (volume) in 2020 – the journal follows the tradition of the *Československý statistický věstník* (Czechoslovak Statistical Bulletin), established back in 1920.

Through this special journal issue of our scientific peer-reviewed quarterly (which is last in 2019), we want to complete this jubilee year of the 100th anniversary of the Czechoslovak statistics – to commemorate and remind events, previous developments as well as current quality and state of research in official statistics in our countries. Therefore, it is symbolically composed of articles by Czech and Slovak authors only.

We believe that papers published in this special anniversary issue will be interesting and beneficial for all its readers. We are looking for further cooperation (not only) with authors (and reviewers) from our two countries and wish all our colleagues, partners, and collaborators plenty of creative thoughts, professional success, and satisfaction.

Marek Rojíček
President of the Czech Statistical Office

Alexander Ballek
President of the Statistical Office of the Slovak Republic

CONTENTS

ANALYSES

350 Stanislava Hronová, Richard Hindls, Luboš Marek
Economic Behaviour of the General Government and Sustainability of Public Finances – Comparative Analysis of the Czech Republic and Selected EU Countries

369 Petra Cisková, Ina Ďurčeková
Determinants of Firms´ Innovation Activities in V4 Countries

383 Martina Mysíková, Tomáš Želinský
On the Measurement of the Income Poverty Rate: the Equivalence Scale across Europe

398 Viera Labudová, Ľubica Sipková
Housing Affordability in Slovakia: what Factors Affect it?

417 Jaroslav Kraus
Spatial Autocorrelation of a Demographic Phenomenon: a Case of One-Family Households and One-Person Households

434 Mária Vojtková, Eva Kotlebová, Daniela Sivašová
Determinants Affecting Health of Slovak Population and their Qualification

451 Marek Strežo, Vladimír Mucha, Erik Šoltés, Michal Páleš
Risk Premium Prediction of Motor Hull Insurance Using Generalized Linear Models

CONSULTATION

468 Miluše Kavěnová
Recent Developments and Challenges in Energy Statistics in the Czech Republic

100th ANNIVERSARY

475 Marek Rojíček
Official Statistics between Past and Future

INFORMATION

481 Publications, Information, Conferences

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.

Acknowledgement

The journal of Statistika, its Executive Board, and the Czech Statistical Office would like to thank to the Faculty of Economic Informatics of the University of Economics in Bratislava, Slovak Republic, and to the Slovak Society for Economic Informatics for the funds provided for preparation of this special anniversary issue.

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

Economic Behaviour of the General Government and Sustainability of Public Finances – Comparative Analysis of the Czech Republic and Selected EU Countries

Stanislava Hronová[1] | *University of Economics, Prague, Czech Republic*
Richard Hindls[2] | *University of Economics, Prague, Czech Republic*
Luboš Marek[3] | *University of Economics, Prague, Czech Republic*

Abstract

The economic behaviour of the general government sector is manifested in the indices such as the government revenue, government expenditure, government deficit and government debt. It is an important tool for evaluating the sustainability of public finances and the orientation of the economic policy. All developed countries were hit by the crisis in 2009 and its continuation in the years 2011 through 2013; it was reflected, in particular, in high values of the government deficit and debt. The European economies have gotten out of this crisis by now, but a question remains: what means did the government institutions use in the respective countries to cope with the unfavourable values of the deficit (and debt)? In this paper, we will make use of the data on the national accounts to show the economic evolution of the general government in the Czech Republic after 2009 and compare it with certain other EU countries.

Keywords

National accounts, general government, government deficit, government debt, sustainability of public finances

JEL code

E21, C82

[1] Faculty of Informatics and Statistics, Department of Economic Statistics, University of Economics, Prague, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: hronova@vse.cz.
[2] Faculty of Informatics and Statistics, Department of Statistics and Probability, University of Economics, Prague, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: hindls@vse.cz.
[3] Faculty of Informatics and Statistics, Department of Statistics and Probability, University of Economics, Prague, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: marek@vse.cz.

INTRODUCTION

The general government is one of the institutional resident sectors. The importance of that sector, measured by its proportion in the gross domestic product, ranks as third, after the non-financial corporations and households. Nonetheless, its economic results are monitored with great interest not only by the creators of the economic policy but also by the top executive authorities in each country. Both the national and EU institutions carefully watch the deficits and debts of the general government.

In each calendar year, the evolution of the government deficit results in a relationship between its revenue and expenditure,[4] which sensitively respond to changes in any index entering into the total revenue and total expenditure. The evolution of the government debt depends not only on the year-to-year government deficit but also on the ability and options the general government has at its disposal to pay up the debt, as well as on other factors based on the definition of the debt. At times of favourable economic development, the government revenue should be growing faster than their expenditure (possibly even create a government surplus), their debts should (under comparable conditions) be decreasing. On the contrary, at times of recessions/crises, the general government sector falls into a deep deficit, and the debt suddenly goes up. A solution should include revenue growing faster than expenditure and stimulating economic growth. However, if the revenue does not grow fast enough, the only remaining way of decreasing the deficit is that of limiting the expenditure. This way, as a rule, leads to inhibiting the economic development and the consequent slowdown in the recovery of the national economy as a whole (which was the case of the Czech Republic in 2012 and 2013). An exception from the deficit-decreasing concept prevailing at times of reduced economic performance, may include preference on investment activities, in particular, those focused on new technologies, science, research, transportation infrastructure, etc. – such activities may temporarily increase the deficit but are aimed at its long-term reduction, and therefore at reducing the debt.

The monetary and financial crisis that, about ten years ago, hit all developed countries was manifested in the Czech Republic and other EU countries by a drop in economic activities (decreasing GDP) and a sudden deterioration of the government deficit (in absolute numbers and relatively with respect to the GDP value). The transition to the recovery stage was different in each country; it was especially complicated and lengthy in the Czech Republic. Despite that factor, in the Czech Republic, the general government's activities resulted in surplus as early as in 2016. In the present paper, we will have a look at the path the Czech general government took after 2009 and the methods of coping with their respective economic crises chosen by general governments in other EU countries. Our analysis will be based on the data from the national accounts of the Czech Republic and of selected EU countries.

1 THE TASK AND MISSION OF THE GENERAL GOVERNMENT SECTOR[5]

The general government sector puts together all institutional units whose main economic function is the provision of non-market services and/or the distribution of the national income and worth, as well as the units administering the social security funds. These units' main scope of activities follows from the mandatory direct and indirect payments (taxes and social contributions) from units ranging over all sectors. The institutional units included in this sector are non-market producers, whose production goes to the individual and collective final consumption. This sector mainly contains the state with all its authorities having general and specific areas of competency and directly subordinated to the state administration. It further contains social security authorities, local administration and different institutions directly governed by them; mainly organisations that are independent institutional units and, to a prevailing extent, funded by the state (from the central or local budget).

[4] The revenue and expenditure of the general governments are, throughout the entire text, understood as entered on the national accounts, that is, based on the accrual and not the cash principle.

[5] Loosely following the contents of Hronová, Sixta, Fischer, Hindls (2019).

The units in the general government sector mainly provide non-market services. However, this sector's production also has its market portion: in its institutional units, we can find those producing goods and market services. The proportion of this market output is negligible in comparison with the volume of the non-market output (this proportion in the Czech Republic does not get over 7% on a long-term basis). The economic significance of the general government sector – measured by the proportion of its gross value added in the gross value added of the total economy – is between 10% and 20% in the EU counties. There is large variability in this value among the EU countries depending on the different scopes of the production created and provided in favour of the society as a whole. A high proportion prevails in the "traditional social states" such as France and Scandinavian countries (around 18%–20%). The general government sector's proportion in the gross value added of the total economy in the Czech Republic fluctuates around 15% (the EU-28 average value is between 14% and 15%).

The main resources for funding the general government's activities come from the mandatory payments, which the other sectors must pay to it, that is, taxes and social contributions. Out of such resources, the government mainly:

- provides the funding for its activities – this is mainly seen in the intermediate consumption and compensation of employees indices;
- redistributes the income by providing subsidies and investment grants, as well as social benefits;
- ensures the functions of the national economy via investments into the infrastructure, environment, science and research, and defence and security;
- provides the funding to the health-care system, education, culture and sports – such funding is manifested in the final consumption expenditure indices.

The balance of the general government (surplus/deficit) is, in every year, given by a difference between its revenue and expenditure. On the national accounts, this result is recorded as its net lending/borrowing; the proportion of that index in the GDP is one of the so-called Maastricht criteria. From the above-mentioned considerations, it is clear that the economic result will be found directly on the general government sector's account, unlike the values of its revenue and expenditure – those are not explicitly stated in the annual report of the national accounts[6]. The rules for computing the general government's revenue and expenditure values[7] are based on the data entered on the sector's account so that their difference corresponds to the net lending/borrowing with respect to the realistic amounts of the total revenue/expenditure.

2 GOVERNMENT REVENUE AND EXPENDITURE[8]

Net lending/borrowing of the general government sector is the balance of the non-financial and financial account of that sector. When identifying which indices should be included in the government revenue and which in its expenditure, we have to keep in mind that certain indices occur twice on the general government's account (individual consumption expenditure vs. social transfers in kind, or collective consumption expenditure vs. actual final consumption); moreover, some of them do not have a character of real monetary flows (non-market output and final consumption expenditure). For this reason, it is necessary to exactly say which items and to what extent will actually be included in the government revenue and expenditure (in the sense of the national accounts); and we must first identify *internationally comparable values of the government revenue and expenditure*.[9]

---

[6] The Czech Statistical Office, as a rule, publishes such data only relative with respect to the GDP value; in certain countries (such as France), data of selected items and total amounts of revenue and expenditure are published with the frameworks of the so-called sector analyses.

[7] Cf. ESA 2010, Chap. 20.

[8] Loosely following the contents of Hronová, Sixta, Fischer, Hindls (2019).

[9] The international comparability is based on the rules implied by the ESA 2010 Standard, Chap. 20.

The first problem, i.e., the double occurrence of final consumption expenditure, can be resolved easily: we exclude from the total expenditure the social transfers in kind and the actual final consumption from among the indices present on the "uses" side of the general government sector's account;[10] we only leave there – with an exception mentioned below – the final consumption expenditure. The second problem is in reflecting the non-market output in the total revenue and the final consumption expenditure in the total expenditure because the said indices do not correspond to real receivables (payables) and their inclusion in the total values would, as an "artefact", make those total values apparently higher than they really are. The requirement that the total values of the revenue and expenditure should be realistic is very important because in the Czech Republic the government revenue is included in a basis for the Derivation of Expenditure Frameworks of the State Budget and State Funds, submitted by the Ministry of Finance of the Czech Republic within the framework of the Budget Strategy for the Public Institutions Sector.[11]

If the government revenue included the entire value of the non-market output (as given on the production account; let us denote it by *P.13*),[12] the overall amount of the revenue would be overvalued. The non-market output does not generate any revenue because the general government does not "sell" this type of production. It is concerned with the value of the goods and services provided by the general government to the society as a whole for free (or nearly for free). The government revenue, therefore, includes not the total value of this non-market output but only its part representing the actual income generated by the non-market activities. These are the so-called *payments for non-market output* (*P.131*). It is the part of the non-market output provided to households; in return, the general government obtains the payments that correspond to the relevant revenue item. In other words, the payments for the non-market output equals a remainder after "subtracting" the "real" non-market output for which the general government will not obtain any payments. This "real" non-market output consists of the collective consumption expenditure (*P.32*) and the social transfers in kind–non market production (*D.631*). For the payments for non-market output (*P.131*), it is thus true that

$$P.131 = P.13 - (P.32 + D.631). \tag{1}$$

If the entire final consumption expenditure were included in the total expenditure, the latter would again be overvalued. Hence only the "real expense" is entered into the expenditure, which equals the social transfers in kind–purchased market production (*D.632*). As a logical consequence the value that enters into the final balance of a difference between the government revenue and government expenditure (*P.131–D.632*) thus equals a value obtained by inclusion of the total non-market output (*P.13*) in the government revenue and the final consumption expenditure (*P.3*) in the government expenditure; at the same time, the total values of expenditure and revenue are not overestimated. In other words, the balance (expressed as the net lending/borrowing value) is the same as if we included the entire final consumption expenditure (*P.3*) into the total expenditure and the non-market output from the production account (*P.13*) into the total revenue. The following formula holds

$$P.131 - D.632 = (P.13 - P.32 - D.631) - D.632 = P.13 - P.3. \tag{2}$$

[10] Altogether they correspond to the value of the final consumption expenditure.

[11] The "public institutions sector" is a term introduced in Act No. 23/2017 Coll., on the budget responsibility rules for the general government sector (S.13, cf. ESA 2010). Nevertheless, the terms "public institutions" pursuant to Act No. 23/2017 Coll. and "general government" pursuant to ESA 2010 both refer to the same group of subjects; for more details, cf. Vebrová and Rybáček (2018).

[12] For the indices here and in Formulas (1) and (2) and in Table 1 we make use their national account codes – cf. ESA 2010.

To sum up the considerations mentioned above, the indices from the non-financial account of the general government are included in the government expenditure: intermediate consumption + compensation of employees + taxes on production and imports (payable) + subsidies (payable) + property income (payable) + current taxes on income and worth (payable) + social benefits other than social transfers in kind + other current transfers (payable) + capital transfers (payable) + gross capital formation + acquisition less disposal of non-produced assets + social transfers in kind–purchased market production.

The government revenue includes the following indices taken from the non-financial account of the general government: market output + output for own final use + taxes on production and imports (receivable) + subsidies (receivable) + property income (receivable) + current taxes on income and worth (receivable) + social contributions + other current transfers (receivable) + capital transfers (receivable) + payments for non-market output. Table 1 shows the values of the indices entering the total amounts of the government revenue and expenditure taken from the national accounts of the Czech Republic in 2018.

Table 1 Items of the government revenue and expenditure in the Czech Republic in 2018 (mil. CZK, current prices)

Code	Expenditure		Code	Revenue	
P.2	Intermediate consumption	324 994	P.11	Market output	28 063
D.1	Compensation of employees	520 623	P.12	Output for own final use	34 988
D.29	Taxes on production and imports	1 116	D.2	Taxes on production and imports	658 487
D.3	Subsidies	120 684	D.4	Property income	35 274
D.4	Property income	40 444	D.5	Current taxes on income and worth	417 057
D.5	Current taxes on income and worth	4 829	D.61	Social contributions	833 820
D.62	Social benefits[13]	628 600	D.7	Other current transfers	50 342
D.7	Other current transfers	102 912	D.9	Capital transfers	43 216
D.9	Capital transfers	33 912	P.131	*Payments for non-market output*	109 575
P.5	Gross capital formation	224 233			
NP	Acquisition less disposal of non-produced assets	–1 606			
D.632	*Social transfers in kind – purchased market production*	162 654			
Total expenditure		**2 163 395**	**Total revenue**		**2 210 822**
Revenue – Expenditure		47 427			

Explanations: From Formula (1), it is true that: *P.131 = P.13 – (P.32 + D.631)* = 1 011 052 – (500 191 + 401 286) = 109 575. From Formula (2), it is true that: *P.131 – D.632 = P.13 – P.3* = 109 575 – 162 654 = 1 011 052 – 1 064 131 = –53 079, where *P.3 = D.631 + D.632 + P.32* = 401 286 + 162 654 + 500 191 = 1 064 131 and *P.13* = 1 011 052.
Source: <www.czso.cz>

The internationally comparable values of the government revenue and expenditure enable us to carry out time- and space-based analyses of relative indices. As already pointed out in the Introduction, we will

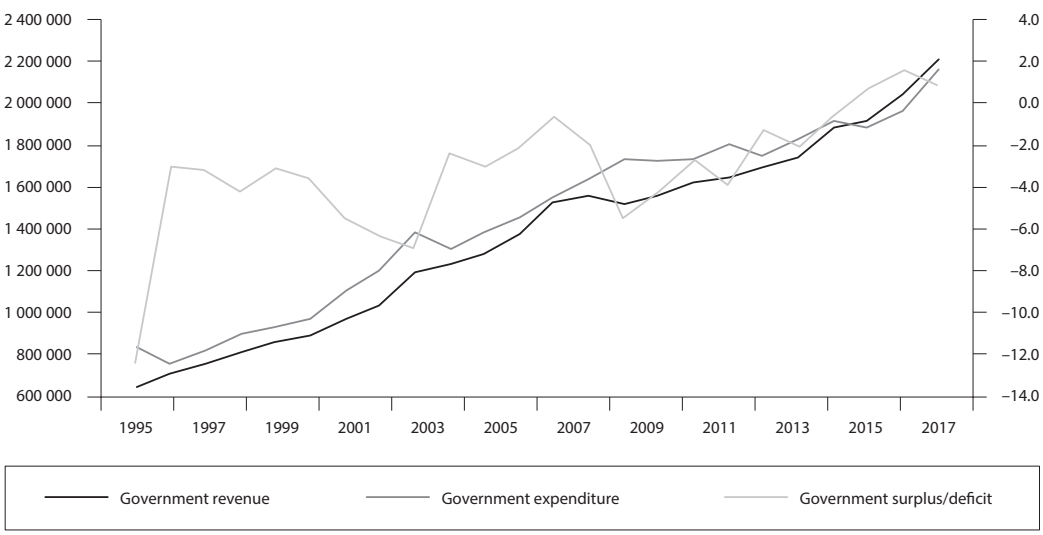
[13] Social benefits other than social transfers in kind.

show the way in which the general government in the Czech Republic coped with the economic deficits after 2009 and compare it with certain other EU countries.

3 ECONOMIC BEHAVIOUR OF THE GENERAL GOVERNMENT SECTOR IN THE CZECH REPUBLIC

Having a look at the long-term evolution (since 1995) of the government revenue and expenditure in the Czech Republic, we can say that the current-price revenue values were growing in the entire period in question except for 2009, when the year-to-year decrease (by 32.6 bil. CZK, i.e., by 2.1%) was predominantly caused by a drop in collected income tax and social contributions due to a drop in economic activities (with a year-to-year decrease in the GDP by 4.8%). The government expenditure in current prices has grown every year except for years 1996, 2004, 2010, 2013, and 2016. The decreasing expenditure values in 2004, 2010, and 2016 were caused by a significant drop in the gross fixed capital formation (by 62.6 bil. CZK, i.e., by 28.8% in 2004; by 34.6 bil. CZK, i.e., by 14.6% in 2010; and by 81.1 bil. CZK, i.e., by 34.3% in 2016). The decreasing expenditure values in 1996 and 2013 were mainly caused by a decrease in the amount of the payable capital transfers (by 166.9 bil. CZK, i.e., by 77.0% in 1996; and by 85.1 bil. CZK, i.e., by 67.5% in 2013). In both of these instances, extraordinary circumstances were connected with the economic and political transformation in the Czech Republic – the amount of other capital transfers included in 1995 the value (of approx. 190 bil. CZK) of the shares transferred to households within the framework of the second wave of the Voucher Privatisation; within the so-called Church Restitutions, churches obtained the first instalment of 59.5 bil. CZK in 2012; smaller instalments followed as late as 2014 (22.2 bil. CZK); 2015 (28.0 bil. CZK); and 2016 (15.7 bil. CZK); they, however, did not significantly affect the evolution of the total government expenditure. Figure 1 illustrates the evolution of the mutual relationship between the government revenue and expenditure in current prices.

Figure 1 Revenue and expenditure (mil. CZK, current prices) and deficit/surplus (in % of GDP) of the general government in the Czech Republic



Source: <www.czso.cz>

The mutual relationship between the government expenditure and revenue values is reflected in the government net lending/borrowing, which is a proportion of the government deficit/surplus expressed

in % of the GDP.[14] Let us identify the causes for the significant fluctuations in this index value: except for 2009, they are again given by the extraordinary circumstances related to the economic and political transformation in the Czech Republic. Apart from the already mentioned years 1996 (when the deficit amounted to 12.4% of GDP) and 2012 (with a deficit at 3.9% of GDP), high values of the deficit occurred in 2001 through 2003 due to the increased expenditure included in other capital transfers. Namely, there were concerned with the stabilisation of the banking sector at about 100 bil. CZK in each of the above-mentioned years with the consequent deficit values at more than 6% of GDP in 2002 and 2003, and 5.5% in 2001.

In 2009, the high value of the government deficit (5.5% of GDP) was caused by a drop in the economic performance of the Czech Republic, reflected in a year-to-year decrease in the revenue by 2.1% while the expenditure went up by 6.2%. Since that year, the deficit with respect to the GDP has been going down (except for 2014 when the collected excise taxes were lower on the revenue side, and the paid Church Restitutions were higher – cf. above). Table 2 illustrates the evolution of the government's revenue, expenditure, deficit/surplus and debt in the Czech Republic.

Table 2 Selected indices of the general government, Czech Republic (in % of GDP)

Index	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Government surplus/deficit	–0.7	–2.0	–5.5	–4.2	–2.7	–3.9	–1.2	–2.1	–0.6	0.7	1.5	0.9
Government revenue	39.7	38.7	38.7	39.3	40.3	40.5	41.4	40.3	41.1	40.2	40.5	41.5
Government expenditure	40.4	40.6	44.2	43.5	43.0	44.5	42.6	42.4	41.7	39.5	38.9	40.6
Government debt[15]	27.5	28.3	33.6	37.4	40.0	44.5	44.9	42.2	40.0	36.8	34.7	32.6

Source: <www.czso.cz>, the authors' own calculations

The considerations mentioned above imply that the fluctuations in the values of the government revenue and expenditure, as well as the deficit, were often caused by extraordinary circumstances not directly related to the economic behaviour of the sector. Let us now have a closer look at the situation after 2009 and study the factors that significantly affected the evolution in the government balance.

The government revenue (in current prices) went up by 45.2% in 2018 as compared with 2009; the same comparison in the expenditure amounted to 24.5%. The most quickly growing components of the expenditure were taxes on production and imports[16] (higher by 54.9%), current taxes on income and worth (higher by 49.9%), and the social contributions (higher by 49.0%).[17] The volume of the collected taxes[18] from production and imports was growing in the entire period in question after 2009 (with the sole exception of 2014 – a decrease by 2.2%). However, the high growth rates of the collected current taxes and social contributions are mainly implied by the low comparison base of 2009 (with a year-to-year drop in the amount of the collected current taxes by 11.1%); the volume of current taxes equal to that

[14] The government deficit/surplus proportion with respect to the GDP is one of the so-called Maastricht criteria; its value should not exceed a level of 3%.

[15] This is consolidated gross debt for the purposes of the EDP (Excessive Deficit Procedure); for more details, cf. Hronová, Sixta, Fischer, Hindls (2019).

[16] The value added tax has the highest proportion in the taxes on production and imports.

[17] The most quickly growing item of the expenditure in the period 2009–2018 was that of the miscellaneous current transfers (higher by 71.3%); however, the amount of this item is approx. 6% of the social contributions' volume.

[18] Here and below we use the term "collected taxes" even if the national accounts do not record data based on cash principle.

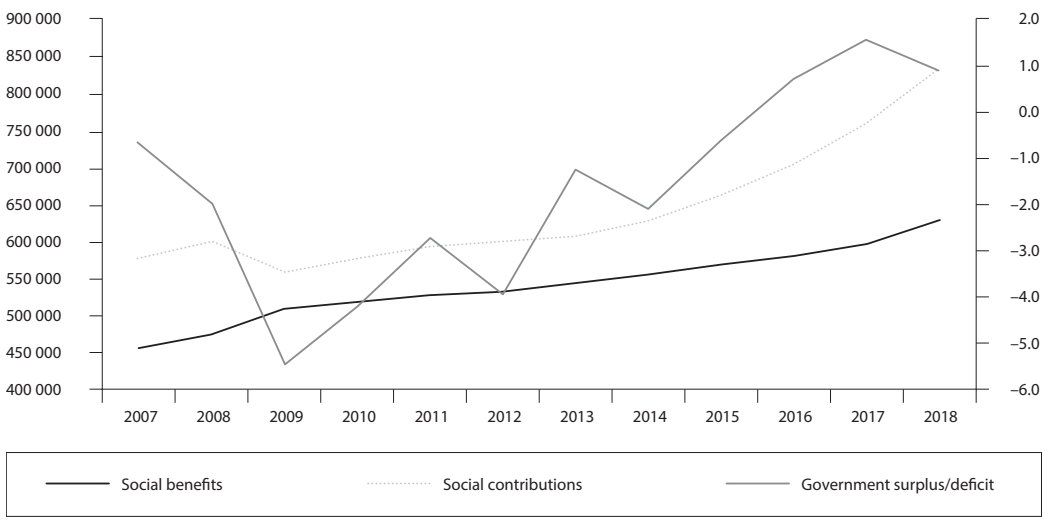
of the pre-crisis year 2008 occurred as late as in 2013. A similar phenomenon is valid for the collected social contributions: a year-to-year decrease in the social contributions amounted to 6.6% in 2009; the amounts of the social contributions equal to that of 2008 was achieved as late as in 2012.

The most important items in the government expenditure are social benefits other than social transfers in kind (hereinafter called just social benefits for the sake of simplicity).[19] The volume of the social benefits payable was continuously growing in the entire period in question (higher by 23.5% as compared with 2009) due to changes in the social policy and the ageing of the population. The most quickly growing component of the expenditure was that of the subsidies (higher by 75.0%) and miscellaneous current transfers (higher by 49.7%). Their total volume amounts to about a third of the expenditure incurred on social benefits. The amount of the compensation of employees also grew faster (higher by 47.9%) than the total expenditure especially due to high year-to-year increases in 2017 and 2018 (on average, by approx. 10% a year) in connection with the salary increases in public institutions.

The opposite direction (decrease in expenditure) can be observed in the gross fixed capital formation (down by 7.8%) and property income, or interest, related to gradually decreasing the government debt (down by 17%). However, a drop in investments into fixed capital cannot be viewed as a positive feature.

A large difference between the revenue growth and the expenditure growth of the general government in the Czech Republic (20.7 p.p. – percentage points) was, logically, manifested in the gradual improvement of the government balance and the consequent decrease in the government debt (cf. Table 2). Let us now have a look at the evolution of the most important items occurring within the government expenditure, and at the related evolution of the government balance (deficit/surplus).

Figure 2 Social benefits (payable) (mil. CZK, current prices) and government deficit/surplus, Czech Republic (in % of GDP)[20]



Source: <www.czso.cz>

[19] Social benefits paid in old age, invalidity, disease, maternity, unemployment, occupational accident or disease, etc., within the framework of the mandatory social security insurance. The general government is the payer and households represent the payee.

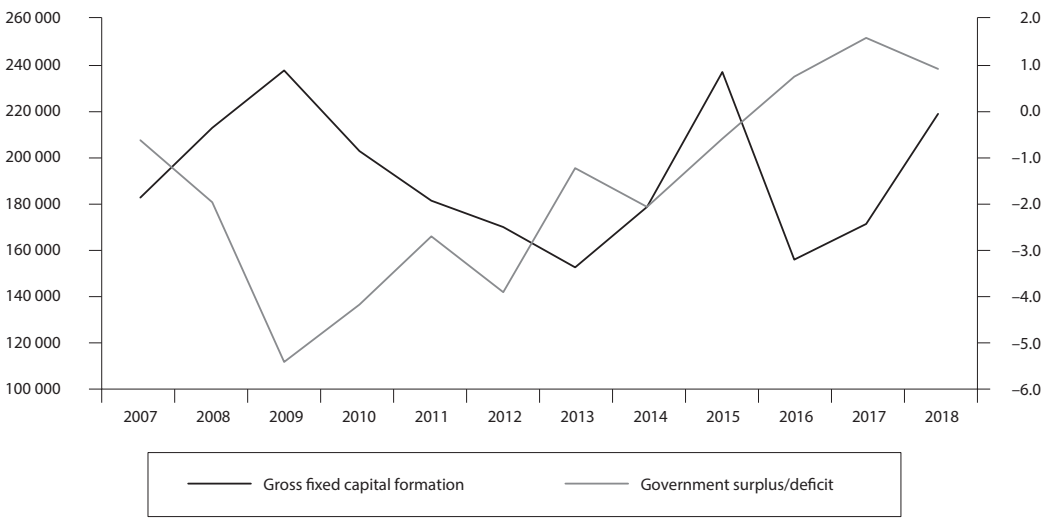
[20] For a better idea of the development of the volume of social benefits (payable), we also added the development of the volume of social contributions (receivable) to the chart.

The continuity of the social benefits' growth (as the most important component of the expenditure) is not compliant with the evolution of the government balance – cf. Figure 2. A similar result, showing a low level of mutual dependency, is obtained when comparing this balance with the compensation of employees (as the second most significant item of the government expenditure). An analogous conclusion is valid for the instance of the intermediate consumption, even though its evolution was not as smooth as that of social benefits and compensation of employees.

Investments into fixed capital represent a factor that significantly influences the government balance. Out of those, the largest proportion (three-fifths to three-quarters on a long-term average) goes to buildings and constructions, including the transportation ones. Figure 3 illustrates the sensitive response of the government deficit/surplus to the investments into the fixed capital.

The gross fixed capital formation does not cover a dominant part of the government expenditure (as compared with social benefits and compensation of employees); nevertheless, the influence of the investments on the government balance is obvious. This phenomenon is also implied by the fact that the social benefits and compensation of employees are mandatory expenses whose amounts are given by legal regulations and agreements. The investments into the fixed capital, i.e., the most significant part of the investment volume, can, to a certain extent, be controlled (boosted or inhibited) based on the expected evolution of the government revenue and expenditure.

Figure 3 Gross fixed capital formation by the general government (mil. CZK, current prices) and the government deficit/surplus, Czech Republic (in % of GDP)



Source: <www.czso.cz>

Figure 3 clearly implies that the growing gross fixed capital formation (GFCF) is reflected in a higher value of the government deficit. An exception is the year 2015 when the year-to-year growth of investments into the fixed capital was 32.8% while the general government's deficit went down by 1.5 p. p. (to a value of 0.6%). A reason for that extraordinary situation was a year-to-year growth of the government revenue by 8.5% (mainly due to a growing volume of the investment subsidies from the EU, revenues from taxes on products, and collected social contributions; the growth of those was implied by the growing wages). Despite the above-mentioned significant increase in the GFCF volume, the government expenditure

only grew by 4.7%, and the government deficit was decreased. In the year after that, on the contrary, the government revenue went up by a mere 1.5%. The investment subsidies (notably from the EU) went down (by 58.5 bil. CZK, i.e., by 72.1%) and, consequently, the volume of investments into the fixed capital was also significantly lower (with a year-to-year decrease by 81.1 bil. CZK, i.e., by 34.3%); the rate of investments of the general government thus went down from 38.5% to 24.4% (which has been the lowest level of the rate of investments of this sector since 1995). Such a large drop in the GFCF volume (despite a significant increase in the compensation of employees by 42.4 bil. CZK) meant a decrease in the government expenditure by 1.8%; in consequence, the government deficit of 0.6% in 2015 was turned to a surplus of 0.7% in 2016. The positive economic result was achieved by markedly attenuating the investments into the fixed capital; this arrangement should not be viewed as positive from the viewpoint of the economic policy. A low rate of investments in 2017 (25.0%) helped keep a positive government balance. On the contrary, the increased rate of investments in 2018 (29.2%) reduced the government surplus by nearly 40%, to 0.9% of GDP.

A certain exception from the GFCF evolution and its influence on the government deficit was the year 2012, in which the GFCF volume went down (by 6.5%) but the deficit was increased (from 2.7% to 3.9%). This increase of the government deficit was caused by the above-mentioned year-to-year growth of the capital transfers (payable) by 76.1 bil. CZK,[21] out of which the Church Restitutions amounted to 59.5 bil. CZK. Moreover, the Czech economy suffered another recession in 2012 (GDP went down by 0.8%, and GFCF by 3.1%).[22]

Summing up the Czech general government sector's situation after 2009, we can characterise the period in question as positive for the overall evolution of the revenue and expenditure because of the deficit and debt having been reduced (or the deficit even turning into surplus) Our analysis has shown that a factor strongly influencing the government balance is the volume of the investments into the fixed capital (in particular, buildings and constructions) and the latter's fluctuations are reflected in the changes of the government deficit/surplus with reciprocal proportion.

4 EXAMPLES OF OTHER EUROPEAN COUNTRIES[23]

We have chosen for our comparison those EU countries whose economic recession in 2009 (measured by the GDP growth rate) and the increased government deficit (as related to the GDP) were comparable with (or even higher than) those of the Czech Republic and in which the recovery after 2009 (similar to the Czech Republic) brought the government balance to a limit given by the convergence criterion. Each of the countries we have selected in this paper took its specific way to reducing its government deficit after 2009. In all instances, we will follow the concept of the national accounts and give our data in current prices. The average inflation rate values in the evaluated countries were not significantly different from each other in the period under assessment;[24] hence the evolution of the chosen absolute indices can be compared.

The first country we will focus on is **France**. Reasons for this selection are given not only by the economic development after 2009, characterised by reducing the government deficit every year, but also the abundant data available at the website of the Institut National de la Statistique et des Etudes Economiques (INSEE).

---

[21] When investment subsidies fall, especially from the EU.

[22] Both these values are given in the comparable prices.

[23] When selecting the countries for this analysis, the authors have been rather restricted by (non)availability of detailed data concerning the general government shown at the websites of the national statistical offices. Regarding the data of the national accounts, the Czech Statistical Office's database published at its website can undoubtedly be considered the best with respect to the presence of required details and user friendliness.

[24] The average annual inflation rate values in the period 2007–2018 was 1.9% in the Czech Republic, 1.4% in France, 1.9% in Belgium, and 1.6% in Slovakia.

Low but stable GDP growth rates are typical for the French economy (the average annual GDP growth rate has been 1.6% in the most recent 20 years). The year of crisis 2009 was the only one in this period in which the GDP went down (by 2.9%). In 2012 and 2013, when the Czech economy again slowed down to negative values of the GDP growth rate (–0.8% and –0.5%), the French economy also stagnated (with 0.3% and 0.6% year-to-year GDP growth rate). The recovery was slow in France; a value above 2% of the year-to-year GDP growth rate was achieved as late as in 2017 (and it went back to a 1.7% the year-to-ear growth value in 2018).

The French general government's proportion in the gross value added of the total economy is at about 18% on a long-term basis, as compared with 15% in the Czech Republic; this difference is implied by a wider redistribution role of the French state. In the latest decade, the government expenditure has been between 53 and 57% of GDP, and the revenue between 50 and 53% of GDP, out of which the mandatory payments (taxes and social contributions) amounted to a value between 44 and 48% of GDP.[25]

Since the early 1990s, the economic development in France has been accompanied by growing values of the government deficit and debt. The values of the government deficit were high in the early 1990s (with the maximum at 6.4% in 1993) to values below 3% (2.4% in 1998) in the effort to fulfil the convergence criteria when entering the EMU. The deficit went in 2017 (after years of recession) below 3% of GDP (and remained below this limit in 2018 as well). As early as 1996, the government debt first touched upon the limit of 60% of GDP (while it was a mere 36.1% of GDP in 1991) and has continuously been growing since that time (except for 2000 and 2001, when its value got slightly below 60% of GDP). The government debt in France has currently exceeded 98% of GDP.

Focusing on the period after 2009, we can clearly see that the government deficit in France was gradually going down and it got the level of the pre-crisis year 2007 in 2018. The slow rate in which the deficit proportion in the GDP was going down was caused by a relatively small lead of the government revenue growth (by 30.1%) before the expenditure growth (19.1%). The growth of the volume of the collected current taxes (by 51.6%) and taxes on production and imports (by 34.2%) were both growing faster than the revenue as a whole. Within expenditure, the fastest-growing items were those of the social benefits (by 23.4%) and subsidies (by 82.3%; but the subsidies only accounted for one-seventh of the social benefits' volume).

Table 3 Selected general government indices, France (in % of GDP)

Index	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Government surplus/deficit	–2.6	–3.3	–7.2	–6.9	–5.2	–5.0	–4.1	–3.9	–3.6	–3.5	–2.8	–2.5
Government revenue	49.9	50.0	50.0	50.0	51.1	52.1	53.1	53.3	53.2	53.1	53.6	53.5
Government expenditure	52.6	53.3	57.2	56.9	56.3	57.1	57.2	57.2	56.8	56.6	56.4	56.0
Government debt[26]	64.5	68.8	83.0	85.3	87.8	90.6	93.4	94.9	95.6	98.0	98.4	98.4

Source: <www.insee.fr>, the authors' own calculations

As already stated above, the situation of the French general government can be viewed both positively – because of the decreasing proportion of the deficit in the GDP, and negatively – because of the ever-growing

[25] In the Czech Republic, the proportion of the government expenditure was between 40% and 42% in the same period, and the proportion of the revenue oscillated around 40% of GDP; the mandatory payments' proportion was more or less stable at approx. 34% of GDP.

[26] This is consolidated gross debt for the purposes of the EDP (Excessive Deficit Procedure); for more details, cf. Hronová, Sixta, Fischer, Hindls (2019).

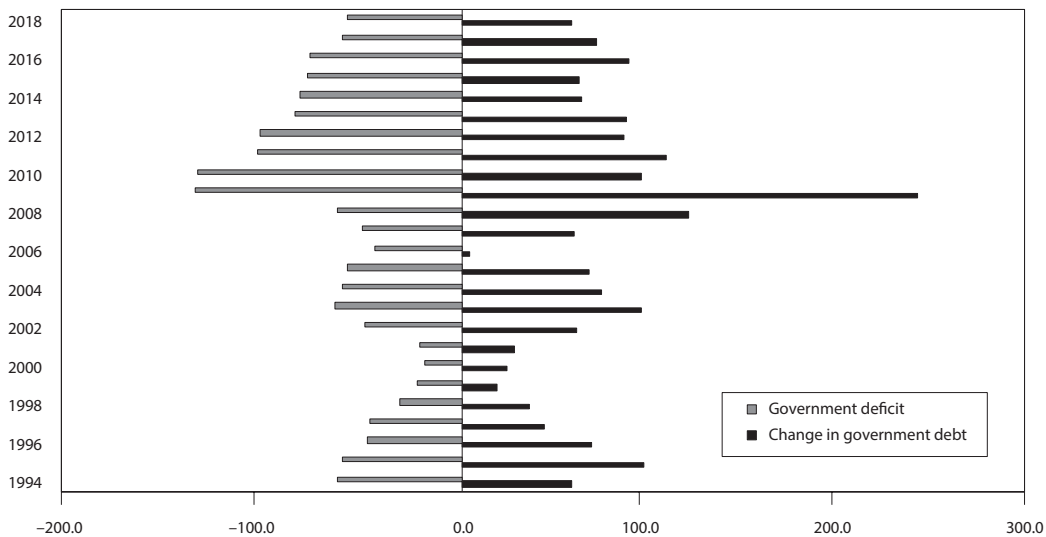
debt (with respect to the GDP). The explanation looks simple – the year-to-year deficit must be covered by the revenues from the newly issued bonds. The reality is not that simple, though. Let us recall a relationship prevailing between the debt increments and the amount of the deficit.

A change in the government debt equals the deficit/surplus only if no other changes occur implied by the government deficit and debt notification (adjustments, transactions affecting the government balance whose counter-items do not enter the government debt,[27] re-evaluating foreign currency liabilities, statistical differences, etc.), or in case of no changes in financial assets or liabilities that do not affect the government balance but affect the amount of the debt.[28]

Net changes in financial assets are among important causes for the existence of a difference between the government debt changes and government deficit/surplus. As a rule, such setup occurs when a general government issues bonds in a certain year but utilises the income from selling those bonds not in the same accounting period but in future years (i.e., it creates a financial reserve that is manifested as a growing value of the financial assets). If this is the case, the deficit may get reduced and the debt unchanged; or the available means may have been used to pay up the debt and then the debt is decreased while the deficit remains unchanged.

In France, the revenue is not sufficient to cover the expenditure from the viewpoint of the balance; this fact has led the French general government to seek new resources by issuing bonds. As Figure 4 shows, the annual increases in the debt value are "consumed" by payments on the deficit, and no money is left for paying

Figure 4 Changes in government debt and government deficit, France (bil. EUR)



Source: <www.insee.fr>, the authors' own calculations

[27] An example in the Czech Republic is represented by the above-mentioned Church Restitutions, included in the other capital transfers on the non-financial account. The corresponding counter-item on the financial account was the change in other liabilities, not entering the amount of the general government's debt.

[28] All such extraordinary operations expressing a difference between a change in the debt on the one hand and the deficit/surplus on the other hand are summed up to an adjustment item denoted by SFA (stock-flow adjustment) – cf. Eurostat (2019) for more details. Detailed data can also be found there concerning individual items entering the difference between the change in government debt and the government deficit/surplus in the entire EU.

up the government debt. This phenomenon leads to new issue of government bonds and the continuing growth of the government debt.

As of the end of 2007, the French general government's indebtedness in the form of bonds amounted to 1 019.9 bil. EUR, i.e., 52.5% of GDP); as of the end of 2018, this value was nearly doubled (to 1 993.0 bil. EUR, i.e., 86.8% of GDP);[29] the long-term bond indebtedness has been growing the fastest. The net increment of the indebtedness (change on the debit side minus the change on the credit side) of the French general government in the form of long-term bonds amounted to 76.6 bil. EUR in 2018. Similar amounts were valid in the years 2015 through 2017. The highest increments in the net indebtedness in the form of long-term bonds could be seen in the years of the crisis and shortly afterwards, i.e., in the years 2009 through 2011 (e.g., the bond-indebtedness was increased by 141.5 bil. EUR in 2009).[30] Since the value of assets (both financial and non-financial) only went up by one-fifth in the period under assessment, the total net worth of the French general government went down by 83.2%, amounting to a mere 1.3% of the national-economy net worth at the end of 2017.[31]

To sum up, the French government was successful in its effort to cover the high and ever-growing expenditure of the general government and, at the same time, to keep the deficit below the critical level of 3% of GDP, but only at the cost of a growing indebtedness in the form of bonds. Since the creditors are mainly foreign financial corporations,[32] the situation of the French general government, i.e., of the French public finances, is hardly sustainable on a long-term scale.

Another country that was hit by the 2009 recession is **Belgium**, with a drop in the GDP by 2.3% and a sudden surge of the government deficit (up to 5.4% of GDP, as compared with the 2007 surplus of 0.1%); its deficit is currently smaller than 1% of GDP.

The size of the Belgian economy is comparable to that of the Czech Republic; the former has been growing in the most recent 20 years at a relatively stable, but rather low rate (with the average GDP growth rate at 1.3% in the years 2000 through 2018). The Belgian economy only achieved the GDP growth rate values higher than 3% in 2000, 2004, and 2007; a year-to-year growth value of 2–3% only occurred in 2005, 2006, and 2010. The stability of the economic growth may have been one of the reasons why the 2009 crisis' impact on the GDP was relatively small in comparison with the other EU countries and the quick recovery as early as in 2010 (with the GDP growth at 2.7%, when the Eurozone average value was 2.1%).

The Belgian general government sector's proportion in the gross value added of the total economy was, on a long-term basis, at 15–16%; the government revenue has, in the most recent decade, amounted to values between 48% and 52%, and the government expenditure between 48% and 56%; the revenue grew by 36.9% and the expenditure by 24.9% from 2009 to 2018. A difference between the revenue and expenditure, i.e., the government deficit went down from a value of 5.4% of GDP in 2009 to 0.7% of GDP in 2018. From the viewpoint of the government deficit, the evolution of this value brought about gradual moderate improvements in the time period after 2009; but from the time of Belgium joining the European Monetary Union until 2007 the government deficit was undergoing significant year-to-year changes.

The most quickly growing item of the Belgian government revenue when comparing the years 2018 and 2009 was current taxes (by 47.8%), out of which the legal-entity income taxes' growth rate was equal

[29] We give here consolidated data to enable comparability with the data on the government debt shown in Table 3.

[30] The annual amount of the interest paid by the French general government has gradually been decreasing (from 57.3 bil. EUR in 2008 to 40.3 bil. EUR in 2018).

[31] Here we take a basis of the final annual balance sheet of 2017, i.e., non-consolidated data; the data from 2018 was not available at the time of writing this paper. In comparison, the net worth of the general government in the Czech Republic at the end of 2017 amounted to 41.6% of the corresponding national-economy value; that would be by 10 p. p. lower than in 2007.

[32] Foreign financial corporations are estimated to hold approx. 55–60% of the general government debt. Cf., e.g., *<https://www.lesechos.fr/2016/07/pourquoi-letat-ignore-qui-detient-sa-dette-215170>* or *<https://www.francetransactions.com/le-saviez-vous/surendettement-des-etats-qui-detient-la-dette-de-la-france.html>*.

to 139.6%. The volume of the collected taxes on production and imports grew at a rate identical with that of the total revenue. Among the expenditure items, the social benefits grew the fastest by 33.7%; out of these, the fastest were old-age pensions (by 44.9% – the pensions make up about two-fifths of the social benefits); there was a significant drop of 24.3% in the unemployment benefits (the unemployment rate went down by 2 p. p. in the same time period, but it still remains at a high level of 6%).

Table 4 Selected general government indices, Belgium (in % of GDP)

Index	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Government surplus/deficit	0.1	–1.1	–5.4	–4.0	–4.2	–4.2	–3.1	–3.1	–2.4	–4.4	–0.8	–0.7
Government revenue	48.3	49.2	48.8	49.3	50.3	51.6	52.7	52.2	51.3	50.7	51.4	51.7
Government expenditure	48.2	50.3	54.2	53.3	54.5	55.9	55.8	55.3	53.7	53.1	52.2	52.4
Government debt[33]	87.0	92.5	99.5	99.7	102.6	104.3	105.5	107.5	106.3	106.1	103.6	102.0

Source: <www.nbb.be>, the authors' own calculations

Another characteristic feature of the general government sector in Belgium is its high level of debt; it has been high since the creation of the EMU, when Belgium did not pass the government debt criterion – this debt was high above the critical 60% level (the Belgian government debt was 118.2% of GDP in 1998). All the same, Belgium became an EMU member state, but the country's effort to reduce its government debt was disrupted by the 2009 crisis. The lowest value of the Belgian government debt occurred in the pre-crisis year of 2007 (at 87.0% of GDP); the highest in 2014 (at 107.5% of GDP); now it is still higher than 100% of GDP (cf. Table 4); out of this value, 78.0% of GDP is the indebtedness in the form of long-term bonds. Belgium thus ranks with Greece, Italy, Cyprus, and Portugal among EU countries whose government debt is, on a long-term basis, higher than 100% of GDP.

The Belgian government debt in 2018 was by 32.4% higher than the 2009 value (and by 53.2% higher than the 2007 value); out of this value, 38.6% was the indebtedness in the form of long-term bonds. The general government's proportion in the overall financial liabilities of the national economy was 11.6% in 2018 (i.e., by 1.6 p. p. more than in 2009).[34] When evaluating the economic development of the Belgian general government, the deficit was being reduced after 2009 by faster growth in revenue than in expenditure. The general government in Belgium, similar to France, looks for the resources to cover the expenses incurred on issue of bonds, in particular, long-term ones. Figure 5 implies that the income from the debt increase after 2009 was mainly utilised on covering the deficit and the debt itself was being reduced only gradually.

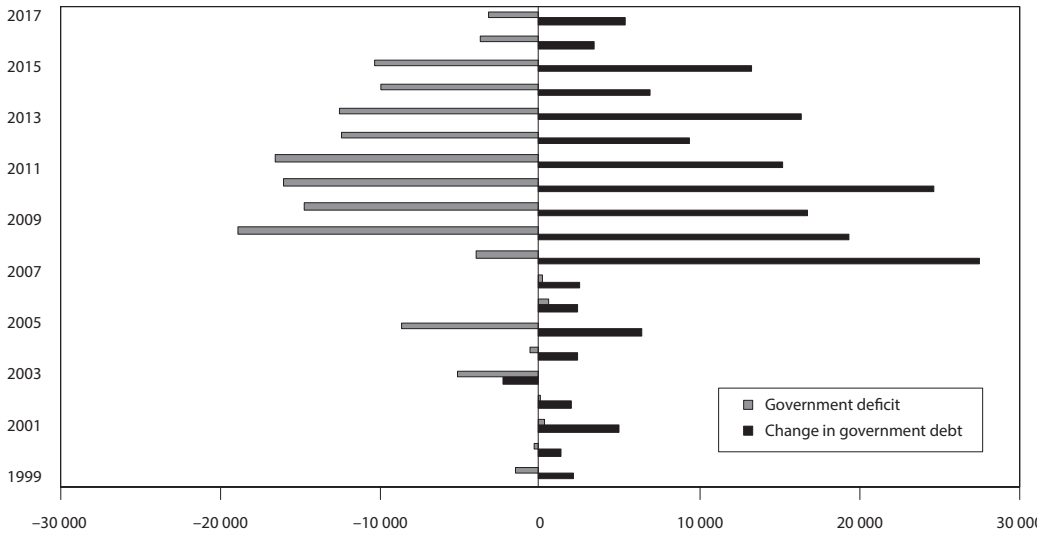
However, we should view on the low changes of the Belgian government debt before 2009 keeping in mind the amount of that debt (whether absolute or relative with respect to the GDP) – it exceeded the critical level of 60% of GDP by tens of percentage points.

The last country we have included in our comparative analysis is **Slovakia**. This choice is based not only on the data availability at the website of the Slovak Statistical Office and the National Bank of Slovakia but mainly because the drop in the Slovak economy after 2009, measured by the GDP growth rate (–5.4%),

[33] This is consolidated gross debt for the purposes of the EDP (Excessive Deficit Procedure); for more details, cf. Hronová, Sixta, Fischer, Hindls (2019).

[34] In 2018, this proportion was 10.3% in France and 9.2% in the Czech Republic.

Figure 5 Changes in government debt and government deficit, Belgium (mil. EUR)



Source: <www.stat.nbb.be>, the authors' own calculations

was one of the highest amount the EU countries;[35] and the Slovak government deficit in that year also ranked among the highest values in EU (7.8% of GDP).[36]

The Slovak economy has been among countries with the highest year-to-year growth rate values; the average GDP growth rate in the period 2002–2018 was 3.4%, which is the highest among the countries we analyse in the present paper.[37] Slovakia was also able to quickly recover its economy after 2009; in 2010, its year-to-year GDP growth was 5.0% – apart from Sweden, this was the largest such value in the entire EU. The Czech economy got to negative growth rates in 2012 and 2013; in the same period the Slovak economy slowed down to 1.7% and 1.5% year-to-year GDP growth rates; and it has been achieving growth rates of more than 3% of GDP since 2014.

The Slovak general government's proportion in the gross value added of the total economy is, on a long-term basis, between 12% and 14% and the government revenue have, in the most recent decade, fluctuated within a rather wide range between 34% and 43% of GDP; the expenditure between 36% and 45% of GDP (cf. Table 5). The government revenue growth (higher by 55.0% as compared with 2009) has been faster than the expenditure growth by 25.2 p.p. – this difference is the highest among the countries we analyse in the present paper. It is logical that, under such circumstances, the government deficit was going down, getting below the critical limit of 3% as early as in 2013. The high rates of the economic growth led to a growing volume of collected current taxes (higher by 76.0%), social contributions (by 66.1%), and taxes on production and imports (by 47.8%). On the other hand, the collected property income and miscellaneous current transfers significantly went down (both by 24.0%); these items only make up less than 5% of the government revenue.

[35] The largest economic drop in 2009 occurred in the Baltic states (nearly 15%); it was between 6% and 8% in Croatia, Hungary, Finland and Iceland.

[36] The highest value of the government deficit with respect to the GDP in 2009 occurred in Greece, Ireland, Spain and the United Kingdom (above 10%); its value was around 9.5% in Lithuania, Latvia, Portugal and Romania. Slovakia, France and Poland had that value between 7% and 8%.

[37] The average GDP growth rate in the period 2000–2018 was 2.5 % in the Czech Republic, 1.6% in France, and 1.3% in Belgium.

Table 5 Selected general government indices, Slovakia (in % of GDP)

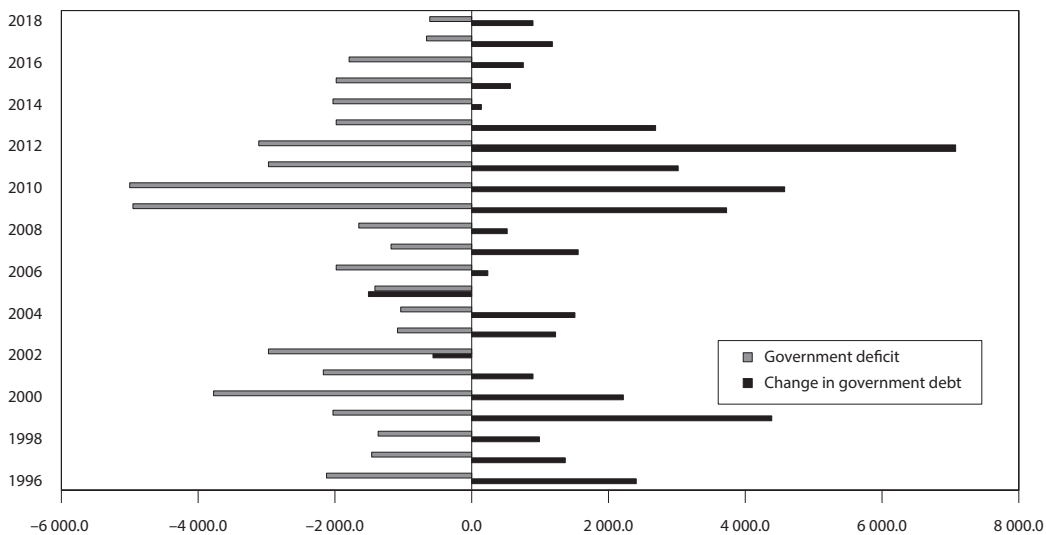
Index	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Government surplus/deficit	–1.9	–2.4	–7.8	–7.5	–4.3	–4.3	–2.7	–2.7	–2.6	–2.2	–0.8	–0.7
Government revenue	34.4	34.5	36.3	34.7	36.5	36.3	38.7	39.3	42.5	39.2	39.4	39.9
Government expenditure	36.3	36.9	44.1	42.1	40.8	40.6	41.4	42.0	45.1	41.5	40.2	40.6
Government debt[38]	30.1	28.5	36.3	41.2	43.7	52.2	54.7	53.5	52.2	51.8	50.9	48.9

Source: <www.nbs.sk>, the authors' own calculations

Regarding the Slovak government expenditure, which has grown by 29.8% as compared with 2009, the highest growth occurred in the compensation of employees (by 53.2%) and social benefits (by 31.0%); the investments into the fixed capital went up as quickly as the total expenditure, and – unlike in the Czech Republic – the year-to-year fluctuations in their volume were not in reciprocal proportion to the changes in the government deficit.

Having in mind the large difference between the Slovak government revenue growth and expenditure growth and the decreasing deficit, it is surprising that the debt was growing as far as 2013 and that its high values still prevail (cf. Table 5).

Figure 6 Changes in government debt and government deficit, Slovakia (mil. EUR)



Source: <www.nbs.sk>, the authors' own calculations

[38] This is consolidated gross debt for the purposes of the EDP (Excessive Deficit Procedure); for more details, cf. Hronová, Sixta, Fischer, Hindls (2019).

[39] Activation of the so-called debt brake pursuant to the constitutional act on budget responsibility means taking a number of steps aimed at the stabilisation of the public finances. If a pre-set limit is exceeded by the debt, the government will have to announce austerity measures and put forth a proposal of how the situation should be resolved at the levels of both the central and the local government authorities.

The Slovak general government tried to resolve a sudden increase in its deficit in 2009 (to 7.8% of GDP); it was aiming at decreasing the deficit in the subsequent years by issuing long-term bonds. This approach was, of course, reflected in the government debt growing every year; the most significant year-to-year change occurred in 2012 (by 22.9%, or by 8.5 p.p. as related to the GDP). The government debt exceeded a level of 50% of GDP, and Slovakia put on the debt brake.[39] The debt increment in 2012 was larger than the amount of the deficit and made it possible for the obtained financial means to be used for covering the expenditure in the subsequent years and to pay up the debt. This way, the deficit was gradually reduced (absolutely and with respect to the GDP); and the debt's proportion in the GDP was also reduced. Figure 6 best illustrates the relationship between the changes in the government debt and the government deficit in Slovakia.

Even though the evolution of the Slovak government debt and deficit after 2009 may be viewed as positive, the debt amount still remains high and only in 2018 got below the sustainability limit.[40]

CONCLUSIONS

The formal fiscal rules setting out the critical levels for the government deficit and debt to a certain extent regulate the general government's behaviour in the respective country; nonetheless, infringements on such rules (especially the long-term exceeding of the government debt value in certain Eurozone countries) are more or less tolerated (e.g., in Belgium and France), unless such infringements are accompanied by additional significantly negative phenomena (such as in Greece).

In 2009, a drop in economic activities occurred in all European countries (except for Poland) – the GDP went down on the EU-average by 4.3%. The general government in each country was hit by the drop in GDP, increased unemployment and other symptoms of the economic crisis. The subsequent drop in revenue from taxes and social contributions, as well as the increased expenditure incurred on social benefits, were manifested in a significant increase of the government deficits (6.6% of GDP on the EU-28-average). A solution was mainly seen in stimulating the economic activities – it may have been supported by the general governments' interventions (which increased their expenditure). An alternative was to cut down the expenditure; this approach was applied in a number of countries and attenuated their economic activities and brought back the crisis in 2012 and 2013 (this development occurred not only in the Czech Republic but also, e.g., in Hungary, Italy, the Netherlands, Spain, and other countries).

Despite that and the "wavering" in 2012 and 2013, all EU countries (except for Spain) achieved values of their government deficit below the critical 3% limit as early as in 2016; and there is another exception of Cyprus in 2018. Each country chose its own specific way to get rid of the crisis and to reduce the too-high deficit prevailing in 2009. The goal of the present paper is to point out the general government's economic behaviour in the Czech Republic and compare it with those of several selected countries – France, Belgium, and Slovakia. When selecting those countries, we looked for meeting a criterion of a high government deficit in 2009 and its reduction below the critical 3% limit by 2018. Unfortunately, we were restricted in our choice by the fact that, in many countries, detailed data of the national accounts are not available at the websites of the respective national statistical offices.

The Czech general government was undergoing a difficult stage of its development after 2009. The high deficit of 2009 (at 5.5%, and with the GDP decreased by 4.8%) had to be covered by a growing indebtedness. Consequently, the government debt underwent significant changes from 2009 to 2012; the gain generated by the issued bonds did not cover the deficit in the first two years. The effort aimed at reducing both the deficit and the debt led to a growth in the expenditure slower than that in the revenue,

[40] It is a pre-set fiscal limit – the maximum level of the debt considered sustainable from the viewpoint of the general government. The Slovak general government's goal is to reduce the debt below a limit of 40% of GDP, and subsequently to put on the debt brake when the debt exceeds 40% of GDP. Cf. the Constitutional Act on Budget Responsibility, Article 13 (Act No. 493/2011 Coll.).

and finally the government deficit was turned into a surplus in 2016 (as well as in 2017 and 2018). Our analysis has, however, shown that the changes in the Czech government deficit respond very sensitively (in addition to the extraordinary circumstances such as the Church Restitutions) to changes in the volume of the gross fixed capital formation. The latter's large decrease by 34.3% in 2016 significantly contributed to the government surplus at 0.7% of GDP; an increase in investments into the fixed capital by 27.7% in 2018 led to a decrease in the surplus by 0.7 p.p. It is not sustainable, on a long-term basis, to reduce the government deficit and, at the same time, to suppress investments into the fixed capital (if the latter were a rule); nevertheless, the structure of assets and liabilities, as well as the scope of the revenue and the expenditure, and their time evolution set up (at least currently) a prerequisite for a favourable development of the Czech general government's economic result.

There are certain common features characterising the evolution after 2009 in France, Belgium and Slovakia. They include a high value of the government debt (more than 100% of GDP in Belgium, nearly that much in France, and at the fiscal limit of 50% of GDP in Slovakia). All the countries we study in the present paper have been trying to reduce the government deficit by issuing bonds, but with different results in each of these countries.

The French general government has been struggling with high values of debt and deficit on a long-term basis; in 2017, the deficit got below a level of 3% of GDP, but only at the cost of increasing the debt to nearly 100% of GDP. The high liabilities of the general government led to a continuing decrease in its net worth as far as 1.3% of the national-economy value.

The Belgian general government had low deficit/surplus values until 2008, with small changes in a very high debt (of more than 100% of GDP); after 2009, it tried to alleviate the impact of the crisis by stimulating a faster growth in the revenue than in the expenditure, and, in particular from 2009 to 2012, by issuing bonds. The changes in government debt were higher than the deficit value, which was going down, the reduction of the debt went rather slowly; its value in 2018 remained higher than 100% of GDP, and the deficit was at 0.7% of GDP.

The Slovak general government took a way of substantially advancing its revenue growth over its expenditure (by 25.2 p. p.). Despite the increasing volume of collected direct and indirect taxes and social contributions based on the positive growth in the economy as a whole, the Slovak general government had to address the problem of high deficit values by emitting bonds from 2009 to 2012. This approach suddenly increased the government debt above 50% of GDP in 2012, and the debt remained higher than this fiscal limit until 2017.

If we sum up the evolution of the Czech general government's economic results and compare it with the circumstances in France, Belgium, and Slovakia, the Czech evolution seems to be sustainable (except for the large year-to-year changes in the gross fixed capital formation) because the government deficit reduction has been accompanied by a decrease in the government debt that is more substantial than in the other countries we study in the present paper. In this article we analyzed the long-term sustainability of public finances only based of national accounts data, ie only based of historical data. Another important issue concerning the long-term sustainability of public finances, which we have not examined here, is the aging population. However, this is a very complex problem requiring separate analyzes and other data than could be obtained only from national accounts.

ACKNOWLEDGMENT

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

References

ASPDEN, C. The Revision of the System of National Accounts. What does it change? *Statistics Brief*, No. 13, Paris: OECD, 2007.

ARVAY, J. The Material Product System (MPS): A Retrospective. *Twenty-second General Conference of the International Association for Research of Income and Wealth* (IARIW), Switzerland, 1992.

CARSON, C., KHAWAJA, S., MORRISON, T. K.: *Revisions Policy for Official Statistics: A Matter of Governance.* Washington: IMF, 2004.

Comptes de la Nation en 2018. Insee Première, No. 1754, INSEE, 2019.

Comptes des administrations publiques en 2018. Insee Première, No. 1753, INSEE, 2019.

CONNOLLY, M. The Challenges Facing Globalization. *Actes du 16ème colloque de Comptabilité Nationale*, Paris, 2017.

European System of Accounts – ESA 2010 (Système Européen des Comptes – SEC 2010). Luxemburg: Eurostat, 2013.

HARVEY, R. Comparaison des taux d´épargne des ménages. Zone EUR/Etas-Unis/Japon. Cahiers statistiques, No. 8, OECD, 2005.

HRONOVÁ, S. AND HINDLS, R. Czech Households in the Years of Crises [online]. *Statistika: Statistics and Economy Journal*, 2013, Vol. 93, No. 4, pp. 4–23.

HRONOVÁ, S., HINDLS, R., MAREK, L. Reflection of the Economic Crisis in the Consumer and Entrepreneurs Subsectors [online]. *Statistika: Statistics and Economy Journal*, 2016, Vol. 96, No. 3, pp. 4–21.

HRONOVÁ, S., SIXTA, J., FISCHER, J., HINDLS, R. *Národní účetnictví – od výroby k bohatství* [National Accounts – From Production to Wealth]. Prague: C. H. Beck, 2019.

Insee Première. No. 1753, INSEE, 2019.

LEQUILLER, F. AND BLADES, D. *Comptabilité nationale.* Paris: Economica, 2004.

LEQUILLER, F. Quelques propositions d'amélioration technique sur le compte des administrations publiques dans le SCN. *Actes du 16ème colloque de Comptabilité Nationale*, Paris, 2017.

Manual on Government Deficit and Debt. Implementation of ESA2010. 2014 Ed. Eurostat, 2014.

MAREK, L., HRONOVÁ, S., HINDLS, R. Structure of Final Consumption in Light of GDP Dynamics [online]. S*tatistika: Statistics and Economy Journal*, 2017, Vol. 97, No. 3, pp. 5–15.

Monitoring economic performance, quality of life and sustainability. German Council of Economic Experts, Wiesbaden and Conseil d´Analyse économique, Paris, 2010.

SINGER, M. A Comparison of the Rates of Growth of Post-Transformation Economies: What Can (Not) Be Expected From GDP? *Prague Economic Papers*, 2013, Vol. 12, No. 1, pp. 3–27.

SKALÁK, Z. AND RYBÁČEK, V. Pension Liabilities in the Czech Republic [online]. *Statistika: Statistics and Economy Journal*, 2018, Vol. 98, No. 3, pp. 209–222.

VEBROVÁ, L. AND RYBÁČEK, V. Public Finance, the Public Sector and the General Government Sector [online]. *Statistika: Statistics and Economy Journa*l, 2018, Vol. 98, No. 4, pp. 362–368.

Stock flow adjustment for the Member States, the EUR area (EA-19) and the EU-28, for the period 2015–2018 as reported in the April 2019 EDP notification. Eurostat, 2019.

Determinants of Firms´ Innovation Activities in V4 Countries

Petra Cisková[1] | *Matej Bel University, Banská Bystrica, Slovakia*
Ina Ďurčeková[2] | *Matej Bel University, Banská Bystrica, Slovakia*

Abstract

Innovation is considered to be the driving force of competitiveness and growth of firms as well as countries. However, despite these benefits of innovation, not all firms undertake innovation projects. There are several barriers and factors determining the involvement of firms in innovation activities. The aim of the paper is to examine determinants affecting involvement of firms in innovation activities in V4 countries. The emphasis is put on issues that present the most pronounced barriers to commercialization of innovation. The analysis is based on data obtained from the Innobarometer 2016 survey. The paper is focused on examination of several determining factors that are studied for a variety of firms. These factors are represented mainly by type of innovation or innovation barriers and their impact on involvement of firms in innovation activities. The analysis is based on several probit models of micro-level data. It seems that R&D, turnover and innovation investments are among the main determinants of innovation activities of firms in V4 countries. We have also found that in V4 countries, product innovation was introduced mostly by smaller firms while larger firms tend to focus on process innovation. The main major barriers of innovation encountered by firms seem to be the lack of human resources and the fact that the market is dominated by competition.[3]

Keywords

Innovation activities, determinants, innovation barriers, Innobarometer 2016

JEL code

C32, O31, O32

INTRODUCTION

Innovation is a key factor affecting competitiveness and growth of firms. Firms therefore put emphasis on introducing new innovation to support their growth and reinforce their position on the market. Innovation can be defined as application of new or improved ideas, products, services or processes that bring increased utility or quality (Mataradzija et al., 2013). The importance of innovation in business environment is constantly increasing. This is also confirmed by the fact that business innovation activities not only lead

[1] Matej Bel University, Faculty of Economics, Department of Quantitative Methods and Information Systems, Tajovského 10, 975 90 Banská Bystrica, Slovakia. Corresponding author: e-mail: petra.ciskova@umb.sk, phone: (+421)484466617.
[2] Matej Bel University, Faculty of Economics, Department of Finance and Accounting, Tajovského 10, 975 90 Banská Bystrica, Slovakia. E-mail: ina.durcekova@umb.sk, phone: (+421)484466314.
[3] This article is based on contribution at the *International Scientific Conference for Doctoral Students and Young Scientists – Scientia Iuventa* in April 2019, Banská Bystrica, Slovakia.

to the generation of knowledge, which may manifest itself in new products and improved production methods used in the production process, but they also lead to higher productivity (Zemplinerová and Hromádková, 2012; Polder et al., 2010; Hashi and Stojcic, 2010; Mairessee and Robin, 2009; Van Leeuwen and Klomp, 2006; Lööf and Heshmati, 2002; Crépon, Duguet, Mairessee, 1998).

According to related literature, there are four main types of innovation activities: product, process, marketing and organizational innovation (OECD, 2005). This is in line with the types of innovation examined in Innobarometer 2016 survey, which defines five types of innovation, since it distinguishes two types of product innovation – significantly improved goods and significantly improved services – in addition to other three aforementioned types.

Even though the introduction of various types of innovation depends on different determinants, there are several factors affecting whether a firm introduces an innovation in general. These factors can generally be divided into three categories: macroenvironmental factors (such as political, economic or social factors), microenvironmental factors (such as suppliers, consumers or competitors) and internal factors (such as production, finance or personal of a firm) (Yachmeneva and Vol's'ka, 2014). These factors can also be divided into internal and external ones. Internal factors reflect various characteristics of a firm, such as its size or age, or decisions made by a firm. External factors describe the environment surrounding a firm, such as customs (EBRD, 2014). This paper is mostly focused on examination of internal factors influencing innovation activities of a firm, such as demographic factors. External factors affecting involvement of firms in innovation activities researched within the paper are mostly focused on problems firms consider to be barriers to introducing innovation.

There have been many studies focused on examining determinants of firm innovation in various countries and regions. Review of the main findings is summarized in Table 1.

Table 1 Determinants of firm innovation – review of the main findings

Determinant	Authors
Past innovation activities	Baldwin, Gu, 2004; Vega-Jurado et al., 2008; Romjin, Albaladejo, 2002
Technological competencies	Baldwin, Gu, 2004; Vega-Jurado et al., 2008
Intensity of R&D	Raymond, St-Pierre, 2010; Baldwin, Gu, 2004
Size of a firm	Fritz, 1989; Baldwin, Gu, 2004; de Jong, Vermeulen, 2004; Rosa, 2002; Oum, Narjoko, Harvie, 2014
Age of a firm	de Jong, Vermeulen, 2004
Foreign control of a company	Fritz, 1989; Guadalupe, Kuzmina, Thomas, 2012; Baldwin, Gu, 2004
Sector	Vega-Jurado et al., 2008; de Jong, Vermeulen, 2004; Rosa, 2002
Type of innovation	Raymond, St-Pierre, 2010; Fritz, 1989; Rosa, 2002
Access to finance	Oum, Narjoko, Harvie, 2014
Human capital	Oum, Narjoko, Harvie, 2014

Source: Authors

Some of the main findings related to the paper are as follows: Baldwin and Gu (2004) found that large firms have higher rates of process innovation than smaller ones and that foreign-controlled firms have

higher innovation rates than domestic ones. These conclusions are partially in line with results of Fritz (1989) who found that smaller, owner-run firms facing less competitive pressure have higher rate of product innovation. Even though the focus of research is often on large enterprises and the innovation they create, many authors emphasize the importance of micro-enterprises and SMEs in the area of innovation. SMEs are often seen as a valuable source of innovation, since their flexibility and simpler organization structure allows them to overcome innovation barriers easier than it is in larger enterprises (Czarniewski, 2016; Stephens, 2016; Lesáková et al., 2010).

The paper is structured as follows: Section 1 provides the details of methodology and describes the data and the details of the probit models. In Section 2 the introduced probit models are applied to data of V4 countries and the results are discussed. Main findings of the paper are summarized in the part "Conclusion".

1 METHODOLOGY

In this paper we use data from Flash Eurobarometer 433 – Innobarometer 2016 – EU Business Innovation Trends survey[4] that was held between February 1st and February 19th of 2016. Innobarometer survey gathers a firm-level data from 28 Member States of European Union, Switzerland and United States concerning information about innovation, design, plans for future investments in innovation and the problems encountered with introducing a new – innovative or non-innovative – goods and services into the market. The methodology of the Eurobarometer was used in the survey and the interviews were conducted with the key decision makers of companies. Innobarometer survey data is used within analyses published by many authors (e.g. Tether, 2005; Lorenz, 2011; Filippetti and Archibugi, 2011; Trigo, 2013; Montresor and Vezani, 2016; Božić and Botric, 2017; Guerzoni, 2014). The data used in the paper was obtained from GESIS based on the instructions from the official website of the European Union in addition to official aggregated data published by the European Union. However, some inconsistencies can be seen between data provided by GESIS and aggregated data published by the European Union, probably due to methodology used to summarize the findings. In our paper, we follow the data provided by GESIS.

We focus on analysis of the Visegrad Group countries consisting of the Czech Republic, Hungary, Poland and Slovakia (hereinafter referred to as "V4"). The countries are selected based on their similar levels of innovation performance according to Summary Innovation Index which stems from their similar geographic and economic positions. Since the survey questions are changed between years, we decided to focus on comparison of selected countries within one year (2016). The data is analyzed using descriptive statistics and the probit models.

The survey covers a wide range of questions related to innovation. Respondents are asked several questions concerning the firms´ innovation activities, whether the company introduced any new or significantly improved goods, services, or processes. In addition, respondents provide various types of demographic information (such as size of a firm, sector in which the firm operates, year of establishment) and information directly connected to innovation activities of a firm (such as type of innovation or innovation barriers).

In V4 countries, the questionnaire of Innobarometer survey was answered by 500 firms from the Czech Republic, 500 firms from Slovakia, 500 firms from Hungary and 501 firms from Poland, which gives us a total number of 2 001 observations. For the purposes of the paper, innovative firm is defined based on the Innobarometer survey question Q2 regarding introduction of types of innovation by a firm. Innovative firm in the paper is defined as a firm that undertook any type of innovation. Innovation activities are crucial to increase market share and competitiveness of a firm which is shown by the fact that approximately 68% of the surveyed firms in V4 countries were involved in innovation activities

[4] Available at: <*https://zacat.gesis.org*>.

in the year under review. Table 2 shows that out of 2 001 firms, 1 359 were involved in any type of innovation activity. The highest number of innovative firms was in Poland, closely followed by the Czech Republic and Slovakia, while the least innovative firms were in Hungary. Overall, more than half of the surveyed firms undertook some type of innovation in all countries. It can also be seen that the structure of the innovative firms consisted of over 43% of microenterprises, 29% of SMEs and 27% of large firms. We assume that the high representation of microenterprises within innovative firms stems from the fact that many start-ups, which are mostly innovative firms, belong to the group of microenterprises.

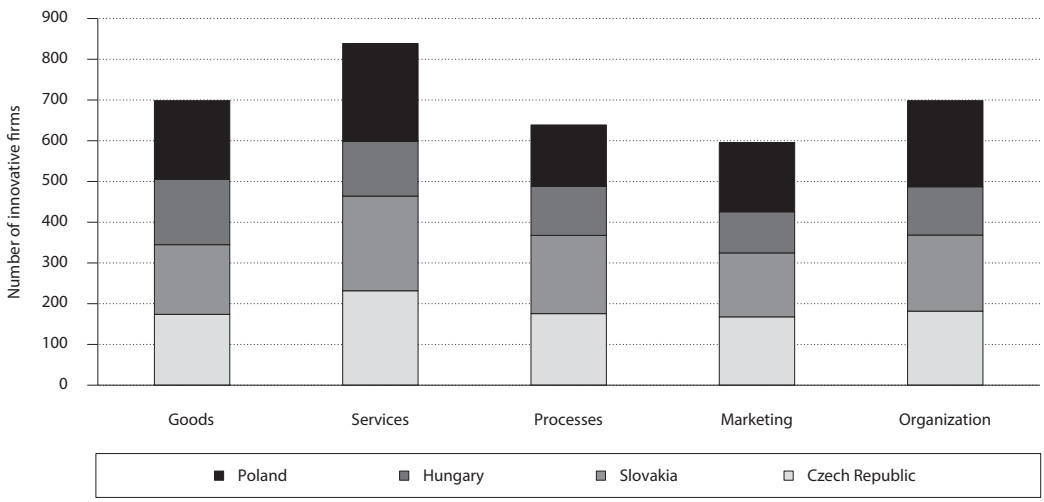
Table 2 Number of innovative and non-innovative firms based on their size in V4 countries

Country	Innovative firms			Non-innovative firms
	Microfirms	SMEs	Large firms	
Czech Republic	139	117	108	136
Slovakia	155	114	88	143
Hungary	112	82	76	229
Poland	185	85	97	134
Total	**591**	**399**	**369**	**642**

Notes: Microfirms: 1–9 employees; SMEs: 10–49 employees; large firms: 50 and more employees.
Source: Own calculations based on Innobarometer 2016

However, it is not only important to look at the aggregate number of how many firms were involved in innovation activities, but to also examine the types of innovation they introduced. Overall, the surveyed firms were mostly focused on innovating their services. The only country where service innovations were

Figure 1 V4 innovative firms according to type of innovation they introduced

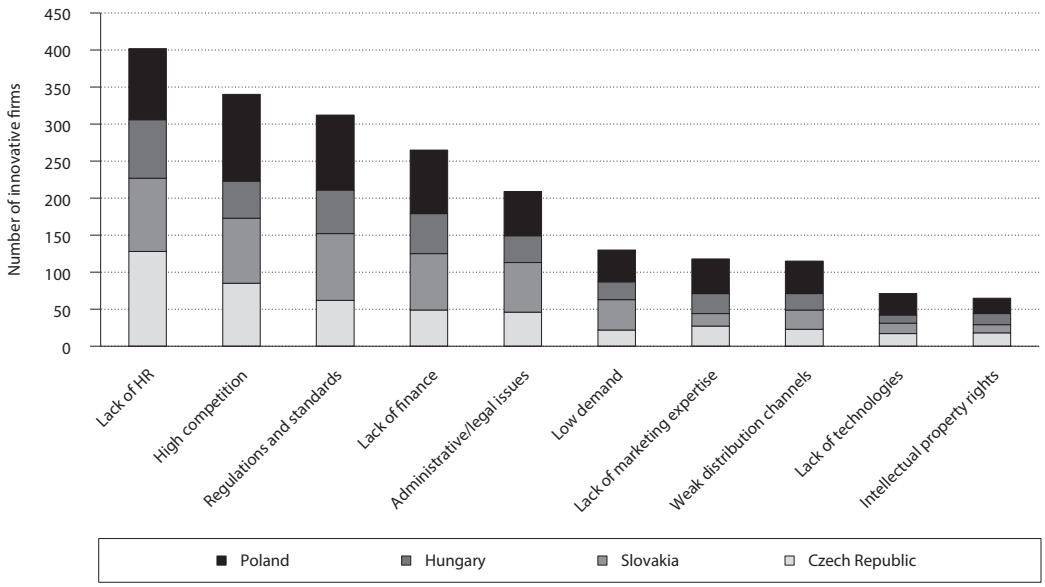


Source: Own calculations based on Innobarometer 2016

not dominant was Hungary, where firms were mostly involved in innovations focused on improving their goods. On the other hand, the V4 firms were least interested in launching marketing innovation.

It it obvious that innovation is an important element of increasing competitiveness of a firm and therefore is very beneficial to a firm. However, despite these benefits, not all firms introduce new innovations. That raises the question of why do not all firms undertake innovation activities. The answer is that firms encounter several difficulties while launching innovation projects. The most pronounced barriers to innovation according to surveyed firms are summarized in Figure 2.

Figure 2 Major barriers to innovation according to innovative firms in V4 countries



Source: Own calculations according to data from Innobarometer 2016

It is apparent that the main major barrier to innovation according to innovative firms in V4 countries is the lack of human resources, closely followed by the fact that the market is dominated by competition. Standards and regulations as well as lack of financial resources were also found to be problematic by innovative firms.

1.1. Descriptive analysis

Our strategy to choose variables is based on similar studies examining the determinants of firm innovation. In addition, we have been strongly influenced by the findings in Capozza and Divella (2017), Rehman (2016) and Montresor and Vezzani (2016).

We employ three probit models where the dependent variables are one of the three following types of innovation that company introduced since January 2013:

- product innovation (y_1),
- service innovation (y_2),
- process innovation (y_3).

Then, we use wide range of independent variables, that were divided into three groups according to their similar characteristics:

- demographic variables,
- variables of innovation impact,
- barrier variables.

The characteristics of a list of variables are described in Table 3.

Table 3 Variable description

Dependent variables	
Product innovation (y_1)	1 if company introduced a new product since January 2013; 0 otherwise
Service innovation (y_2)	1 if company introduced a new service since January 2013; 0 otherwise
Process innovation (y_3)	1 if company introduced a new process since January 2013; 0 otherwise
Independent variables	
Demographic variables	
Firm´s size (x_1)	1 if number of employees are between 1 to 9; 2 for companies with 10 to 49 employees and 3 for companies with more than 50 employees
Young (x_2)	1 if company was established after 1 January 2010; 0 otherwise
Group (x_3)	1 if company belongs to a business group; 0 otherwise
Turnover (x_4)	–1 if company´s turnover has decreased since January 2013; 0 if turnover remained approximately the same; 1 if turnover has increased
Variables of innovation impact	
Innovative products and services (x_5)	1 if 0% of company's turnover was due to innovative goods or services that have been introduced since January 2013; 2 if the percentage of turnover was between 1 and 5%; 3 if the percentage of turnover was between 6 and 10%; 4 if the percentage of turnover was between 11 and 25%; 5 if the percentage of turnover was between 26 and 50%; 6 if the percentage of turnover was 51% and more
Investing in innovation (x_6)	1 if company has not invested in innovation activities; 2 if company has invested in innovation activities less than 1% of turnover in 2015; 3 if company has invested between 1 and 5%; 4 if company has invested between 6 and 10%; 5 if company invested more than 11%
R&D (x_7)	1 if company has not invested in research and development since January 2013; 2 if company invested in R&D less than 1% from turnover; 3 if company has invested between 1 and 5% of turnover; 4 if company has invested more than 5% of turnover
Training (x_8)	1 if company has not invested in training since January 2013; 2 if company invested in training less than 1% from turnover; 3 if company has invested between 1 and 5% of turnover; 4 if company has invested more than 5% of turnover
Organization investments (x_9)	1 if company has not invested in organization or business process improvements since January 2013; 2 if company invested less than 1% from turnover; 3 if company has invested between 1 and 5% of turnover; 4 if company has invested more than 5% of turnover
Acquisition of assets (x_{10})	1 if company has not invested in acquisition of machines, equipment, software or licenses since January 2013; 2 if company invested less than 1% from turnover; 3 if company has invested between 1 and 5% of turnover; 4 if company has invested more than 5% of turnover
Marketing innovation (x_{11})	1 if company introduced a new marketing strategy since January 2013; 0 otherwise
Organization innovation (x_{12})	1 if company introduced a new organizational method since January 2013; 0 otherwise
Barrier variables	
Lack of HR (x_{13})	1 if company considers the lack of human resources as a major problem in the commercialization of company´s innovative goods and services; 0 otherwise
Regulations and standards (x_{14})	1 if company considers the cost or complexity of meeting regulations or standards as a major problem in the commercialization of company´s innovative goods and services; 0 otherwise
Competitors (x_{15})	1 if company considers the market dominated by established competitors as a major problem in the commercialization of company´s innovative goods and services; 0 otherwise

Source: Authors

Innobarometer survey is a structured type of questionnaire, where the respondents select (mostly) one-choice or multiple-choice answers. If some questions are linked to previous question and the answers are not applicable, or if respondents chose not to answer, we decided to exclude these observations from our sample. Choices are often offered as intervals, with different widths of scale (respondents are subsequently divided into several categories, e.g. according to their R&D investments, with the R&D investment being 0%, lower than 1%, lower than 5% or higher than 5% of the turnover, etc.). Considering

Table 4 Description of NACE classification and corresponding categories

NACE classification	Categories
Manufacturing	C – manufacturing
Industry	D – electricity, gas, steam and air conditioning supply E – water supply, sewerage, waste management and remediation F – construction
Retail	G – wholesale and retail trade, repair of motor vehicles and motorcycles
Service	H – transportation and storage I – accommodation and food service activities J – information and communication K – financial and insurance activities L – real estate activities M – professional, scientific and technical activities N – administrative and support service activities R – arts, entertainment and recreation

Source: Authors based on Innobarometer 2016

Table 5 Number of innovative firms based on NACE classification and firm size (in regards to the data used in probit analysis)

NACE classification	Microfirms	SMEs	Large firms	Total
Manufacturing	28	47	96	171
Industry	59	43	76	135
Retail	157	119	45	321
Service	128	105	82	315
Total	**372**	**314**	**256**	**942**

Notes: Microfirms: 1–9 employees; SMEs: 10–49 employees; Large firms: 50 and more employees.
Source: Own calculations based on Innobarometer 2016

the different widths of intervals, it is difficult to statistically evaluate the results of the survey. However, Innobarometer survey does not determine the exact share, only an interval to which the surveyed firm falls under. Thus, it is not possible to unify the methodology of scaling variables and we therefore must use the scales provided by the survey. This methodology is also used by other papers studying various Eurobarometer surveys (e.g. Ehrmann, Soudan, Stracca, 2013; Horváth and Katuščáková, 2016; Capozza and Divella, 2017).

After data cleansing, we worked with 942 observations. Two types of control variables – country dummies and NACE variables – were also included in models. Economic agents in the paper are clustered in line with NACE classification shown in Table 4. Table 5 shows the number of innovative firms (regardless of their country of origin) based on NACE classification and size of a firm in regards with the cleansed number of data used in probit analysis. In Table 6 we present correlation analysis of all variables used in models.

Table 6 Correlation matrix

	y_1	y_2	y_3	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}
y_1	1																	
y_2	–0.472	1																
y_3	0.121	0.055	1															
x_1	–0.005	–0.057	0.199	1														
x_2	–0.032	0.057	0.007	–0.152	1													
x_3	0.037	–0.080	0.099	0.327	–0.012	1												
x_4	–0.008	0.029	0.171	0.061	0.149	0.099	1											
x_5	0.144	0.050	0.222	–0.035	0.132	0.031	0.166	1										
x_6	–0.020	0.142	0.244	0.040	0.071	0.045	0.126	0.277	1									
x_7	0.212	–0.060	0.251	0.205	0.060	0.130	0.099	0.253	0.340	1								
x_8	–0.063	0.145	0.120	0.141	–0.026	0.085	0.034	0.075	0.195	0.224	1							
x_9	–0.022	0.175	0.197	0.114	0.077	0.077	0.073	0.166	0.305	0.220	0.356	1						
x_{10}	–0.020	0.110	0.172	0.117	0.001	0.083	0.155	0.131	0.356	0.187	0.280	0.249	1					
x_{11}	0.092	0.131	0.243	0.031	0.103	0.020	0.058	0.177	0.082	0.124	0.087	0.265	0.039	1				
x_{12}	–0.038	0.193	0.262	0.136	0.025	0.018	0.066	0.142	0.145	0.120	0.179	0.354	0.130	0.265	1			
x_{13}	0.099	–0.049	0.086	0.042	0.005	0.006	0.043	–0.025	0.098	0.054	0.075	0.077	0.107	0.046	0.099	1		
x_{14}	–0.006	0.027	0.014	–0.062	0.017	–0.067	–0.007	–0.004	0.041	–0.023	0.034	0.079	0.034	0.106	0.022	0.163	1	
x_{15}	**0.056**	**–0.032**	**0.041**	**0.035**	**–0.030**	**0.028**	**–0.153**	**–0.011**	**0.021**	**0.081**	**0.063**	**0.073**	**0.048**	**0.106**	**0.083**	**0.110**	**0.133**	**1**

Source: Authors

1.2. Model specification

For analyzing the determinants of firms´ innovation activities in V4 countries we use binary probit models, that correspond to a probabilistic model with the form:

$$P(y_{ij} = 1|x_{ik}, \beta_{jk}) = \Phi(c_j + \beta_{j1}\, x_{i1} + \beta_{j2}\, x_{i2} + \ldots \beta_{j15}\, x_{i15}), \tag{1}$$

where: $\Phi(\cdot)$ is distribution function of a normal distribution $N(0, 1)$.

Our models can be written:

$$y_{ij} = f(dem'_{ij}; inno'_{ij}; bar'_{ij}; c_{ij}; nace_{ij}; country_{ij}) + \varepsilon_{ij}\,, \tag{2}$$

where:

$$y_{ij} = \begin{cases} y_{i1} - \text{product innovation}, \\ y_{i2} - \text{service innovation}, \\ y_{i3} - \text{process innovation}, \end{cases} \tag{3}$$

$$dem'_{ij} = (x_{i1}, x_{i2}, x_{i3}, x_{i4})',$$
$$inno'_{ij} = (x_{i5}, x_{i6}, x_{i7}, x_{i8}, x_{i9}, x_{i10}, x_{i11}, x_{i12})',$$
$$bar'_{ij} = (x_{i13}, x_{i14}, x_{i15})'.$$

Symbol i means the response of a company, j corresponds to a type of innovation, k is a number of a variables, β_{jk} denotes the regression coefficients, vector dem'_{ij} signs demographic variables, $inno'_{ij}$ is a vector of variables of innovation impact, vector bar'_{ij} designates the barrier variables, c_j is an intercept, control variables $nace_{ij}$ and $country_{ij}$ represent a NACE and country dummies and ε_{ij} is an estimate error.

Now we can rewrite a system (1) corresponding to (2) into the following probabilistic models:

$$\begin{aligned} P(y_{i1} = 1|\cdot) &= F(f(dem'_{i1}; inno'_{i1}; bar'_{i1}; c_{i1}; nace_{i1}; country_{i1})) \\ P(y_{i2} = 1|\cdot) &= F(f(dem'_{i2}; inno'_{i2}; bar'_{i2}; c_{i2}; nace_{i2}; country_{i2})) \\ P(y_{i3} = 1|\cdot) &= F(f(dem'_{i3}; inno'_{i3}; bar'_{i3}; c_{i3}; nace_{i3}; country_{i3})) \end{aligned} \tag{4}$$

2 RESULTS

In many studies, the researchers have tried to explain why companies innovate and what are the main drivers of innovations. In this paper we focused on the firms´ innovation activities in V4 countries. Using the maximum likelihood estimation (MLE), we found interesting results. Table 7 presents the results from three probit regression analyses introduced in previous section.

Table 7 Results from probit regression

Explanatory variable	Explained variable		
	Product innovation (y_1)	***Service innovation*** (y_2)	***Process innovation*** (y_3)
Firm´s size (x_1)	–0.1560** (0.0619)	–0.0048 (0.0677)	0.2037*** (0.0616)
Young (x_2)	–0.2200 (0.1382)	0.1244 (0.1565)	–0.1790 (0.1388)
Group (x_3)	0.1397 (0.1277)	–0.2823** (0.1336)	0.0354 (0.1281)

Table 7 (continuation)

Explanatory variable	Explained variable		
	Product innovation (y_1)	***Service innovation*** (y_2)	***Process innovation*** (y_3)
Turnover (x_4)	–0.0841 (0.0642)	0.0486 (0.0683)	0.2060*** (0.0642)
Innovative products and services (x_5)	0.1630*** (0.0357)	–0.0296 (0.0379)	0.1143*** (0.0353)
Investing in innovation (x_6)	–0.1027** (0.0467)	0.1491*** (0.0515)	0.1556*** (0.0471)
R&D (x_7)	0.3146*** (0.0503)	–0.1986*** (0.0529)	0.1616*** (0.0479)
Training (x_8)	–0.1194** (0.0522)	0.1504*** (0.0574)	–0.0119 (0.0521)
Organization investments (x_9)	–0.0421 (0.0526)	0.1293** (0.0575)	–0.0307 (0.0518)
Acquisition of assets (x_{10})	–0.0569 (0.0502)	0.1183** (0.0527)	0.0557 (0.0504)
Marketing innovation (x_{11})	0.2881*** (0.0964)	0.2243** (0.1037)	0.4723*** (0.0942)
Organization innovation (x_{12})	–0.1416 (0.0989)	0.3752*** (0.1060)	0.4408*** (0.0965)
Lack of HR (x_{13})	0.3558*** (0.0976)	–0.2823*** (0.1028)	0.0955 (0.0960)
Regulations and standards (x_{14})	–0.0502 (0.1021)	–0.0264 (0.1115)	0.0002 (0.1016)
Competitors (x_{15})	0.1174 (0.1007)	–0.1819* (0.1068)	0.0234 (0.1008)
Intercept (c)	1.0647*** (0.2701)	–1.1630*** (0.2918)	–1.6774*** (0.2783)
Control variable			
NACE	Yes	Yes	Yes
Country	Yes	Yes	Yes
Log likelihood	–540.95	–462.86	–543.63
AIC	1.1847	1.0187	1.1904
SIC	1.2872	1.1211	1.2929
McFadden *pseudo* R^2	0.1385	0.1568	0.1706
% Correctly predicted	69.80	75.84	69.80

Notes: Standard errors are in parentheses. * p < 0.1, ** p < 0.05, *** p < 0.01.
Source: Authors

If we look at the first group of explanatory variables, namely demographic variables, we find that firm´s size is statistically significant in two types of innovations – product and process. It seems that bigger firms invest more money to introduce a new technology or method (process innovation) than smaller firms. On contrary, smaller firms are more efficient in introducing new products. There findings are in line with Baldwin and Gu (2004) and Fritz (1989). The variable *young* is not statistically significant and the variable *group* is significant and negatively associated with service innovation. It means that being

a part of a business group seems to be a disadvantage when it comes to introducing new services. One of the most important financial indicators is a firm´s turnover. Stable and sustainable turnover growth is a key factor for long-turn success of firm. In our analysis, increasing (decreasing) *turnover* growth is positively (negatively) associated with the process innovation.

Many ratios are calculated on the turnover basis. Our models are no exceptions. We have included several variables. For example, greater percentage of turnover due to *innovative products and services* has a positive impact on product and process innovation. For a company to gain a competitive advantage, it is necessary to make an effort to improve its innovation activity. It is also important to create a business strategy to identify key factors affecting the level of innovation activities. A statistically significant variable supporting these claims is *investing in innovation*. This variable is positively associated with service and process innovation, but it has negative impact on product innovation. Many studies (such as Zemplinerová and Hromádková, 2012; Vokoun, 2014; Griffith et al., 2003; Crépon, Duguet, Mairessee, 1998) focus on examining the impact of R&D in innovation activities. R&D helps stimulate the innovation performance to make the business processes more efficient. We found that the higher investment of company´s turnover in R&D has a positive impact on product and process innovations and negative impact on service innovation. This may be due to the fact that the R&D is frequently oriented towards products and processes rather than services. We also found that higher percentage of turnover spent on *training* employees has a positive impact on service innovation. Training programs help employees improve their knowledge and skills and, consequently, lead to higher productivity. Service innovation also seems to be positively impacted by *organization improvements* and *acquisition of assets*.

In addition to product, service and process innovations, the Innobarometer 2016 survey also examined an introduction of marketing and organization innovations. *Marketing innovation* represents the implementation of new marketing methods such as design creation, product promotion and placement. We can describe *organizational innovation* as an introduction of a new organizational method or improvement of business relationships. We used these variables as explanatory variables to our three main types of innovation – product, process and service innovation. *Marketing innovation* seems to have a positive impact on product, service and process innovation. This means that the introduction of marketing method is an important innovation activity. The explanatory variable *organizational innovation* is statistically significant and means that if a company introduces an *organization innovation*, the probability of introduction a service and process innovation will also increase. We can therefore state that marketing and *organizational innovation* are crucial determinants of innovation activities in firms and support the introduction of other types of innovation.

The last group of variables presented in the paper are barrier variables. Figure 2 in section Methodology illustrates the major barriers that firms face. In our analysis we used three most relevant barrier variables: *lack of HR*, *regulations and standards* and *competitors*. Based on the results, we can state that the *lack of HR* is significant in two output innovation variables: product and service innovation. Human resources are very important especially for the service sector. At present, many companies are struggling with the problem of lack of skilled human resources. In V4 countries, this is most apparent in health and IT sector. The lack of human resources is mainly due to the lack of labor force, the migration of more skilled labor force abroad and inability to adapt to the dynamics of innovation changes. However, we find it interesting that the *lack of HR* is positively associated with the product innovation. This means that the *lack of HR* is not a barrier to product innovation, but, on contrary, is a factor that positively affects introduction of new products. This may be due to the fact that, at present, we face Industry 4.0 and many processes are being automatized. Labor force is being replaced by fully automated lines and machines, hence the *lack of HR* ceases to be a problem in product innovation to some extent. The variable *regulations and standards* is not statistically significant and the variable *competitors* is significant and negatively associated with the service innovation, which means that competition proves to be a significant barrier to service innovation.

CONCLUSION

It is indisputable that innovation is crucial in terms of growth and competitiveness of firms and thus for the whole economy. However, despite these benefits brought by innovation, just a few firms are involved into innovation activities. The aim of the paper was to examine determinants affecting involvement of firms in innovation activities in V4 countries.

We analyzed data from Innobarometer 2016 survey for V4 countries. We used probit models to determine key factors affecting the involvement of V4 firms in innovation activities. Determinants were divided into three categories: demographic variables, variables of innovation impact and barrier variables. We examined the impact of these variables on three different innovation activities firms could have undertaken: product innovation, service innovation and process innovation. We found that product innovation is mostly introduced by smaller firms oriented towards R&D that also introduced new marketing methods. On the other hand, process innovation is mostly developed in larger firms with higher turnover that also invest more in innovation. R&D, marketing innovation and organization innovation are also important determinants of process innovation in V4 countries. Service innovation can mostly be found within firms that invest in innovation and focus on training their employees. Introduction of new organization and marketing methods are also drivers of service innovation in firms. However, being a part of a business group seems to be a disadvantage when it comes to introducing service innovation. Main barriers of innovation were lack of human resources, regulations and standards and competition on the market.

Even though there are many papers focused on examining determinants of innovation, only a small percentage of them uses Eurobarometer surveys in their analysis. The papers aimed at examination of Eurobarometer surveys are mostly focused on analysis of all 28 EU countries, which present an important transnational overview, but sometimes provide overly generalized results and recommendations. We think that the use of firm-level data is significant in finding the key drivers of innovation, while using data for a smaller group of countries (such as a sample of V4 countries) provides specific results that have direct implications related to innovation activities of firms in these countries. Our results can be therefore further used by policy makers in creating optimal innovation policies in a country. However, we realize that our research also has its restrictions. Since the questions asked in Eurobarometer surveys change annually, it is difficult to compare the results between years, so our research is only based on data obtained within one year. We therefore think that it would be interesting to select several questions that are repeated in the questionnaire for more than one year and analyze the data in the longer run. It may be useful to see how the answers of surveyed firms change between years and to look for the changes that may have influenced respondents' answers.

ACKNOWLEDGMENT

The paper was financially supported by the grant scheme VEGA 1/0385/19 *"Determinants of business innovation performance on the basis of Quadruple helix model"* of the Ministry of Education, Science, Research and Sport of the Slovak Republic.

References

BALDWIN, J. R. AND GU, W. Innovation, Survival and Performance of Canadian Manufacturing Plants. *SSRN Electronic Journal*, 2004, 22, 49 p.

BOŽIĆ, L. AND BOTRIC, V. Innovation investment decisions: are post(transition) economies different from the rest of the EU? *Eastern Journal of European Studies*, 2017, 8(2), pp. 25–43.

CAPOZZA, C. AND DIVELLA, M. Organizational changes and innovation: firm-level evidence from European countries. *Annali del Dipartimento Jonico*, 2017, pp. 41–54.

CRÉPON, B., DUGUET, E., MAIRESSE, J. Research, Innovation, and Productivity: An Econometric Analysis at the Firm Level. *Economics of Innovation and New Technology*, 1998, 7, pp. 115–158.
CZARNIEWSKI, S. Smal and medium-sized enterprises in the context of innovation and entrepreneurship in the economy. *Polish Journal of Management Studie*s, 2016, 13(1), pp. 30–39.
D'ESTE, P., IAMMARINO, S., SAVONA, M., TUNZELMANN, N. V. What hampers innovation? Revealed barriers versus deterring barriers. *Research Policy*, 2012, 41, pp. 482–488.
DE JONG, J. P. J. AND VERMEULEN, P. A. M. Determinants of product innovation in small firms: A Comparison Across Industries. *International Small Business Journal*, 2006, 24 p.
EBRD. *Innovation in Transition* [online]. EBRD Transition Report 2014. [cit. 20.3.2019]. <https://www.ebrd.com/publications/transition-report-2014-english.pdf>.
EHRMANN, M., SOUDAN, M., STRACCA, L. Explaining European Union Citizens' Trust in the European Central Bank in Normal and Crisis Times. *Scandinavian Journal of Economics*, 2013, 115(3), pp. 781–807.
FILIPPETTI, A. AND ARCHIBUGI, D. Innovation in times of crisis: National Systems of Innovation, structure and demand. *Research Policy*, 2011, 40(2), pp. 179–192.
FRITZ, W. Determinants of Product Innovation Activities. *European Journal of Marketing*, 1989, 10, pp. 32–43.
GRIFFITH, R., REDDING, S., VAN REENEN, J. Mapping the two faces of R&D: Productivity growth in a panel of OECD industries. *Review of Economics and Statistics*, 2004, 86(4), pp. 883–895.
GUADALUPE, M., KUZMINA, O., THOMAS, C. Innovation and Foreign Ownership. *The American Economic Review*, 2012, 7, pp. 3594–3627.
GUERZONI, M. An Application of Graphical Models To The Innobarometer Survey: A Map of Firms's Innovative Behaviour. *SSRN Electronic Journal*, 2014, 22 p.
HASHI, I. AND STOJIC, N. The impact of innovation activities on firm performance using a multi-stage model: Evidence from the Community Innovation Survey 4. *CASE Network Studies & Analyses,* 2010, 410, 39 p.
HORVÁTH, R. AND KATUŠČÁKOVÁ, D. Transparency and trust: the case of the European Central Bank. *Applied Economics*, 2015, 58(57), pp. 1–14.
LESÁKOVÁ, Ľ. et al. *Determinanty inovačnej aktivity malých a stredných podnikov v SR.* 1st Ed. Banská Bystrica: Matej Bel University, 2010.
LORENZ, E. Do Labour Markets and Educational and Training Systems Matter for Innovation Outcomes? A multi-level analysis for the EU27. *Science and Public Policy*, 2011, 38(9), pp. 691–702.
LÖÖF, H. AND HESHMATI, A. On the relationship between innovation and performance: a sensitivity analysis, *Economics of Innovation and New Technology*, 2006, 15, pp. 317–344.
MAIRESSE, J. AND ROBIN, S. Innovation and productivity: a firm-level analysis for French Manufacturing and Services using CIS3 and CIS4 data (1998–2000 and 2002–2004). *Working paper,* Paris: CREST-ENSAE, 2009, 20 p.
MATARADZIJA, A. et al. *Innovation and innovative performance in the European Union* [online]. International Conference 2013: Active Citizenship by Knowledge Management & Innovation. [cit. 12.2.2019]. <http://www.toknowpress.net/ISBN/978-961-6914-02-4/papers/ML13-225.pdf>.
MONTRESOR, S. AND VEZZANI, A. Intangible investments and innovation propensity: Evidence from the Innobarometer 2013. *Industry and Innovation*, 2016, 23(4), pp. 331–352.
OECD. *Oslo Manual. Guidelines for Collecting and Interpreting Innovation Data.* 3rd Ed. Paris: OECD Publications, 2005.
OUM, S., NARJOKO, D., HARVIE, CH. Constraints, Determinants of SME Innovation, and the Role of Government Support [online]. *ERIA Discussion Paper Series*, 2014. [cit. 20.3.2019] <http://www.eria.org/ERIA-DP-2014-10.pdf>.
POLDER, M., LEEUWEN, G., MOHNEN, P., RAYMOND, W. Product, Process and Organizational Innovation: Drivers, Complementarity and Productivity Effects. *MPRA Paper*, University Library of Munich, Germany. 2010.
RAYMON, L. AND ST-PIERRE, J. R&D as a determinant of innovation in manufacturing SMEs: An attempt at empirical clarification. *Technovation*, 2010, 30, pp. 48–56.
REHMAN, N. U. Innovation performance of Chilean firms, a bivariate probit analysis. *Journal of Entrepreneurship in Emerging Economies*, 2016, 8(2), pp. 204–224,
ROMIJN, H. AND ALBALADEJO, M. Determinants of innovation capability in small electronics and software firms in southeast England. *Research Policy*, 2002, 31, pp. 1053–1067.
ROSA, J. M. Determinants of Product and Process Innovation in Canadas' Dynamic Service Industries [online]. *Working papers – Science, Innovation and Electronic Information Division*, 2002. [cit. 20.3.2019] <https://pdfs.semanticscholar.org/5123/d88298e9ae2bd2677d785679abade75c6f73.pdf>.
STEPHENS, S. Innovation in micro enterprises: reality or fiction? *Journal of Small Business and Enterprise Development*, 2016, 23(2), pp. 349–362.
TETHER, B. S. Do services innovate (Differently)? Insights from the European innobarometer survey. *Industry and Innovation*, 2005, 12(2), pp. 153–184.
TRIGO, A. Mechanisms of learning and innovation performance: The relevance of knowledge sharing and creativity for Non-Technological Innovation. *International Journal of Innovation and Technology Management*, 2013, 10(6), article number 1340028.

VAN LEEUWEN, G. AND KLOMP, L. On the contribution of innovation to multi-factor productivity growth. E*conomics of Innovation and New Technology*, 2006, 15, pp. 367–390.
VEGA-JURADO, J. The effect of external and internal factors on firms' product innovation. *Research Policy*, 2008, 37, pp. 616–632.
VOKOUN, M. R&D and innovation activities – search for better definitions and an economic-historical approach. *2nd Economics & Finance Conference*, 2014, pp. 553–576.
YACHMENEVA, V. AND VOL'S'KA, G. Factors influencing the enterprise innovation. *Econtechmod. An International Quarterly Journal*, 2014, 1, pp. 133–138.
ZEMPLINEROVÁ, A. AND HROMÁDKOVÁ, E. Determinants of firm´s innovation. *Prague Economic Papers*, 2012, 4, pp. 487–503.

On the Measurement of the Income Poverty Rate: the Equivalence Scale across Europe

Martina Mysíková[1] | *Institute of Sociology of the Czech Academy of Sciences, Prague, Czech Republic*
Tomáš Želinský[2] | *Institute of Sociology of the Czech Academy of Sciences, Prague, Czech Republic*

Abstract

The methodology used to determine the at-risk-of-poverty rate commonly applied in the European context is often criticised for arbitrary steps in its construction. This study questions the first step – the equivalence scale applied to transform the disposable income of households of different sizes into comparable units. First, we hypothesise that economies of scale are lower in Central-Eastern European countries than in their Western counterparts. We assess the hypothesis using a simple descriptive analysis of the structure of household consumption expenditures based on Household Budget Survey data. Second, we demonstrate the sensitivity of the at-risk-of-poverty rate to an equivalence scale based on the Statistics on Income and Living Conditions data. We identify three different groups of countries according to the sensitivity of the income poverty rate to the relative adult and child household member weights assigned by the equivalence scale. The study contributes to the discussion on defining accurate, country-specific equivalence scales.

Keywords

Central-Eastern Europe, equivalence scale, income poverty rate, sensitivity, Western Europe

JEL code

I32

INTRODUCTION

Income has been thoroughly analysed from numerous perspectives. For instance, total household income is examined in studies on income inequality and income sources, and individual income and earnings are included when researchers are interested in its contributory factors. However, calculating an equivalent income per household member is often a more convenient measure, for instance, in studies on income poverty indicators. Income poverty can be assessed using objective or subjective, and relative or absolute approaches. The objective and relative approach prevails in the European environment, where the at-risk-of-poverty rate is derived as the share of people whose equivalised disposable household income

[1] Institute of Sociology of the Czech Academy of Sciences, Jilská 1, 110 00 Prague 1, Czech Republic. Corresponding author: e-mail: martina.mysikova@soc.cas.cz.
[2] Institute of Sociology of the Czech Academy of Sciences, Jilská 1, 110 00 Prague 1, Czech Republic.

falls below 60% of the relevant national median income. The absolute level of the poverty threshold thus differs for each country. This relative approach, therefore, captures income disparity across countries "to some extent". Determination of the poverty line and estimation of the poverty rate depend heavily on the equivalence scale used to obtain the "equivalised" household income.

The commonly used OECD-modified equivalence scale was adopted in the EU in the 1990s (as a modification of the original 1980s OECD scale), and even the authors of the scale warned that *"...more research efforts should be devoted to the choice of equivalence scales which can be used for cross-country comparisons. One principal issue to be resolved is whether in the cross-country comparisons we should use a single equivalence scale for all the Member States, or whether a single methodology should be applied to estimate equivalence scales which can be different across different countries."* (Hagenaars et al., 1994, p. 194). It is understood that economies of scale can be strongly country-specific, depending on the national structure of living costs, consumption of durable and non-durable goods, and goods with different economies of scale in general.

The literature on the sensitivity of income-based poverty and inequality measures to equivalence scales was relatively rich up to two decades ago (Buhmann et al., 1988; Coulter et al., 1992; Jenkins et Cowell, 1994; Banks et Johnson, 1994; Lanjouw et Ravallion, 1995; Burkhauser et al., 1996; de Vos et Zaidi, 1997; Aaberge et Melby, 1998). Most of the 1980s and 1990s studies took into account a very limited number of equivalence scales, and only a minority considered analysing a wider range of weights. Recently, scholars have been more focused on construction of equivalence scales based on different approaches, while comparing their sensitivity to commonly adopted equivalence scales (see, e.g., Bishop et al., 2014), assessing the robustness of poverty rates (Cheung et Chou, 2017), analysing differences in income characteristics between subpopulations (see, e.g., Posel et al., 2016), or cross-country comparisons with respect to the sensitivity to equivalence scales (Dhongde et Minoiu, 2013; Ravallion, 2015).

There is a wide range of possible scales between the extremes of ignoring household size (i.e., using a total household income) and applying income per capita. The scale can be derived according to equivalence elasticity, by a rule of thumb, or developed empirically based on survey data. The choice of the scale substantially influences cross-country comparisons, the ranking of countries on both poverty and inequality scales, and the demographic composition of poor populations (Buhmann et al., 1988). Scales have usually been estimated based on consumption/expenditure data (Lazear et Michael, 1980; Van der Gaag et Smolensky, 1982) or subjective data such as income evaluation question (Kapteyn et al., 1988; Van Praag et al., 1982), minimum income question (Danziger et al., 1984), or income satisfaction (van Praag et Ferrer-i-Carbonell, 2004). The literature on equivalence scales in the CEE countries, or the Czech Republic particularly, is scarce. Partially, the topic has been examined by Brázdilová et Musil (2017) and previously by Želinský et Tartaľová (2012), in the Czech and Slovak contexts, respectively; while empirical research has been focused on income poverty more generally (for instance, Bartošová et Želinský, 2013; Večerník et Mysíková, 2016; Mysíková et al., 2019).

The OECD (-modified) equivalence scale was established long before the current European Union was formed. Research in that period was mainly driven by the leading Western European countries. The former socialist Central and Eastern European block then adopted the "Western European" equivalence scale when they joined the EU, regardless of differences in the structures of household consumption expenditures which inevitably existed. Even if we assume that the 1990s equivalence scale fits the current Western European consumption structure, it is very likely that the scale does not accurately reflect the current structure of consumption in Central and Eastern European countries.

First, this paper aims to justify the hypothesis that the same set of equivalence scales should not be used uniformly across Europe. The methodological and empirical literature on equivalence scales was booming more than two decades ago, but has taken a backseat since. We highlight the differences between Central-Eastern and Western European regions to motivate the current research to focus specifically on national

equivalence scales. We argue that equivalence scales should reflect the economies of scale of a particular country, and thus should be based on the consumption structure of that particular country. In order to assess this hypothesis, we perform a descriptive analysis of consumption expenditure structures in Central-Eastern and Western European countries (Section 1). The second goal of this study is to demonstrate the sensitivity of the impact of the equivalence scale applied on the resulting at-risk-of-poverty rate. The sensitivity analysis aims to identify countries which should be cautious about interpreting their income poverty rate and applying anti-poverty policies based on the OECD-modified equivalence scale (Section 2). The final section summarizes, concludes, and describes further steps that should be undertaken in order to achieve more comparable indicator of income poverty across Europe.

1 CONSUMPTION EXPENDITURE STRUCTURE

The Household Budget Survey (HBS) is conducted in EU countries every five years, and provides information on the detailed structures of household consumption expenditures.[3] The structure of household expenditures can serve as an appropriate tool to define at least the basic features of country-specific expenditure behaviour – and so is a clue in indicating country specific or regional differences in equivalence scales. First, we hypothesise that economies of scale are substantially different

Table 1 Structure of consumption expenditure by COICOP (%) – regional averages (weighted by country population share)

	2005				2010				2015			
	CEE		WE		CEE		WE		CEE		WE	
CP01 Food and non-alcoholic beverages	29.0*	(9.0)	12.7*	(2.6)	24.4*	(5.0)	14.0*	(2.6)	23.2*	(4.4)	14.0*	(3.0)
CP02 Alcoholic beverages, tobacco and narcotics	3.5	(1.4)	2.3	(0.6)	3.4	(1.4)	2.2	(0.5)	3.3	(1.6)	1.9	(0.5)
CP03 Clothing and footwear	5.3	(1.1)	5.6	(1.0)	4.5	(1.0)	5.1	(0.9)	4.6	(0.7)	4.7	(0.3)
CP04 Housing, water, electricity, gas and other fuels	25.2	(7.6)	28.2	(2.3)	32.9*	(5.0)	27.6*	(4.8)	32.5	(5.0)	32.5	(2.4)
CP05 Furnishings, household equipment and routine maintenance of the house	4.5*	(1.0)	5.8*	(0.7)	4.2*	(1.2)	5.4*	(1.0)	4.2	(1.0)	4.7	(0.7)
CP06 Health	3.8	(1.0)	3.1	(1.4)	3.9	(0.8)	2.9	(1.3)	4.0	(1.0)	3.9	(1.1)
CP07 Transport	8.6*	(2.8)	12.9*	(1.4)	8.1*	(2.5)	13.5*	(1.5)	8.2*	(2.7)	12.3*	(1.4)
CP08 Communications	4.9*	(0.6)	2.9*	(0.3)	4.2*	(0.5)	2.8*	(0.4)	4.4*	(0.6)	2.7*	(0.4)
CP09 Recreation and culture	6.2*	(2.3)	9.5*	(2.7)	6.2	(2.7)	9.1	(3.0)	5.6	(2.2)	7.9	(2.8)
CP10 Education	0.9	(0.3)	1.0	(0.5)	0.8	(0.4)	1.1	(0.7)	0.7	(0.3)	1.0	(0.6)
CP11 Restaurants and hotels	2.6*	(1.5)	6.2*	(2.2)	2.7*	(1.5)	6.5*	(2.1)	3.4*	(1.6)	6.1*	(1.8)
CP12 Miscellaneous goods and services	5.5*	(2.3)	9.9*	(2.9)	4.7*	(1.5)	9.8*	(2.7)	5.7*	(1.7)	8.4	(1.9)

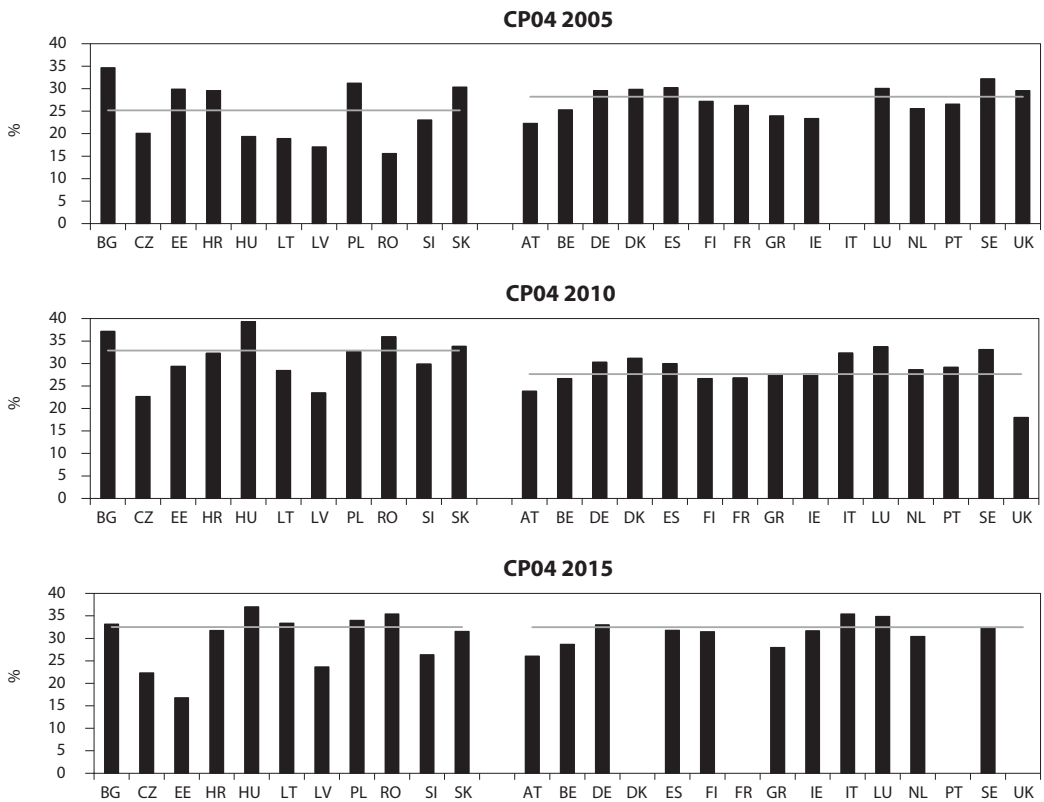
Notes: * The means in Eastern and Western Europe are statistically different at the 5% level (t-test). Standard deviations in parentheses.
Source: Eurostat database (variable hbs_str_t211) based on the Household Budget Survey; average population (Eurostat database, variable demo_gind) used for weights; authors' calculations

[3] HBS is not fully harmonised by Eurostat, meaning that countries have a certain degree of freedom in the survey outcomes they deliver (e.g., CZ used quota sampling up to the HBS 2015; next HBS wave will have been conducted on random sampling). These possible differences must be taken into consideration.

between the Central and Eastern (CEE) and Western (WE) European regions.[4] The Central-Eastern region is composed of post-communist countries distinguished by relatively low wages, while the Western region includes "old" EU-member states with typically higher wages. However, for the purposes of our study, the structure of consumption expenditures together with the related economies of scale of the most substantial consumption expenditures categories (COICOP classification) are of greater importance than income level.

Table 1 shows the differences in consumption structure between Central-Eastern and Western Europe according to the basic COICOP classification (twelve categories). The largest share of consumption expenditures is represented by "Housing, water, electricity, gas and other fuels" ("Housing" hereafter, COICOP 4), which comprises on average about 30% of household expenditures in both CEE and WE, with a statistically significant difference only in 2010. Though the housing consumption expenditures are relatively similar at the regional level, countries in the CEE region exhibit a substantially higher variance than those in WE. The bar charts in Figure 1 support this, suggesting a few different clusters

Figure 1 Consumption expenditure on Housing, water, electricity, gas and other fuels (% of total expenditures)



Source: Eurostat database (variable hbs_str_t211) based on the Household Budget Survey; average population (Eurostat database, variable demo_gind) used for weighted mean (depicted by the horizontal lines); authors' calculations

[4] The division corresponds to the new and old EU member countries. However, we exclude Malta and Cyprus from the analysis, as they are not post-communist countries.

of countries within the region. Clearly, one group would consist of CZ and LV, as these countries are located far below the CEE average in all three time periods observed. The opposite group of countries, which are always above or around the CEE average would include BG, HR, PL, and SK.[5] The rest of the CEE countries are more difficult to evaluate at first glance as, for instance, the share of expenditure on housing was substantially decreasing over time in EE.

Housing expenditures can be expected to exhibit large economies of scale; for instance, the costs of a single individual change only marginally when a second person moves into the household. The structure of consumption expenditures is relevant for economies of scale: the larger the share of housing expenditures in the total household budget is, the higher the overall economies of scale are. Therefore, we suppose that at least a part of the CEE[6] has lower economies of scale from cohabitation than is typical in WE countries. Consequently, with respect to the main idea of the equivalence scale concept, the weight of second (and additional) person/s in the household should be higher in these CEE countries than in WE countries.

The consumption expenditure on "Food and non-alcoholic beverages" ("Food" hereafter, COICOP 1) is the second largest item in household budgets. On average, across all European countries included in the analyses, it comprises 17% of household budgets, but the differences between the CEE and WE regions are substantial: "Food" accounts for roughly 25% of household expenditures in CEE countries, but only about 14% in WE (see Table 1), with the difference being highly statistically significant. As opposed to housing expenditures, food is expected to exhibit very low economies of scale. Though joint cooking might be more efficient than cooking separately, we can assume that individuals consume the same volume of food regardless of whether they live separately or in a shared household. With the higher share of expenditures on food in the CEE, we again assume that economies of scale arising from shared living situations are lower in CEE than in WE countries, with almost complete uniformity across all CEE countries.

Similarly to housing, the variability of food expenditures among CEE countries is somewhat greater than in WE (see Figure 2). The largest share of consumption expenditure on food is in RO, BG, and LT, and the smallest in SI. No CEE country spends lower share on food than any WE country, except SI. Therefore, we assume that food consumption expenditures considerably support our hypothesis that there are lower economies of scale in CEE countries and, thus, the greater weight of the second (and additional) person/s in the household on the equivalence scale.

Each of the remaining categories of consumption expenditures comprise about 10% or less of household budgets. The following categories, in descending order of their share of the total expenditures, are: Transport (COICOP 7), Miscellaneous goods and services (COICOP 12), Recreation and culture (COICOP 9), and Restaurants and hotels (COICOP 11).[7] Inhabitants of WE countries spend, on average, a larger share of their household budgets on these four categories than those living in CEE countries. These categories are miscellaneous in nature, and we do not intend to speculate about their economies of scale at this level of our analysis.

However, the two main consumption categories (Housing and Food), which have clearly predictable directions of economies of scale, account for about 55% of all household expenditures in CEE, and roughly 45% in WE. Though this descriptive analysis does not provide any "proof", it clearly indicates that economies of scale can be expected to be lower in the CEE than in the WE region, and that the weight of the second (and additional) household member/s should be higher in CEE. The next section focuses on the consequences of using different equivalence scales.

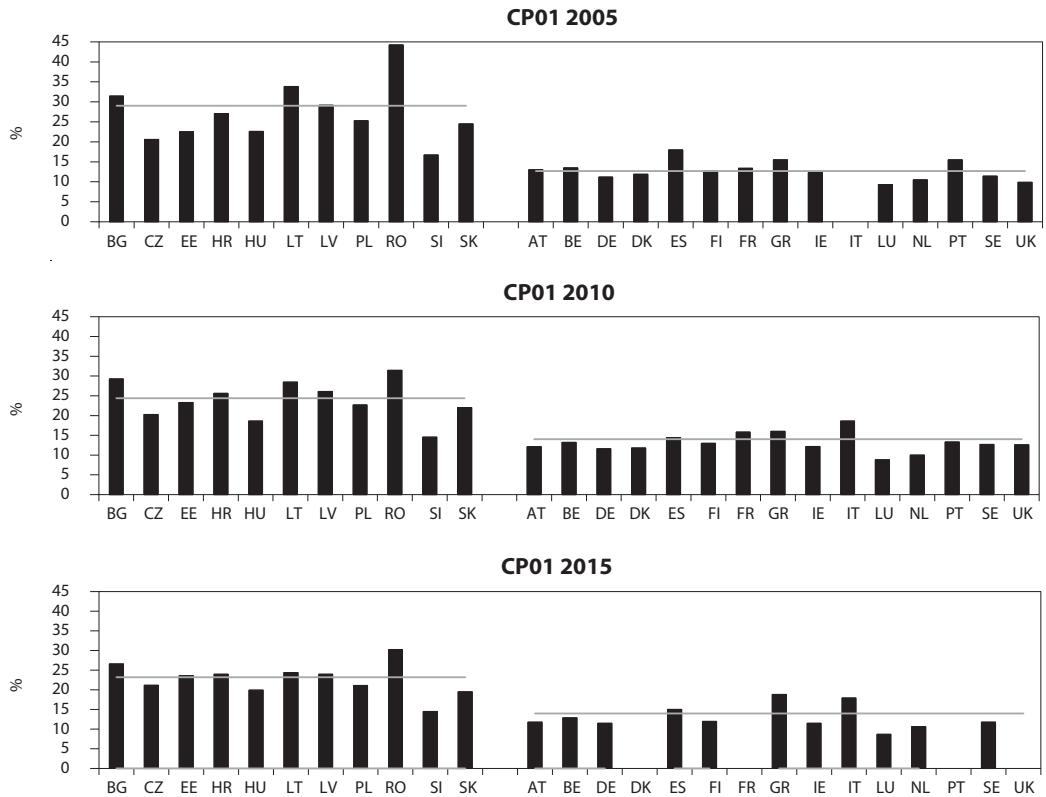
---

[5] Country abbreviations are stated in Table 2.

[6] Countries with substantially lower shares of consumption expenditures on housing than are common in WE countries.

[7] The other six categories (COICOP 2, 3, 5, 6, 8, and 10) do not reach 5% of consumption expenditure share in either region.

Figure 2 Consumption expenditure on Food and non-alcoholic beverages (% of total expenditures)



Source: Eurostat database (variable hbs_str_t211) based on the Household Budget Survey; average population (Eurostat database, variable demo_gind) used for weighted mean (depicted by the horizontal lines); authors' calculations

The CEE countries with statistics that most strongly support our assumptions are those with below-average shares of expenditures on Housing and above-average expenditures on Food: LT and LV. In WE, the opposite direction of shares of expenditures conforming to our assumptions, i.e., above-average shares of expenditures on Housing and below-average expenditures on Food: LU, DE, DK, and SE play into our hands. On the other side, data on PL and SK in the Central-Eastern region and GR in the Western region contradict our assumptions.

2 SENSITIVITY ANALYSIS OF THE IMPACT OF EQUIVALENCE SCALES ON THE AT-RISK-OF-POVERTY RATE

In the previous section, we described clues that signal lower economies of scale in CEE than in WE. Now we proceed to illustrate the sensitivity of the resulting at-risk-of-poverty rates to equivalence scale. We focus mainly on the difference between the CEE and WE regions, though the CEE region seems to be more heterogeneous, and will require more focused distinctions in future analyses.

2.1 Data and methodology

We use the European Union – Statistics on Income and Living conditions (EU-SILC, known in CZ as "Životní podmínky") survey data for 2016 (and partially for 2006 and 2011). The survey is compulsory

for all EU member countries and is harmonised by Eurostat. It is thus a convenient data source for international comparisons, and is utilized to determine official poverty statistics. Information is collected at the household and individual levels, and includes core and basic socio-demographic characteristics along with detailed information on income sources and living conditions. The income reference period is the calendar year preceding the dates of the survey in most countries, hence, the income poverty rates from EU-SILC 2016 in fact correspond to 2015, so it fits the HBS 2015 data presented in the previous section of this paper.

The OECD-modified scale, used to calculate the official at-risk-of-poverty rate indicator (income poverty rate, hereafter), assigns a weight of 1 to the first adult household member. All other adults and household members older than 13 are assigned a weight of 0.5, while each child aged 13 or younger has a weight of 0.3. The sum of the weights of all household members then provides the "equivalised household size". The total disposable household income is then divided by the equivalised household size to obtain the equivalised household income.

For a detailed example, imagine a two-adult household, in which each adult has a net monthly income of 10 000 CZK, for a total household income of 20 000 CZK.[8] Their equivalised household size is 1.5, yielding an equivalised income of 20 000/1.5 = 13 333; the equivalent of the income of each adult household member. Computing the income poverty rate as a percentage of persons in the population below the poverty line thus takes into account the economies of scale from living together: the amount of 13 333 CZK is calculated for both adults (rather than the actual income of 10 000 CZK), since they save some costs by living together, though they would each have an income of 10 000 CZK if they lived separately and alone. The poverty line is then expressed as 60% of the median of the equivalised disposable income.

Our main hypothesis is that the weights assigned by the OECD-modified equivalence scale do not necessarily properly reflect the economies of scale from cohabitation and cost-sharing, particularly in Central-Eastern European countries. At this stage of the research, our aim is not to provide new, country-specific equivalence scales. We limit our contribution to providing evidence that the income poverty rates can be highly sensitive to the equivalence scale used. We believe that one of the requirements of a good equivalence scale is low sensitivity of the income poverty rate to the relative weights of adult and child household members. When the income poverty rate changes substantially in response to a moderate change in the equivalence scale, the explanatory power of the income poverty rate is very low and cannot be accurately used to inform social policies.

In order to demonstrate the sensitivity of income poverty to the equivalence scale, we compute the income poverty rates for a wide range of combinations of the weights assigned to adult and child household members. Specifically, we simulate poverty rates on a grid with adult and child weights ranging from 0 to 1 by 0.01 unit. Put differently, we estimate the income poverty rate for each combination of adult and child household member weights in {0, 0.01, 0.02, …, 1}, i.e., we generate a grid of 10 201 different combinations. For instance, were the weights of both (and additional) adults and children equal to zero, the income considered would correspond to total household income (the equivalised household size would equal one), and the economies of scale would be at their maximum (bottom left corners in Figure 3). However, were the weights of both adults and children equal to one, the income considered would correspond to income per capita, meaning that there are no economies of scale at all (right top corners in Figure 3).

Using this approach, we present the main results visually, i.e., we construct level plots with income poverty rate as the response variable, while adult and child household member weights are evaluated on a grid (as described above). We include only selected plots in this paper, but all available plots are available from the authors upon request.

[8] The income poverty rate is calculated from annual income, but monthly income serves better for illustration.

In addition to the visual outputs, for each country we report selected characteristics. We first show the official income poverty rate based on the OECD-modified scale, and the mean income poverty rate based on values of potential poverty rates from our grid of different combinations of adult/child household member weights. Next, we present two simple measures of the sensitivity of the income poverty rate to adult/child household member weights. (1) The overall coefficient of variation of the potential poverty rate (based on the grid) reflects the overall level of the sensitivity of the income poverty rate to adult/child household member weights. Higher values are associated with higher sensitivity to weights. This measure, however, does not allow us to identify whether the resulting level of sensitivity is primarily caused by greater sensitivity to adult or child household member weights. For that reason, we also (2) compute separate coefficients of variation of the income poverty rate for the adult household member weight ranging from 0 to 1, while keeping child household member weight constant (repeatedly for each value of the child weight), and report the mean coefficient of variation. Similarly, we also compute the mean coefficient of variation of income poverty rate with respect to child household member weight, while keeping adult household member weight constant. Comparing the latter two separate measures of variation (see the two last columns of Table 3) allows us to determine whether the income poverty rate is more sensitive to adult or child household member weights, or whether it is the case that the income poverty rate is sensitive to both weights.

2.2 Results

The intersection of the horizontal and vertical lines in Figure 3 corresponds to the actual income poverty rate based on the OECD-modified equivalence scale. The images typically show a part of a "reversed hill": the brighter the area, the lower the resulting income poverty rate. The units of the scale are the same at all figures, which helps us to show the sensitivity of the income poverty rate to the weights assigned by the equivalence scale in an illustratively convenient way.

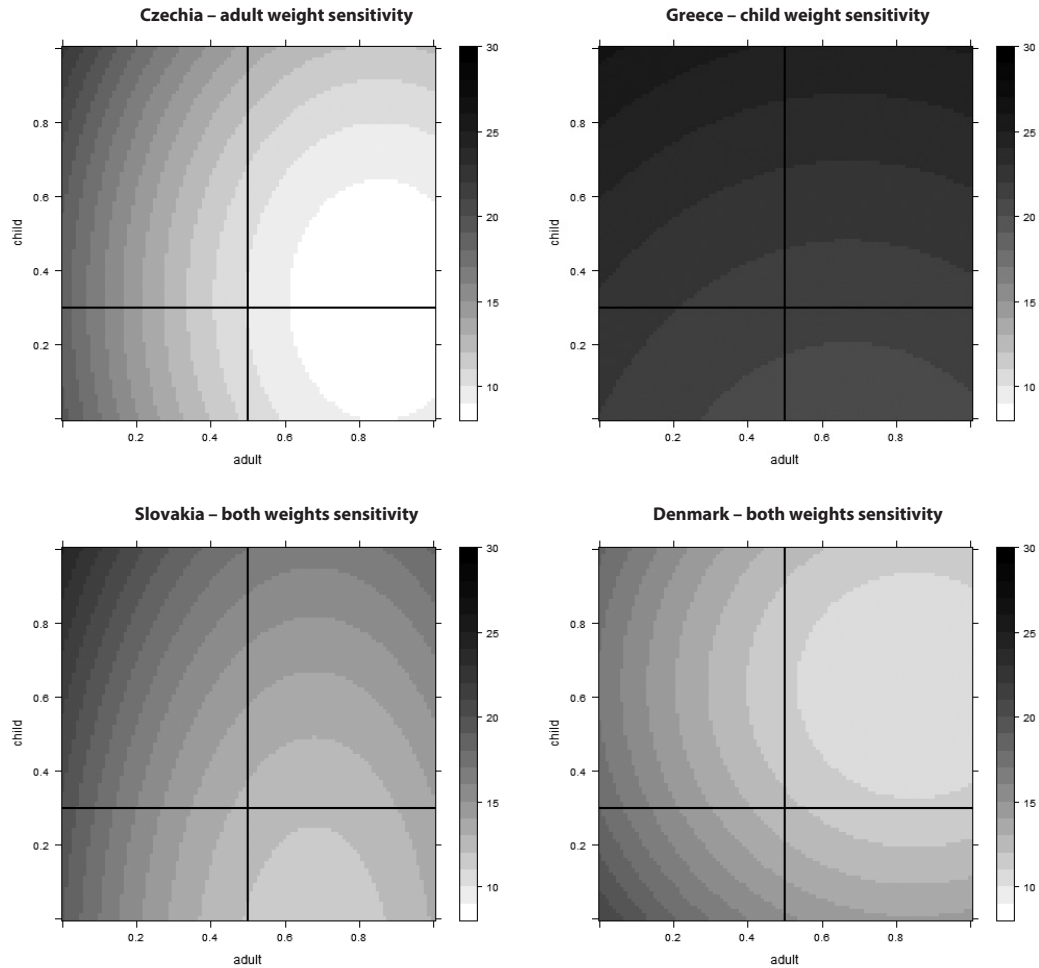
The countries can be roughly divided into three groups. First, countries which exhibit relatively high sensitivity to the adult weight but relatively low sensitivity to the child weight – Czechia is an example of this (see top left panel in Figure 3). Taking the intersection as a starting point (the OECD-modified scale), it is clear that moving along the horizontal line is accompanied by rapid changes in the income poverty rate. On the other hand, moving along the vertical line barely changes it.

Greece serves as an example of the second type of countries – those with relatively high sensitivity to the child weights but very low sensitivity to the adult weights. Here, moving along the horizontal line barely results in a change in the income poverty rate, while moving along the vertical line exhibits rapid changes. The third group of countries can be characterised by a relatively strong sensitivity to both of the weights: changes in either influences the income poverty rate substantially. Slovakia and Denmark form our examples.

Table 2 shows the basic rough division of countries according to their sensitivity to either of the weights, with the OECD-modified equivalence scale as the starting point. Prevailing sensitivity to child weight is rather uncommon - these patterns can be seen only in Greece and Italy. Relatively high sensitivity to adult weights is mildly prevalent in CEE countries, while fewer countries exhibit a sensitivity to both weights. The opposite seems to hold in WE, where sensitivity of the income poverty rate can be assigned to both weights in the majority of countries.

Figure 4 shows how the sensitivity of the poverty rate to equivalence scales developed over time in Czechia and Slovakia. Compared with Figure 3 for CZ and SK, the pictures exhibit relatively stable results. However, from our simple perspective, the income poverty rate was sensitive to both adult and child weights in CZ in 2006, when the intersection is considered a starting point, and the sensitivity to child weights diminished somewhat over time. The Slovakian income poverty rate, on the other hand, gained sensitivity to the child weights (see Figures 3 and 4). The results can be influenced by the combination

Figure 3 Income poverty rate by adult and child weight, 2016



Note: Figures for all countries are not stated due to space restrictions, but are available upon request.
Source: EU-SILC 2016, authors' calculations

Table 2 Sensitivity of income poverty rate by adult and child weight – groups of countries, 2016

	Central and Eastern Europe (CEE)	Western Europe (WE)
Sensitivity to adult weight	Bulgaria (BG)	Belgium (BE)
	Czechia (CZ)	Germany (DE)
	Estonia (EE)	Finland (FI)
	Lithuania (LT)	Ireland (IE)
	Latvia (LV)	
	Slovenia (SI)	

Table 2 (continuation)

	Central and Eastern Europe (CEE)	Western Europe (WE)
Sensitivity to child weight		Greece (GR) Italy (IT)
Sensitivity to both types of weights	Croatia (HR) Hungary (HU) Poland (PL) Romania (RO) Slovakia (SK)	Austria (AT) Denmark (DK) Spain (ES) France (FR) Luxembourg (LU) Netherlands (NL) Portugal (PT) Sweden (SE) United Kingdom (UK)

Source: Authors' classification based on EU-SILC 2016 data

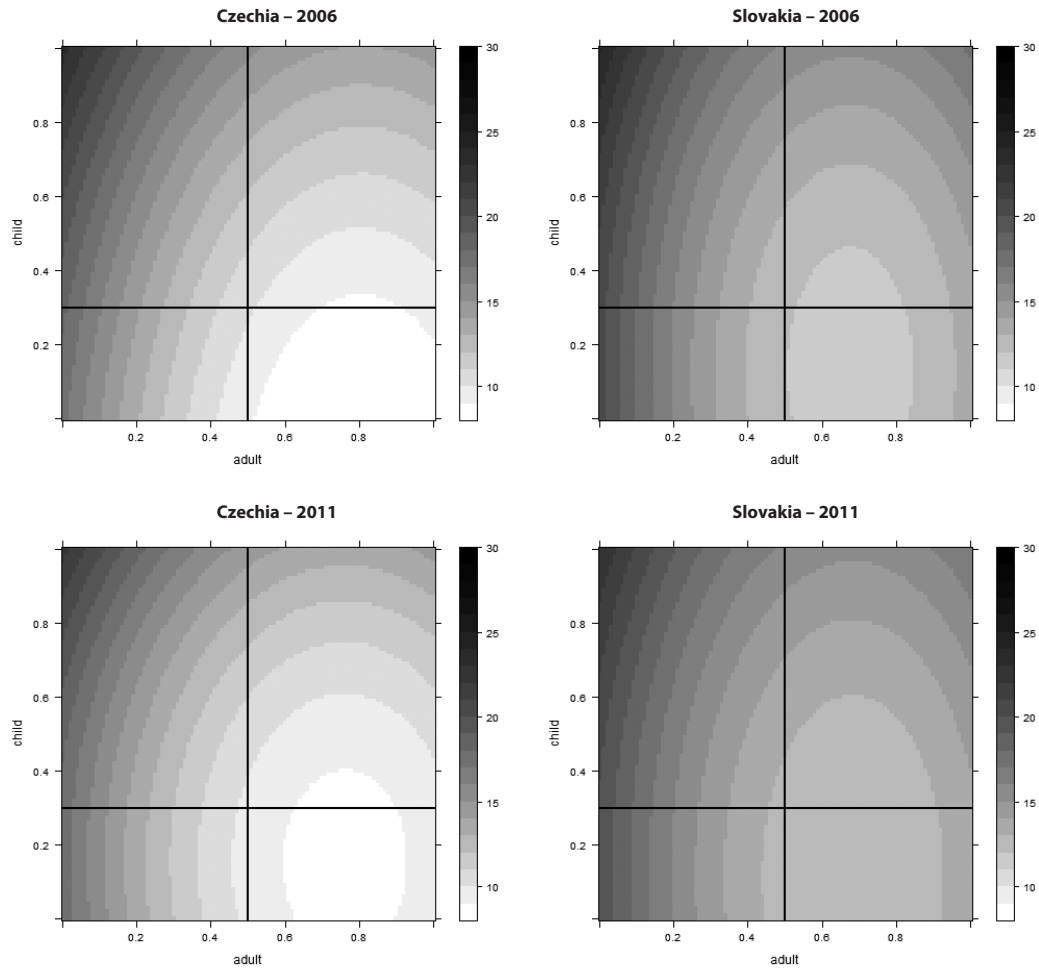
of the structure of consumption expenditures and household composition in a country. This only supports our idea that equivalence scale should not only be country-specific, but should be updated.

It is clear that our simple sensitivity assessment is highly dependent on the starting point, i.e., the currently applied OECD-modified equivalence scale, the validity of which this study questions. Table 3 shows both the income poverty rate for 2016, and its coefficient of variation. In both regions, the lowest income poverty rates are accompanied by the highest coefficient of variation (CZ, SK, and SI in the CEE region, and FI, DK, NL in the WE region), and vice versa (RO and BG in CEE, and ES, IT, GR, PT in WE). In Central-Eastern Europe, the coefficient of correlation of income poverty rate and its variation is –0.81, while it is –0.91 for Western Europe. This means that countries with low income poverty rates tend to have rates that are more sensitive to the equivalence scale applied, while countries with high income poverty rates have rates that are almost insensitive to the scale.

When it comes to particular sensitivity to adult household member weights, the CZ substantially exceeds other CEE countries (followed by SI and EE). Similarly, in the WE region, sensitivity to the adult weight is substantially higher in FI, followed by somewhat lower values in DK and NL. The lowest sensitivity to the adult weight can be seen in RO within the CEE region, and in IT, GR, ES, and PT within the WE region.

Regarding the sensitivity to child household member weights, SK, HU, and CZ are at the top of the ladder in the CEE region, as are LU and AT in the WE region. The bottom of the ladder is occupied by BG, LT, and RO in CEE, and ES and PT in WE. It follows that when we abandon the starting point of the OECD-modified equivalence scale, but consider the whole possible spectrum of weight combinations, Czechia exhibits relatively high sensitivity to both adult and child weights compared to other countries, though the sensitivity to the adult weight prevails in absolute terms.

Figure 4 Income poverty rate by adult and child weight, CZ and SK, 2006 and 2011



Note: Figures for all countries are not stated due to space restrictions, but are available upon request.
Source: EU-SILC 2006, 2011; authors' calculations

Table 3 Income poverty rate characteristics, 2016

	Poverty rate	Coefficient of variation (CV)	Mean poverty rate	Mean CV with respect to adult weight	Mean CV with respect to child weight
CEE					
BG	22.9	0.08	23.2	0.08	0.01
CZ	9.7	0.28	12.1	0.26	0.09
EE	21.7	0.16	20.7	0.16	0.05

Table 3 (continuation)

	Poverty rate	Coefficient of variation (CV)	Mean poverty rate	Mean CV with respect to adult weight	Mean CV with respect to child weight
CEE					
HR	19.5	0.09	20.9	0.08	0.04
HU	14.5	0.13	17.1	0.09	0.09
LT	21.9	0.11	22.8	0.11	0.02
LV	21.8	0.12	21.4	0.12	0.04
PL	17.3	0.10	19.1	0.08	0.06
RO	25.3	0.03	25.7	0.03	0.02
SI	13.9	0.17	15.3	0.17	0.04
SK	12.7	0.18	15.3	0.14	0.10
WE					
AT	14.1	0.14	16.8	0.10	0.10
BE	15.5	0.13	16.8	0.13	0.04
DE	16.4	0.11	17.6	0.10	0.04
DK	11.9	0.18	13.1	0.16	0.09
GR	21.2	0.06	22.6	0.02	0.06
ES	22.3	0.03	23.1	0.03	0.02
FI	11.7	0.23	13.6	0.23	0.06
FR	13.6	0.11	15.5	0.08	0.08
IE	16.6	0.14	17.8	0.13	0.05
IT	20.6	0.05	21.8	0.02	0.04
LU	16.5	0.14	19.1	0.08	0.11
NL	12.7	0.16	14.4	0.15	0.05
PT	19.0	0.07	19.8	0.06	0.03
SE	16.2	0.15	16.9	0.14	0.05
UK	15.9	0.13	18.6	0.09	0.08

Source: EU-SILC 2016; authors' calculations

Though, at this stage of research, we primarily assess the sensitivity of income poverty to equivalence scales using visualisation techniques, our modest results indicate that European countries can be classified into different groups. Our results, showing that the consumption expenditure structure differs across countries, suggest that countries should consider establishing their own national equivalence scales. Moreover, the results described in this section suggest that countries with a high sensitivity of income poverty rate to equivalence scale should pay attention to the selection of adult/child household member weights when defining their national equivalence scales. Otherwise, their official income poverty rates may not necessarily reflect the true nature of income poverty in the country.

CONCLUDING REMARKS AND IMPLICATIONS FOR FOLLOW-UP RESEARCH

This study questions the cross-country comparability of the main, most commonly used indicator of income poverty, the at-risk-of-poverty rate. The construction of this indicator applies a uniform equivalence scale to transform the disposable income of households of different sizes into comparable units. We discuss two different views of reasons to re-examine the OECD-modified equivalence scale and to verify its validity across European countries. First, we provide some insights into why a uniform equivalence scale adopted by all countries should not be used to derive "equivalised" household disposable income, focusing on the apparent differences in consumption expenditure structures between Central and Eastern (CEE) and Western (WE) European regions. Second, we offer a simple analysis of the sensitivity of the income poverty rate to the weights of adult and child household members assigned by the scale in order to identify countries with higher sensitivity to either weight.

Regarding the consumption expenditure structure, the two main categories of goods and services, defined by highest shares of consumption expenditures according to the basic COICOP classification – "Housing" and "Food" – comprise on average half of household expenditures. The share of Housing expenditures, where large economies of scale can be expected, does not exhibit significant differences at the regional level; however, a smaller group of CEE countries with a lower share of expenditures on housing can be identified. Regarding Food, where, on the contrary, relatively low economies of scale are usually expected, CEE countries exhibit substantially higher shares of expenditures than WE countries. These findings strongly indicate lower economies of scale in the CEE than in the WE region. Therefore, it can be concluded that a uniform equivalence scale is not appropriate for all European countries. Moreover, countries with a dynamic change of the structure of consumption expenditures should not only consider to establish their own national equivalence scale, but also to adjust it regularly.

Concerning the sensitivity of the resulting income poverty rates to the equivalence scale, our primary aim was to perform a visual analysis, and to identify groups of countries with similar patterns. We have distinguished three basic groups based on the most recent data. First, countries with relative sensitivity to the adult weights and insensitivity to child weights, which includes most CEE countries. Second, the set of countries with relative insensitivity to adult weights and sensitivity to child weights, which includes only two South-Western European countries. And, third, countries with relative sensitivity to both adult and child weights – WE countries prevail in this group. Ultimately, a uniform pattern can be identified in both regions: the lower the income poverty rate, the higher its variation, and, thus, sensitivity to the equivalence scale. Countries considering establishment of their own country-specific equivalence scale should focus especially on the weights to which their national income poverty rate is sensitive.

Though we do not conclude this study by proposing new country-specific equivalence scales, we believe that a uniform methodology to establish more tailored equivalence scales would be a better way to achieve comparative income poverty indicators than the current use of a uniform equivalence scale. This study only offers reasons and motivation for research which necessarily must continue with identification of national equivalence scales. Our future research studies thus aim to, first, assess the sensitivity of income poverty rates to equivalence scales in a more technical way, and, second, to compare various approaches,

methodologies, and estimation techniques for establishment of national equivalence scales, in conjunction with testing their reliability and validity.

ACKNOWLEDGEMENT

This work was supported by the Czech Science Foundation under Grant No. 18-07036S *"Methodology and reality of poverty: Czech Republic in the European context"*. The EU-SILC datasets were made available on the basis of contract No. 265/14 between the European Commission, Eurostat, and the Institute of Sociology of the Czech Academy of Sciences. Responsibility for all conclusions drawn from the data lies entirely with the authors.

References

AABERGE, R. AND MELBY, I. The sensitivity of income inequality to choice of equivalence scales. *Review of Income and Wealth*, 1998, 4, pp. 565–569.

BANKS, J. AND JOHNSON, P. Equivalence scale relativities revisited. *The Economic Journal*, 1994, 425, pp. 883–890.

BARTOŠOVÁ, J. AND ŽELINSKÝ, T. The extent of poverty in the Czech and Slovak Republics 15 years after the split. *Post-Communist Economies*, 2013, 1, pp. 119–131.

BISHOP, J. A., GRODNER, A., LIU, H., AHAMDANECH-ZARCO, I. Subjective poverty equivalence scales for Euro Zone countries. *The Journal of Economic Inequality*, 2014, 2, pp. 265–278.

BRÁZDILOVÁ, M. AND MUSIL, P. Impact of consumption unit's scale on credibility of the income indicators in the Czech Republic [online]. *Statistika: Statistics and Economy Journal*, 2017, 2, pp. 15–24.

BUHMANN, B., RAINWATER, L., SCHMAUS, G., SMEEDING, T. M. Equivalence scales, well-being, inequality, and poverty: Sensitivity estimates across ten countries using the Luxembourg Income Study (LIS) database. *Review of Income and Wealth*, 1988, 2, pp. 115–142.

BURKHAUSER, R. V., SMEEDING, T. M., MERZ, J. Relative inequality and poverty in Germany and the United States using alternative equivalence scales. *Review of Income and Wealth*, 1996, 4, pp. 381–400.

CHEUNG, K. C. K. AND CHOU, K. L. Measuring child poverty in Hong Kong: Sensitivity to the choice of equivalence scale. *Social Indicators Research*, 2017, 3, pp. 909–921.

COULTER, F. A. E., COWELL, F. A., JENKINS, S. P. Equivalence scale relativities and the extent of inequality and poverty. *The Economic Journal*, 1992, 414, pp. 1067–1082.

DANZIGER, S., VAN DER GAAG, J., TAUSSIG, M. K., SMOLENSKY, E. The direct measurement of welfare levels: How much does it cost to make ends meet? *The Review of Economics and Statistics*, 1984, 3, pp. 500–505.

DE VOS, K. AND ZAIDI, M. A. Equivalence scale sensitivity of poverty statistics for the Member States of the European Community. *Review of Income and Wealth*, 1997, 3, pp. 319–333.

DHONGDE, S. AND MINOIU, C. Global poverty estimates: a sensitivity analysis. *World Development*, 2013, C, pp. 1–13.

HAGENAARS, A. J. M., DE VOS, K., ZAIDI, A. *Poverty statistics in the late 1980s: Research based on micro-data.* Luxembourg: Office for Official Publications of the European Communities, 1994. ISBN 92-826-8982-4.

JENKINS, S. P. AND COWELL, F. A. Parametric equivalence scales and scale relativities. *The Economic Journal*, 1994, 425, pp. 891–900.

KAPTEYN, A., KOOREMAN, P., WILLEMSE, R. Some methodological issues in the implementation of subjective poverty definitions. *The Journal of Human Resources*, 1988, 2, pp. 222–242.

LANJOUW, P. AND RAVALLION, M. Poverty and household size. *The Economic Journal*, 1995, 433, pp. 1415–1434.

LAZEAR, E. AND MICHAEL, R. Family size and the distribution of real per capita income. *American Economic Review*, 1980, 1, pp. 91–107.

MYSÍKOVÁ, M., ŽELINSKÝ, T., GARNER, T. I., VEČERNÍK, J. Subjective perceptions of poverty and objective economic conditions: Czechia and Slovakia a quarter century after the dissolution of Czechoslovakia. *Social Indicators Research*, 2019, 2, pp. 523–550.

POSEL, D., CASALE, D., GRAPSA, E. Re-estimating gender differences in income in South Africa: The implications of equivalence scales. *Development Southern Africa*, 2016, 4, pp. 425–441.

RAVALLION, M. On testing the scale sensitivity of poverty measures. *Economics Letters*, 2015, pp. 88–90.

VAN DER GAAG, J. AND SMOLENSKY, E. True household equivalence scales and characteristics of the poor in the United States. *Review of Income and Wealth*, 1982, 1, pp. 17–28.

VAN PRAAG, B. AND FERRER-I-CARBONELL, A. *Happiness quantified. A satisfaction calculus approach.* New York: Oxford University Press, 2004.

VAN PRAAG, B., HAGENAARS, A. J., VAN WEERDEN, H. Poverty in Europe. *Review of Income and Wealth*, 1982, 3, pp. 345–359.

VEČERNÍK, J. AND MYSÍKOVÁ, M. *Poverty in the Czech Republic: A critical look at EU indicators*. Prague: SOÚ AV ČR, 2016. ISBN 978-80-7330-290-0.

ŽELINSKÝ, T. AND TARTAĽOVÁ, A. Impact of equivalence scale on at-risk-of-monetary poverty rates in the regions of Slovakia. In: PAUHOFOVÁ, I. AND ŽELINSKÝ, T. eds. *Inequality and Poverty in the European Union and Slovakia,* Košice: Technical University of Košice, 2012, pp. 57–66.

Housing Affordability in Slovakia: what Factors Affect it?

Viera Labudová[1] | *University of Economics in Bratislava, Bratislava, Slovakia*
Ľubica Sipková[2] | *University of Economics in Bratislava, Bratislava, Slovakia*

Abstract

Housing affordability represents a challenge everyone faces when covering the costs of their current or potential housing on the one hand and costs unrelated to their housing within the limits of their own income on the other hand. At the international level, two approaches are used to measuring the housing affordability: the ratio approach and the residual approach. According to Eurostat's definition, a household is considered "overburdened" when the total housing costs ("net" of housing allowances) represent more than 40% of disposable income.

The primary objective of this paper was to define the relevant factors affecting the household cost burden in the Slovak Republic and quantifying the intensity of their influence. For this purpose, a logistic regression model and a classification tree model were created, using the sample of the cross-component of the data of the statistical survey EU SILC 2016. The analysis was completed by using SAS Enterprise Guide (SAS EG) and SAS Enterprise Miner (SAS EM).

Keywords

Housing affordability, housing cost overburden rate, EU SILC, logistic regression model, decision tree model

JEL code

C35, I31, R21

INTRODUCTION

Housing is not only a basic human need, it is one of the basic social rights recognised under international legislation (Scanlon, Arrigoitia, Whitehead, 2015). The right to housing is protected by international documents, including the Universal Declaration of Human Rights, International Covenant on Economic, Social and Cultural Rights, the Convention on the Rights of the Child, the International Convention on the Elimination of All Forms of Racial Discrimination, and the Convention on the Elimination of All Forms of Discrimination against Women. While the right to housing is not among the competencies of the European Commission or other institutions of the European Union and its resolution is left completely up to the member states, there are a number of tools related to it. These include the European Social Charter, the Charter of Fundamental Rights of the EU, the EU Treaty, the EU Anti-Discrimination Legislation

[1] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Statistics, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia. E-mail: viera.labudova@euba.sk, phone: (+421)267295733.
[2] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Statistics, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia. E-mail: lubica.sipkova@euba.sk, phone: (+421)267295729.

and the EU Agency for Fundamental Rights. The right to housing is of critical importance to achieving an inclusive and competitive Europe. Access to safe and affordable housing is one of the basic prerequisites for the well-being of European citizens and society (Hegedüs, Elsinga, Horváth, 2016).

Affordability can be evaluated in various ways that lead to different conclusions as to the nature of the problem and the best solutions. Of the objective indicators of housing affordability, the most interesting indicators at EU level are to be found in the SILC survey database. The share of housing costs in disposable income refers to the expenditure on housing compared to the household's income[3]. Housing costs (including utilities) are calculated after deduction of housing allowances[4]. Those who spend more than forty percent of disposable income on housing costs are considered to be burdened with housing costs (Pittini, 2012).

1 HOUSING AFFORDABILITY

Applying the right to housing is linked to two aspects: housing accessibility and housing affordability. There is a fundamental difference between the notions of housing affordability and housing accessibility (Sendi, 2011). Affordability is a market concept related to capacity to pay. Housing is affordable for these, who can afford to pay for it, therefore using this approach they gain access to it. On the contrary, those who cannot afford to pay for housing, lack such access (Sendi, 2011). Accessibility, on the other hand, is a humanitarian concept. The notion of housing accessibility essentially implies the objective to guarantee the right to housing for everyone. Housing is not a market commodity within this concept; rather represents a right that must be guaranteed for every human being. Access to housing relates to the whole population, including those groups of people, who are often limited in the implementation of their rights to have adequate housing in various ways (Sendi, 2011).

The term housing affordability should not be confused with affordable housing, which traditionally refers to a specific kind of housing designed to be affordable for low-income groups. Affordable housing is more of an attempt to alleviate some of the need associated with identified housing affordability problems (Atfield, 2013).

The definition of "affordable housing" varies across economies, but generally it includes a financial component (the share of income devoted to housing), a standard for what constitutes minimum socially acceptable housing with a clear idea of what income groups are affected, and at what income level households should be eligible for housing assistance. The definition should accommodate a range of sizes, tenure options (purchase vs. rental), and affordability thresholds that take into account households of different sizes and incomes in the area. In many parts of the world, "affordability" is defined as housing costs that consume no more than 30 to 40 percent of household income. A basic socially acceptable standard housing unit is defined by a particular community's view of what is required for decent living and this varies by city. How much floor space is required in a standard unit reflecting consumer choices, market conditions, and regulatory constraints. The definition should also include minimum standards for basic amenities (running water, a toilet) as well as access to essential social services such as schools

[3] Housing costs are a substantial component of household expenditures. Those who allocate a large proportion of their income to housing often have to make difficult financial decisions with significant short-term and long-term implications on their live.

[4] Monthly housing costs sustained by owners include the following components: mortgage principal repayment, mortgage interest payments (net of any tax relief), gross of housing benefits, (i.e., housing benefits should not be deducted from the total housing cost), structural insurance, mandatory services and charges (sewage removal, refuse removal, etc.), regular maintenance and repairs, taxes, and the cost of utilities (water, electricity, gas and heating). Monthly housing costs sustained by renters include the following components: rent payments, gross of housing benefits (i.e., housing benefits should not be deducted from the total housing cost), structural insurance (if paid by the tenants), services and charges (sewage removal, refuse removal, etc.) (if paid by the tenants), taxes on dwelling (if applicable), regular maintenance and repairs and the cost of utilities (water, electricity, gas and heating).

and health clinics. An acceptable housing unit should also place workers no more than an hour's commute from centres of employment (Woetzel et al., 2014).

The definition of housing affordability must be based on a concrete concept of housing accessibility. This is given by the relationship between two subjects: people on the one side (specifically their incomes) and their dwelling on the other side (housing expenses and costs associated with housing). This relationship may be mathematically modelled as a ratio or a difference, which forms the formal basis of the predominant paradigm of housing affordability. In practice, there is a broad range of approaches to defining housing affordability and housing unaffordability. A relative approach is primarily used in the real estate market and is based on prototypical housing costs. This permits comparison of two or more periods, considering whether the flats sold became relatively more or less affordable (Stone, Burke, Ralston, 2011; Jewkes and Delgadillo, 2010). The subjective concept is based on the assumption that households make a choice which is the best one given their financial limitations. From this perspective, housing affordability itself does not have any importance; it is not rationally possible or socially acceptable to define a standardised level of affordability that is other than a personal choice. The ratio approach uses the ratio of housing costs to household income (Norazmawati, 2015; Stone, Burke, Ralston, 2011). Its starting point was conception based on the family budget in which household incomes were evaluated, whether they are sufficient to support all basic household expenditures, including housing costs. In the residual concept, there considered whether households after covering their total housing expenses had sufficient income for paying other expenditures (Stone, Burke, Ralston, 2011).

One of the first definitions of housing affordability is presented by Howenstine (1983, p. 20): "The ability of the household to acquire decent accommodation by the payment of a reasonable amount of its income on shelter". The terms "the decent accommodation" and "the reasonable amount of household income" were not more concrete specified.

MacLennan and Williams (1990, p. 9) clarify the meaning of a reasonable amount of income. In a frequently cited definition of housing affordability, they defined housing affordability as: "Affordability is concerned with securing some given standard of housing (or different standards) at a price or rent which does not impose, in the eye of some third party (usually government) as an unreasonable burden on household incomes". Wong et al. (2010) and Sendi (2011) consider as a lack of this definition, the absence of identification of the term "the unreasonable burden" which would be necessary to explain its accurate and useful content.

A more precise definition explaining the unreasonable burden of a household's income is provided by Bramley and Karley (2005). They mentioned that: "Household should be able to occupy housing that meets well-established (social sector) norms of adequacy (given household type and size) at a net rent which leaves them enough income to live on without falling below some poverty standard" (per Lau, 2001, p. 1).

Another definition provides a description, how to quantify the housing affordability. "The comparison relationship between the housing expenditure (rent, mortgage) and household income is the most common way to define and measure the housing affordability" (Whitehead, 1991).

At the international level, two approaches are used to measuring the housing affordability: the ratio approach or indicator approach and the residual approach (Mulliner, 2012). Other sources also mention the reference approach (Lux et al., 2002, p. 14.).

The ratio or indicator approach is primarily applied in Australia and in international comparisons (Chaplin and Freeman, 1999). This approach is based on the calculation of the portion of income used to cover housing-related costs (the ratio method). Spending over a specific limit, is considered as the housing burden of households and on this basis, this is also used to calculate the housing burden rate. These ratios, therefore, address the question of whether households spend an unreasonably large proportion of their income on housing. While such approaches have been modified and adapted

to a variety of contexts and for specific political purposes, they may be grouped into three general types (Burke et al., 2005, p. 22):

- simple "housing cost to income" ratio,
- fixed ratio with a benchmark,
- refined ratio measures.

The residual approach analyses the amount of the specific portion of income remaining after payment of all housing-related costs (Lux et al., 2002, p. 14). The reference approach does not use any limit for defining when the housing is endangered but reflects on the situation in another sector of housing or the need to secure housing for the concrete selected group of population (Lux et al., 2002, p. 14).

With a focus on the North American usage, Hulchanski (1995) identifies six elements of measuring the housing expenditure to income ratio to measure housing affordability: description of household expenditures, analysis of trends, administration of public housing by defining eligibility criteria and subsidy levels, definition of housing need for public policy purposes, prediction of the ability of a household to pay the rent or the mortgage and as part of the selection criteria in the decision to rent or provide a mortgage. Each of the six uses is assessed based on the extent to which it is a valid and reliable measure of what it purports to measure.

Well known and practiced measurement of affordable housing is that housing costs should be less than 30% of household income (in the United States, Australia and Canada) of the occupants in the bottom 40% of household incomes. Those families who pay more than 30% (40%) of their income for housing are considered cost burdened, and may have difficulty affording necessities such as food, clothing, transportation and medical care (Gabriel et al., 2005).

Therefore, in this broad definition, affordable housing means any housing costing less than 30% of household income of the bottom 40% of the community. Nevertheless, this definition is far from being universally accepted, and poses questions on which costs should be included (such as for instance whether to consider utilities bills) (Pittini, 2012).

According to Eurostat's definition, a household is considered "overburdened" when the total housing costs ("net" of housing allowances) represent more than 40% of disposable income ("net" of housing allowances), where housing costs include mortgage or housing loans interest payments for owners and rent payments for tenants. Utilities (water, electricity, gas and heating) and any costs related to regular maintenance and structural insurance are likewise included (Eurostat, 2009).

The household cost burden (HCB) is defined as the ratio between the monthly total housing costs (HH070) multiplied by 12 and diminished by gross housing allowances (HY070G), and the annual disposable income (HY020) diminished by gross housing allowances following the formula (in percentage after multiplying by 100):

$$HCB = \frac{HH070 \cdot 12 - HY070G}{HY020 - HY070G} \cdot 100 \tag{1}$$

Household cost burden has to be calculated by an individual of the population or a subset of the population, and not by household. Individual weights are therefore used and are based on the Adjusted Cross-Sectional Weight (RB050a) (Eurostat, 2009).

One critique of housing cost burden as a standard of housing affordability is that it does not differentiate between those who have sufficient income to meet household needs after shelter expenditures and those who do not (Stone, 2006). Another critique is that spending a large proportion of income on housing does not necessarily reflect a housing affordability problem. For higher-income households, spending thirty percent of income on housing may be a deliberate decision based on preferences for more spacious and higher-quality housing (Kutty, 2005). On the other hand, for lower-income households, spending thirty

percent or more of income on housing likely represents an involuntary allocation of what are already limited economic resources (McConnell, 2013).

Within the international comparisons, indicator of housing cost overburden rate HH_OVERBURDEN (*housing cost overburden rate*) is used, which indicates percentage of the population living in household, where total housing costs (net of housing allowances) represent more than 40% of the total disposable household income (net of housing allowances).

$$HH\_OVERBURDEN = \frac{\sum_{\forall\ i\ in\ relevant\ breakdown\ with\ HCB>40\%} RB050_i}{\sum_{\forall\ i\ in\ the\ same\ breakdown} RB050_i} \cdot 100. \quad (2)$$

In calculation of HCB are used data related to the statistical unit – household. The housing cost overburden rate indicator is calculated on the level of individual person. Therefore, the personal cross sectional weights are used in its calculation RB050.

The housing costs accounted for 22% of disposable household income in the whole EU in 2016. Households in Greece (nearly 42%) and in Bulgaria (approximately 29%) used the largest amount of disposable income to cover their housing costs. Households in Cyprus (12.8%) and Malta (approximately 7.6%) used the lowest amount of their disposable income on housing. The average housing costs for the Slovak population amount to 21% of disposable income, this percentage increases on average to 36.7% if we look at people at risk of poverty (i.e. those with an equalised disposable income below 60% of the national median equalised disposable income).

One factor linked with housing affordability is the stage in the course of life (McConnell, 2013). Persons in later stages of the life course, such as households headed by older persons and married couples versus households headed by younger people or of other marital statuses, tend to allocate a lower proportion of income to housing (DeVaney et al., 2004; McConnell, 2013) and are more likely to be cost burdened than those without children (Elmelech, 2004). Households with one adult, either living alone or single parents with dependent children, spend the largest amount of their disposable income on housing. Households with a single adult below the age of 65 have housing costs that are 12.9 percentage points higher than the cost level for the general Slovak population. In households with two adults, the share of housing costs is higher than the Slovak average (1.6 percentage points higher for households with two adults with two dependent children and 2.1 percentage points higher than the Slovak average for households with two adults and three dependent children). The share of this type of expenditure in terms of disposable income is lower, probably thanks to the income of an additional adult, in households in which three adults live. The distribution of population by type of household in which they live and based on the housing cost overburden rate shows that the greatest share of persons living in households in which the housing costs burden exceeded 40 percent of their disposable income, were in households with 1 parent and with 1 or more dependent children (29.89%), in single-person households (25.76%) and in households with 2 adults and 1 dependent child (11.89%) (Table 1).

Table 1 Distribution of the population by household type and housing cost burden (HCB) (in %), Slovak Republic, 2016

HCB	Household type								
Col Pct	10	11	12	13	5	6	7	8	9
≤ 40%	88.11	92.34	91.72	96.50	74.24	91.14	96.88	98.04	70.11
> 40%	11.89	7.66	8.28	3.50	25.76	8.86	3.12	1.96	29.89

Notes: 10 – two adults with one dependent child, 11 – two adults with two dependent children, 12 – two adults with three or more dependent children, 13 – households with dependent children, 5 – single person, 6 – two adults younger than 65 years, 7 – two adults, at least one aged 65 years or over, 8 – households without dependent children, 9 – single person with dependent children.
Source: Own processing in SAS Enterprise Guide

The housing affordability is influenced by the tenure status. In EU-28, the proportion of the population whose housing costs exceeded 40% of their disposable income was highest for tenants with market price rents (28.0%) and lowest for persons in owner-occupied dwellings with a loan or mortgage (5.4%). In Slovakia, the housing cost overburden rate for persons in owner-occupied dwellings with a loan or mortgage is 14.48%, and for tenants with market price rents 12.33% (Table 2).

Table 2 Distribution of population by tenure status and housing cost burden (HCB) (in %), Slovak Republic, 2016

HCB	Tenure status			
Col Pct	outright owner	owner paying mortgage	tenant/ subtenant paying rent at prevailing or market rate	accommodation is rented at a reduced rate or accommodation is provided free
≤ 40%	93.85	85.52	87.67	81.99
> 40%	6.15	14.48	12.33	18.01

Source: Own processing in SAS Enterprise Guide

2 ANALYTIC APPROACH

2.1 Database

Research on factors affecting affordability are mostly focusing on rent, income and housing related cost (Howenstine, 1983; Maclennan and Williams 1990; Hancock, 1993). However, other factors are almost being ignored. For example, other non-monetary factors are also playing an important role to determine one's affordability. Without the critical investigation of these other factors, the complete picture of household affordability cannot be shown and needs to be further analysed (Wong et al., 2010).

A system of criteria influencing housing affordability was identified via an extensive review of relevant housing literature. The authors postulate that housing affordability assessment must take a broader view of the wide-ranging criteria that affect households. However, the choice of variables in a multivariable analysis is limited, e.g. by the database used.

Factors of housing unaffordability were analyzed by Jing Li through revision surveys of 112 journal papers over the period from 1990 to 2013. According to Bogdon and Can's research from 1997, he considered for assessment of housing needs three dimensions: amenity, overcrowding, and affordability. The first two "are more prevalent in less developed economies where there is little land for accommodation" (Li, 2014). In this study, he concluded that: "the problem of housing affordability is associated with multi-faceted economic, social, political and demographic considerations". He related "deteriorating housing affordability with low incomes, younger households, elderly and singles" in developed economies with slower GDP growth. His summary of used keywords in housing affordability research over the last two decades contains, for example, homeownership, housing poverty, housing tenure, and demographic factors. Finally, he proposed six major components for a model of affordability: "house price, household formation, housing tenure, migration, demography, and labor".

Lux (2012) concluded that „the structure of the housing market, as measured through housing tenure and partially regionally-based differences in affordability, does influence how workers evaluate participation in the labor market". The author states a decisive effect of housing affordability on the level of structural unemployment and he warns about „the dynamic impact of regional differences in housing affordability on labor mobility concentrated within the most highly skilled segment of the labor force" (Lux, 2012). He also testified that „this relationship was stronger for the house price-to-income ratio than for the rent-to-income ratio. An examination of partial correlation coefficients confirmed the statistical significance of this relationship were control was made for other potentially important confounding variables:

interregional differences in per capita GDP, per capita disposable income, key demographic differences, unemployment rate, and average salary."

According to Nickell, there "was a statistically significant positive correlation between the share of owner-occupied housing and the level of unemployment" across 20 OECD countries during the 1989–1984 period" (Nickel, 1998).

The aim of this paper is to analyse the influence of individual and household characteristics on household cost burden (HCB). The primary objective of this paper was to define the relevant factors affecting the household cost burden in the Slovak Republic and quantifying the intensity of their influence. For this purpose, a logistic regression model and a classification tree model were created, using the sample of the cross-component of the data of the statistical survey EU SILC 2016. The analysis were completed using SAS Enterprise Guide (SAS EG) and SAS Enterprise Miner (SAS EM).

The analysis was carried out using an individual-level data extracted from EU SILC 2016 cross-sectional component provided by the Statistical Office of the Slovak Republic (EU SILC 2016, UDB 27/04/2017). Four types of data sets were used in analysis: Register of persons (R_SILC_2016), Personal data (P_SILC_2016), Household register (D_SILC_2016) and Household data (H_SILC_2016). The combination of all four data sets through the identification numbers of persons and identification numbers of households resulted in a dataset composed of 14,101 records of respondents aged 16 and over.

The household cost burden was calculated by the Formula (1). For the purposes of modelling, its values were substituted with 0 (if HCB < 40%) and 1 (if HCB ≥ 40%). Input variables described the basic characteristics of persons over the age of 16 and the characteristics of the household in which they live: At risk of poverty or social exclusion (AROPE), Household type, Self-defined current economic status (EA_SELF), People living in households with very low work intensity, Tenure status (TENURE_STAT), Dwelling type (DW_T), Region, Sex, NUTS 3 Region, Degree of urbanisation. The description of the input variables is captured in Table 3.

Table 3 Distribution of population by tenure status and housing cost burden (HCB) (in %), Slovak Republic, 2016

Original variables – description	Categories
At risk of poverty or social exclusion *	ARPT60i = 0; SEV_DEP = 0; LWI = 0;
	ARPT60i = 1; SEV_DEP = 0; LWI = 0;
	ARPT60i = 1; SEV_DEP = 1; LWI = 0;
	ARPT60i = 1; SEV_DEP = 0; LWI = 1;
	ARPT60i = 1; SEV_DEP = 1; LWI = 1;
	ARPT60i = 0; SEV_DEP = 1; LWI = 0;
	ARPT60i = 0; SEV_DEP = 0; LWI = 1;
	ARPT60i = 0; SEV_DEP = 1; LWI = 1;
Household type	single person
	two adults younger than 65 years
	two adults, at least one aged 65 years or over
	households without dependent children

Table 3 (continuation)

Original variables – description	Categories
Household type	single person with dependent children
	two adults with one dependent child
	two adults with two dependent children
	two adults with three or more dependent children
	other households with dependent children
Self-defined current economic status	employee working full-time
	employee working part-time
	self-employed working full-time (including family worker)
	self-employed working part-time (including family worker)
	unemployed
	pupil, student, further training, unpaid work experience
	in retirement or in early retirement or has given up business
	permanently disabled or/and unfit to work
	fulfilling domestic tasks and care responsibilities
	other inactive person
Tenure status	outright owner
	owner paying mortgage
	tenant or subtenant paying rent at prevailing or market rate
	accommodation is rented at a reduced rate or accommodation is provided free
Dwelling type	detached house
	semi-detached or terraced house
	apartment or flat in a building with less than 10 dwellings
	apartment or flat in a building with 10 or more dwellings
	some other kind of accommodation
Region	Bratislava Region
	Western Slovakia Region
	Central Slovakia Region

Table 3 (continuation)

Original variables – description	Categories
Region	Eastern Slovakia Region
Sex	man
	woman
NUTS3 Region	Bratislava Region
	Trnava Region
	Trenčín Region
	Nitra Region
	Žilina Region
	Banská Bystrica Region
	Prešov Region
	Košice Region
Degree of urbanisation	densely populated area
	intermediate area
	thinly populated area

Notes: * ARPT60i = 1: person with disposable income below at-risk-of-poverty threshold (ARPT60i = 0: person with disposable income above at-risk-of-poverty threshold); SEV_DEP = 1: person is affected by severe material deprivation (SEV_DEP = 0: person is not affected by severe material deprivation); LWI = 1: person lives in households with very low work intensity (LWI = 0: person does not live in households with very low work intensity).

Source: Methodological guidelines and description of the EU-SILC target variables, own processing

2.2 Logistic regression model

A logistic regression model is a special instance of a generalised linear model. It may be used to explain (dependent) variables with other than normal distribution of probability (binomial, Poisson, exponential, gamma distribution, ...). The selection of our model for analysis was conditioned by the fact that variable of the household cost burden had a binomial distribution of probability.

The logistic regression model may be used to estimate the conditional mean value of a dependent variable $E(Y|x_i) = \pi$ (the conditional probability that a dependent variable will have a value of 1):

$$\pi = \frac{exp\ (\beta_0 + \sum_{j=1}^{k} \beta_j\, x_{ij})}{1+exp\ (\beta_0 + \sum_{j=1}^{k} \beta_j\, x_{ij})}, \tag{3}$$

where x_j (j = 1, 2, …k) are the input variables, β_0 and β_j (j = 1, 2, …k) are the unknown parameters of model. More often, the model is presented in a form used to record the generalised linear model and which expresses the relationship between function of the conditioned mean value of the dependent variable π (in the logit model it is logit: $\ln \frac{\pi}{1-\pi}$ and the linear combination of the independent

variables: $\ln \frac{\pi}{1-\pi} = \beta_0 + \sum_{j=1}^{k} \beta_j x_j = \sum_{j=0}^{k} \beta_j x_j$.

The odds $\frac{\pi}{1-\pi}$ is the probability that the observed event occurs (a person lives in a household where the total housing costs are more than 40% of the total disposable household income) and the probability that the observed event does not occur (a person lives in a household where the total housing costs do not exceed 40% of the total disposable household income).

The Odds Ratio $OR = \frac{odds_1}{odds_2}$ is used to interpret the parameters of the logistic regression model where $odds_1$ indicates the odds that the given event occurs for the first object of comparison and $odds_2$ is the odds that the given event occurs for the second object of comparison (Hosmer and Lemeshow, 2000; Agresti, 1990).

2.3 Decision tree model

In addition to the logistic regression model, the decision tree (classification tree) model was used in analyses. Classification and regression trees are suited for the analysis of complex data. Decision tree models can be effectively used to determine the most important attributes in a dataset (Breiman, 2001).

A decision tree is a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. A decision tree model consists of a set of rules for dividing a large heterogeneous population into smaller, more homogeneous groups with respect to a particular target variable (Berry and Linoff, 2004; Dietterich, 1990).

Ideally, subsets at the end of the branching process, i.e. leaves, should contain only one class (category) of the specified dependent variable. In the case of a decision tree applied to data sourced from the EU SILC survey in which the inhabitants of Slovakia were the objects of investigation, the branches end at leafs, in which the predominant group were people living in households, where the housing cost burden exceeded the 40% of the disposable household income, or the category of persons for whose the housing cost burden was below threshold. The relative frequencies of categories of the explained variables influence the cleanliness of the individual nodes of leaves that can be measured by entropy:

$$H(Y) = \sum_{j=1}^{m} \frac{n_j}{n} log_2 \frac{n_j}{n}, \tag{4}$$

where n_j is the frequency of the class y_j (in our case is the size of the class of persons burdened with the housing costs and the class whose the housing costs exceed the threshold of 40% of the disposable income). In the case of a binary dependent variable, entropy acquires a maximum value of 1 (if both classes have the same frequencies) and a minimum value of 0 (if the set contains only one class).

The created decision tree was not used as a predictive model, we used its ability to classify individual cases (persons) into two classes, according to whether their housing costs could be considered as burdensome or not.

3 RESULTS

3.1 Results of the logistic regression model

For our analysis we used PROC LOGISTICS, that is the most popular SAS procedure for doing Maximum Likelihood Estimation of the logistic regression model (Allison, 2012). The results of the logistic regression analysis are presented in Tables 4–7.

The table "Model Fit Statistics" (Table 4) reports three different model fit statistics: the Akaike's Information Criterion (AIC), the Schwarz Criterion (SC) and the maximized value of the logarithm of the likelihood function multiplied by –2 (–2Log *L*). Values of these fit statistics are displayed for two different models, a model with an intercept but no predictors (covariates), and a model that includes all the specified predictors. Higher values of –2Log *L* mean a worse fit to the data. The problem with –2Log *L* is that models with more predictors tend to fit better by chance alone. The other two fit statistics avoid this problem by penalizing model that have more covariates (Allison, 2012).

Table 4 Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	6 808.464	4 603.882
SC	6 816.018	4 868.272
–2 Log L	6 806.464	4 533.882

Source: Own processing in SAS Enterprise Guide

Table 5 is "Global Zero Hypothesis Testing: Beta = 0". In this table there are three statistics with values of 2272.5814, 3100.3644 and 1502.5312. All three statistics test for the same null hypothesis: that all explanatory variables have a coefficient of 0. The associated p-values are less than 0.01, so we can reject the null hypothesis and conclude that at least one of the coefficients is not 0.

Table 5 Testing Global Null Hypothesis: BETA = 0

Test	Chi-Square	DF	p-value
Likelihood Ratio	2 272.5814	34	<.0001
Score	3 100.3644	34	<.0001
Wald	1 502.5312	34	<.0001

Source: Own processing in SAS Enterprise Guide

From the original set of fourteen input variables, only those with a statistically significant influence on the variable HCB were selected: self-defined current economic status, household type, region, dwelling type, tenure status and at risk of poverty or social exclusion (Table 6).

In the Table 7 there are coefficient estimates, their estimated standard errors, and test-statistics for the null hypothesis that each coefficient is equal to 0. Since categorical variables were involved in the model, they were replaced by dummy (indicator) variables in the model. We inserted the odds ratios (odds) for household cost burden derived from binary logistic regression in association with the 34 indicators (dummy variables) in the Table 7 too. The point estimates of the odds ratios are used to interpret the values of the estimated model parameters.

The dummy indicators created by using their self-defined current economic status had higher odds that their housing costs exceeded 40 percent of the disposable income threshold when compared to the reference category of disabled persons or persons unable to work (permanently disabled persons or persons unfit to work). The odds were up to 6.854 times higher for the self-employed working part-time

Table 6 Testing Global Null Hypothesis: BETA = 0

Effect	DF	Wald Chi-Square	p-value
Self-defined current economic status	9	75.8949	<.0001
Household type	8	580.1947	<.0001
Region	3	27.9035	<.0001
Dwelling type	4	14.0034	0.0073
Tenure status	3	151.1032	<.0001
At risk of poverty or social exclusion	7	839.8892	<.0001

Source: Own processing in SAS Enterprise Guide

Table 7 Analysis of Maximum Likelihood Estimates and Odds Ratio Estimates

	Odds Ratio Estimates				
Effect	**Parameter Estimate**	**Odds Ratio Estimate**	**Standard Error**	**Wald Chi-Square**	**p-value**
Self-defined current economic status					
Employee working full-time	0.2342	1.264	0.2640	0.7870	0.3750
Fulfilling domestic tasks and care responsibilities	1.1184	3.060	0.4512	6.1438	0.0132
Other inactive person	0.7239	2.062	0.3160	5.2477	0.0220
Employee working part-time	0.9873	2.684	0.3848	6.5825	0.0103
Self-employed working full-time	1.2360	3.442	0.2830	19.0725	<.0001
Self-employed working part-time	1.9248	6.854	0.6234	9.5329	0.0020
Unemployed	1.1153	3.051	0.2614	18.2006	<.0001
Pupil, student	0.4800	1.616	0.2866	2.8054	0.0939
In retirement or in early retirement or has given up business	0.8023	2.231	0.2679	8.9700	0.0027
Permanently disabled or/and unfit to work	Reference category				
Household type					
Two adults with one dependent child	–0.5080	0.602	0.2253	5.0862	0.0241
Two adults with two dependent children	–1.2673	0.282	0.2272	31.1175	<.0001
Two adults with three or more dependent children	–2.1108	0.121	0.2709	60.7045	<.0001
Households with dependent children	–2.3383	0.096	0.2320	101.6072	<.0001
Single person	0.9970	2.710	0.2275	19.2090	<.0001

Table 7 (continuation)

Odds Ratio Estimates					
Effect	**Parameter Estimate**	**Odds Ratio Estimate**	**Standard Error**	**Wald Chi-Square**	**p-value**
Permanently disabled or/and unfit to work	Reference category				
Household type					
Two adults younger than 65 years	–0.4863	0.615	0.2309	4.4342	0.0352
Two adults, at least one aged 65 years or over	–1.0158	0.362	0.2640	14.8029	0.0001
Households without dependent children	–1.9340	0.145	0.2455	62.0709	<.0001
Single person with dependent children	Reference category				
Region					
Bratislava Region	0.5421	1.720	0.1361	15.8696	<.0001
Western Slovakia Region	0.3562	1.428	0.1088	10.7139	0.0011
Central Slovakia Region	–0.0391	0.962	0.1175	0.1105	0.7395
Eastern Slovakia Region	Reference category				
Household type					
Detached house	0.1013	1.107	0.0942	1.1576	0.2820
Semi-detached or terraced house	0.7172	2.049	0.2839	6.3800	0.0115
Apartment or flat in a building with less than 10 dwellings	0.1840	1.202	0.1543	1.4228	0.2329
Some other kind of accommodation	1.1253	3.081	0.4106	7.5123	0.0061
Apartment or flat in a building with 10 or more dwellings	Reference category				
Tenure status					
Outright owner	0.6920	1.998	0.3069	5.0846	0.0241
Owner paying mortgage	2.1942	8.973	0.3297	44.2985	<.0001
Tenant or subtenant paying rent at prevailing or market rate	1.4653	4.329	0.3252	20.3051	<.0001
Accommodation is rented at a reduced rate or accommodation is provided free	Reference category				
AROPE					
ARPT60i = 0; SEV_DEP = 0; LWI = 0	–3.1266	0.044	0.2044	234.0613	<.0001
ARPT60i = 0; SEV_DEP = 0; LWI = 1	–2.6165	0.073	0.4005	42.6785	<.0001
ARPT60i = 0; SEV_DEP = 1; LWI = 0	–2.5514	0.078	0.2630	94.1201	<.0001
ARPT60i = 1; SEV_DEP = 0; LWI = 0	–0.0208	0.979	0.1994	0.0109	0.9169

Table 7 (continuation)

Odds Ratio Estimates					
Effect	**Parameter Estimate**	**Odds Ratio Estimate**	**Standard Error**	**Wald Chi-Square**	**p-value**
Accommodation is rented at a reduced rate or accommodation is provided free	Reference category				
AROPE					
ARPT60i = 1; SEV_DEP = 0; LWI = 1	0.3753	1.455	0.2371	2.5062	0.1134
ARPT60i = 0; SEV_DEP = 1; LWI = 1	–2.5011	0.082	0.6549	14.5859	0.0001
ARPT60i = 1; SEV_DEP = 1; LWI = 0	–0.9246	0.397	0.2701	11.7141	0.0006
ARPT60i = 1; SEV_DEP = 0; LWI = 0	–0.0208	0.979	0.1994	0.0109	0.9169
ARPT60i = 1; SEV_DEP = 1; LWI = 1	Reference category				

Source: Own processing in SAS Enterprise Guide

and 3.06 times higher for the persons in the category of persons fulfilling domestic tasks and care responsibilities than the odds were for the reference category. Given the type of household in which a person lives, the highest housing cost burden is faced by persons living in households with 1 adult and at least 1 dependent child (reference category). For nearly all other considered types of households, the odds that the HCB variable exceeded the 40% of available income threshold were lower. Only single-member households had greater odds that the housing costs would be a burden for some of them.

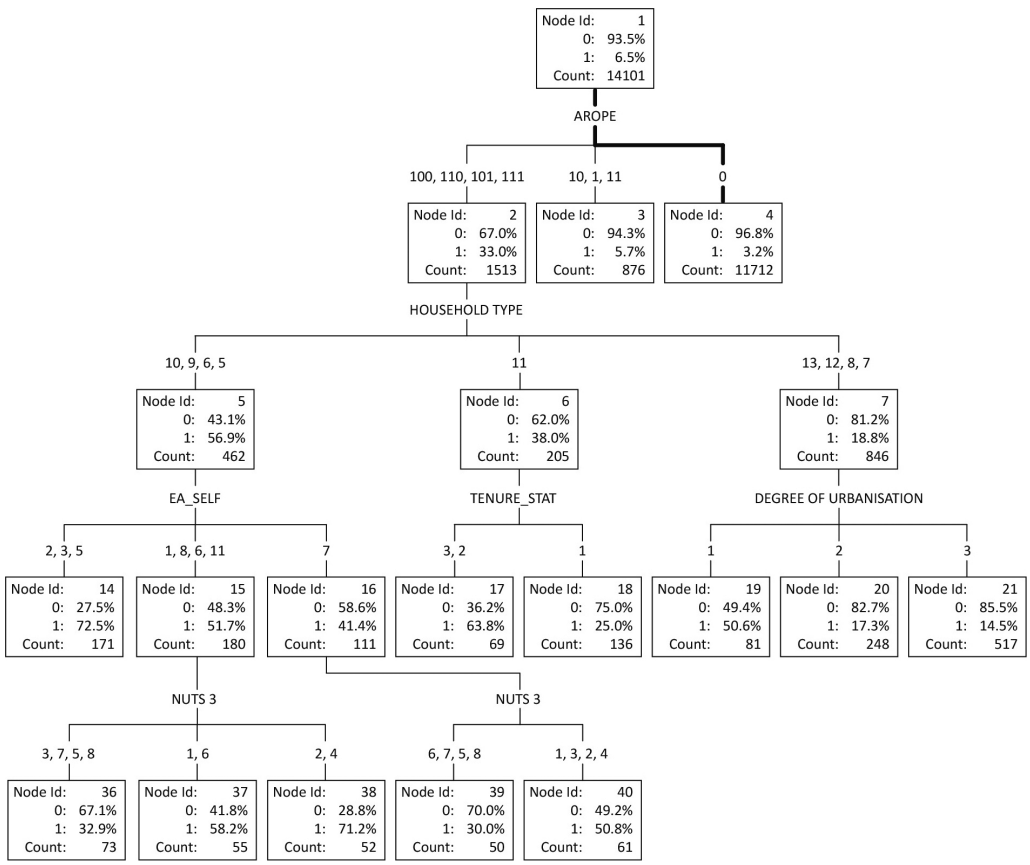
We discovered the statistically significant differences in the housing cost burden in comparison between the Bratislava Region, the Western Slovakia Region, in the Central Slovakia Region, and the Eastern Slovakia Region. The odds, that the housing cost burden exceeded the 40% of disposable household income threshold, are 1.72 times higher for inhabitants living in the Bratislava Region and 1.482 times higher than those living in the Western Slovakia Region in comparison to inhabitants living in the Eastern Slovakia Region. There exists a significant difference in the odds of housing cost burden between categories of population created by the type of dwelling and the odds for inhabitants falling in this reference category of variable. Persons living in apartment or flat in a building with 10 or more dwellings have odds that their housing costs exceed the 40 percent of disposable income and this odds are lower than for other groups categorised by their type of dwelling. The housing costs represent a significant burden for the owners paying the mortgage. Their odds that they spend more than 40% of their disposable income on housing are up to 8.973 times higher than those who rent their housing at a lower price (lower than the market price) or who have the housing free-of-charge. The odds that the housing costs represent a burden are up to four times higher for those who rent their housing at the market price compared with inhabitants who live in social housing (with a reduced rent) or have the housing free-of-charge.

There are also differences in the housing cost burden among the groups formed using the AROPE variable. The category of persons who are currently at risk of poverty, are severely materially deprived and living in households in which the work intensity is defined as low (reference category 111) have odds that their housing costs exceed 40% of their disposable household income, that are higher than the odds for other classes created by using the AROPE category of variables. The only exception are persons who are at risk of poverty but without the risk of severe material deprivation and are living in households with very low work intensity. The odds of this category were 1.455 times higher in comparison with the reference category.

3.2 Results of the decision tree model

The algorithm used in the generation of the decision tree applied a maximum of triple branching of nodes, the growth of the tree was limited by defining its maximum depth[5] (Max Depth = 5 controls the maximum depth of the tree that will be created. The root node is considered to have a depth of 0.) and the selection of branching variables was completed using the values of expected entropy. From the set of potential decision trees that were created in SAS EM, the one with the lowest misclassification rate was selected (details in Neville and de Ville, 213). Decision tree identified the most significant variables (AROPE, household type (HT), self-defined current economic status (EA_SELF), tenure status (TENURE_STAT), NUTS 3, degree of urbanisation) and their values that give the best homogeneous sets of the population. It chose the split which has the lowest entropy compared to the parent node and other splits. The tree structure of the tree contains a total of 13 leaves. Each of them provides information about the relative magnitude of the classes of persons with a housing cost burden (Figure 1). These may then be used to estimate the probability of their occurrence. The properties of persons are contained in the decision-making rules in the individual leaves.

Figure 1 Decision Tree Diagram



Source: Own processing in SAS Enterprise Miner

[5] Max Depth controls the maximum depth of the tree that will be created. The root node is considered to have a depth of 0.

From our perspective, those sets of persons for whom the probability of the housing cost burden is very high or very low were of interest (Figure 2, Figure 3).

Figure 2 Decision-making rules for leafs with the lowest number of people burdened with

Node	Depth	Observations	Percent 1
4	1	11712	0.03
3	1	876	0.06
21	3	602	0.14
14	3	171	0.73
18	3	136	0.25
43	4	119	0.24
37	4	96	0.61
36	4	84	0.4
17	3	69	0.64
42	4	68	0.53
38	4	61	0.51
20	3	57	0.19
39	4	50	0.3

if **AROPE** IS ONE OF: 0 or MISSING
then Tree Node Identifier = 4
Number of Observations = 11 712
Predicted: HH_OV=0 = 0.97
Predicted: HH_OV=1 = **0.03**

if **AROPE** IS ONE OF: 10, 1, 11
then Tree Node Identifier = 3
Number of Observations = 876
Predicted: HH_OV=0 = 0.94
Predicted: HH_OV=1 = **0.06**

Source: Own processing in SAS Enterprise Miner

The lowest number of people burdened with the housing costs (0.03) is in the node that included persons who are not at risk of poverty, severely materially deprived and living in households whose work intensity is not defined as low. The next node with the relatively lowest number of people burdened with the housing costs (0.06) included those who are not at risk of poverty but who are severely materially deprived (AROPE = 10), or living in households whose work intensity is defined as low (AROPE = 1), or who are both severely materially deprived and are living in households with the low work intensity (AROPE = 11) (Figure 2).

The highest share of persons burdened with the housing costs (0.73) is in the group of people living in any of the following types of household: households with 2 adults and 1 dependent child (HT = 10), households with 1 adult and with 1 or more dependent children (HT = 9), households with 2 adults and without dependent children, both under the age of 65 (HT = 6) or single-member households (HT = 5), are employees with the shortened working hours (EA_SELF = 2), or full-time entrepreneurs and self-employed persons (EA_SELF = 3), or unemployed persons (EA_SELF = 5) and who are at risk of poverty according to AROPE (AROPE = 100) and who are concurrently either severely materially deprived (AROPE = 110) or living in households whose work intensity is defined as low (AROPE = 101), or who are concurrently severely materially deprived and living in households with low work intensity (AROPE = 111) (Figure 3).

The second group in order with the highest share of persons burdened by the housing costs (0.64) includes persons whose AROPE indicator of poverty or social inclusion are the same as the previous group. Their another common characteristic is that they live in households with 2 adults and 2 dependent children (HT = 11) and are owners of a flat and repaying a mortgage (TENURE_STATUS = 2) or tenants or sub-lessees who are paying rent or a sub-lease (TENURE_STATUS = 3) (Figure 3).

Figure 3 Decision-making rules for leafs with the highest number of people burdened with housing costs

Node	Depth	Observations	Percent 1
4	1	11712	0.03
3	1	876	0.06
21	3	602	0.14
14	3	171	0.73
18	3	136	0.25
43	4	119	0.24
37	4	96	0.61
36	4	84	0.4
17	3	69	0.64
42	4	68	0.53
38	4	61	0.51
20	3	57	0.19
39	4	50	0.3

if **HT** IS ONE OF: 10, 9, 6, 5
AND **EA_SELF** IS ONE OF: 2, 3, 5
AND **AROPE** IS ONE OF: 100, 110, 101, 111
then Tree Node Identifier = 14
Number of Observations = 171
Predicted: HH_OV=0 = 0.27
Predicted: HH_OV=1 = **0.73**

if **TENURE_STATUS** IS ONE OF: 3, 2
AND **HOUSEHOLD TYPE** IS ONE OF: 11
AND **AROPE** IS ONE OF: 100, 110, 101, 111
then Tree Node Identifier = 17
Number of Observations = 69
Predicted: HH_OV=0 = 0.36
Predicted: HH_OV=1 = **0.64**

Source: Own processing in SAS Enterprise Miner

CONCLUSIONS

The objective of our paper was to identify the factors that have a statistically significant effect on the housing cost burden on the Slovak population in 2016. A logistic model regression was used to identify and quantify the strength of their influence. The variable modelled in the regression model was the household cost burden on housing, which is used by the European Union to measure housing affordability.

By the stepwise method, indicators that had a statistically significant influence on the household burden were selected: self-defined current economic status, type of household, region, type of dwelling, ownership status, and the AROPE variable. The strength of their effects was quantified using Cramer's V (CV) coefficient. Based on its value, it may be said that variable the household cost burden is most strongly influenced by the AROPE (CV = 0.376), type of household (CV = 0.267) and self-defined current economic status (CV = 0.170). Additionally, odds ratios were estimated to facilitate a comparison of the housing cost burden between the individual groups of persons categorized based on their individual characteristics and the typology of the households in which they live.

Decision tree identified the most significant variables: AROPE, household type, self-defined current economic status, tenure status, NUTS 3 and degree of urbanisation. The decision (classification) tree was used as a classifier of persons according to their housing cost burden. It allows for the prediction of the probability of whether a person whose characteristics are expressed using the values of the input variables are burdened by housing costs. The results of the decision tree confirmed that the AROPE variable is the most influential variable, given its ability to differentiate the population according to their housing cost burden.

It is everyone's right to obtain affordable housing, while failure to attain the goal is mainly due to political struggles. Financial deregulation, coupled with an unusual rise in property prices, inappropriately targeted socio-economic, housing policies or fiscal policy together with incompetence to strategically manage affordable housing supply for low-income households with housing access problems raises two general questions: Is the housing affordability problem partly connected to the poverty issue? What other factors also account for housing affordability? Our results partially considered both of these questions. Although examined in the context of the Slovak Republic, a similar analysis of attributes determining housing affordability might be applied in international studies. The results of our analysis of household

patterns related to housing affordability may contribute to some extent to appropriate targeting of the proper regional and state government policies and for defining administrative rules about eligibility for housing programmes.

ACKNOWLEDGMENT

This article is provided as one of the outputs of the research projects: VEGA 1/0770/17: Availability and affordability of housing in Slovakia.

References

AGRESTI, A. *Categorical Data Analysis.* New York: John Wiley & Sons, 1990.

ALLISON, P. D. *Logistic Regression Using SAS: Theory and Application*, 2nd Ed. NC: SAS Institute Inc., 2012.

ATFIELD, M. *Rural Housing Affordability: A Location-Based Investigation of the Characterisitcs of those Experiencing Housing Affordability Problems in Ontario.* Canada: Queen's University Kingston, 2013.

BERRY, M. J. A. AND LINOFF, G. S. *Data mining Techniques. For Marketing, Sales, and Customer Relationship management.* 2nd Ed. Indianapolis, Indiana: Wiley Publishing, Inc., 2004.

BRAMLEY, G. AND KARLEY, N. K. How Much Extra Affordable Housing is Needed in England? *Housing Studies*, 2005, 20(5), pp. 685–715. DOI: 10.1080/02673030500213938.

BREIMAN, L. Random forests. *Machine Learning*, 2001, 45(5), pp. 5–32. DOI: 10.1023/A:1010933404324.

BURKE, T. et al. *Conceptualising and measuring the housing affordability problem. National Research Venture 3: Housing Affordability for Lower Income Australians Research Paper 1* [online]. Melbourne: Australian Housing and Urban Research Institute, 2005. [cit. 12.2.2019] <https://core.ac.uk/download/pdf/43326790. pdf>.

CHAPLIN, R. AND FREEMAN, A. Towards an Accurate Description of Affordability. *Urban Studies*, 1999, 36(11), pp. 1949–1957. DOI: 10.1080/0042098992692.

DEVANEY, S. A. et al. Life Cycle Stage and Housing Cost Burden. *Financial Counseling & Planning* [online]. 2004, 15(1), pp. 31– 9. [cit. 12.1.2019] <https://www.researchgate.net/ publication/2284 61551_Life_cycle_stage_and_housing _cost_burden>.

DIETTERICH, T. G. Machine learning. *Annual Review of Computer Science*, 1990, 4(1), pp. 255–306. DOI: 10.1146/annurev. cs.04.060190. 001351.

ELMELECH, Y. Housing inequality in New York City: Racial and ethnic disparities in homeownership and shelter-cost burden. *Housing, Theory, and Society*, 2004, 21(4), pp. 163–175. DOI: 10.1111/j.1533-8525.2001.tb00028.x.

EP. *Charter of Fundamental Rights of the European Union* [online]. Luxembourg: European Parliament, Office for Official Publications of the European Communities, 2000. [cit. 12.2.2019] <http://www.europarl.europa.eu/charter/pdf/text_en.pdf>.

EU. *Treaty on the Functioning of the European Union* [online]. European Union, 2009. [cit. 8.1.2019] <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A12012E%2FTXT>.

EUROSTAT. *Algorithms to compute social inclusion indicators based on EU-SILC and adopted under the Open Method of Coordination (OMC)* [online]. Luxembourg: Eurostat, 2009. [cit. 12.1.2019] <https://www.dst.dk/ext/747139308/0/ukraine/ENG_Algorithms-to-compute-Social-Inclusion-Indicators-based-on-EU-SILC-and-adopted-under-the-Open-Method-of-Coordination-(OMC)--pdf>.

GABRIEL, M. et al. *Conceptualising and measuring the housing Affordability problem* [online]. Australian Housing and Urban Research Institute, 2005. [cit.10.3.2019] <http://www.ahuri.edu.au/downloads/NRV3/NRV3_Research_Paper_1.pdf.>.

HANCOCK, K. Can Pay? Won't Pay? The Economic Principles of Affordability. *Urban Studies*, 1993, 30(1), pp. 127–145. DOI: 10.1080/00420989320080081.

HEGEDÜS, J., ELSINGA, M., HORVÁTH, V. *Policy Discussion Brief for the European Commission on housing in EU member states. Habitat for Humanity International Europe, Middle East and Africa, June, 2016* [online]. Habitat for Humanity EMEA. [cit. 18.1.2019] <https://www.habitat.org/sites/default/files/EMEA%20Policy%20Brief%20on%20Housing%20in%20EU_24112016.pdf.>.

HOSMER, D. V. AND LEMESHOW, S. *Applied Logistic Regression.* New York: John Wiley & Sons, 2000.

HOWENSTINE, E. J. *Attacking Housing Cost: Foreign Policy and Strategies.* New Jersey: Centre for Urban Policy Research, 1983.

HULCHANSKI, J. D. The concept of Housing affordability: Six contemporary uses of the housing expenditure-to-income ratio. *Housing Studies*, 1995, 10(4), pp. 471–491. DOI: 10.1080/02673039508720833.

JEWKES, M. D. AND DELGADILLO, L. M. Weaknesses of Housing Affordability Indices Used by Practitioners [online]. *Journal of Financial Counseling and Planning*, 2010, 21(1), pp. 43–52. [cit. 25.2.2019] <https://afcpe.org/assets/pdf/volume_21_issue_1/jewkes_delgadillo.pdf.>.

KUTTY, N. A new measure of housing affordability: Estimates and analytical results. *Housing Policy Debate*, 2005, 16(1), pp. 113–142. DOI: 10.1080/10511482.2005. 9521536.

LAU, K. Y. A *Comparison Of Indicators Used In Measuring Housing Affordability In Hong Kong And Their Validity. Working Paper Series 2001 / No. 2* [online]. Department of Public and Social Administration City University of Hong Kong, 2001. [cit. 25.2.2019] <http://www6.cityu.edu.hk/pol/staff/KYLau/wp0102.pdf>.

LI, J. *Recent Trends on Housing Affordability Research: Where are we up to?* [online]. Urban Research Group – CityU on Cities Working Paper series, WP No. 5/2014. [cit. 15.10.2019]. <http://www.cityu.edu.hk/cityuoncities/upload/file/original/70552015012 6143516 .pdf>.

LUX, M. et al. *Bydlení-věc veřejná: sociální aspekty bydlení v České republice a zemích Evropské unie* (in Czech). Prague: Sociologické nakladatelství, 2002.

LUX, M. AND SUNEGA, P. Labour Mobility and Housing: The Impact of Housing Tenure and Housing Affordability on Labour Migration in the Czech Republic. *Urban Studies*, 2012, 49(3), pp. 489–504. DOI: 10.1177/0042098011405693.

MACLENNAN, D. AND WILLIAMS, P. eds. *Affordable Housing in Britain and America.* York: Joseph Rowntree Foundation, 1990.

MCCONNELL, E. D. Who has housing affordability problems? Disparities in Housing Cost burden by Race, Nativity and Legal Status in Los Angeles. *Race and social problems*, 2013, 5(3), pp. 173–190. DOI: 10.1007/s12552-013-9086-x.

MULLINER, E. K. *A model for the complex assessment of sustainable housing affordability: Doctoral thesis* [online]. Liverpool: John Moores University, 2012. [cit. 5.2.2019] <http://researchonline.ljmu.ac.uk/6183/1/589785.pdf.>.

NEVILLE, P. AND DE VILLE, B. *Decision Trees for Analytics Using SAS® Enterprise Miner™*. USA, NC: SAS Institute, 2013.

NICKELL, S. Unemployment: questions and some answers. *Economic Journal*, 1998, 108(448), pp. 802–816. DOI: 10.1111/1468-0297.00316.

NORAZMAWATI, M. S. Price to Income Ratio Approach in Housing Affordability. *Journal of Economics, Business and Management*, 2015, 3(12), pp. 1190–1193. DOI: 10.7763/JOEBM.2015.V3.357.

PITTINI, A. *Housing Affordability in the EU: Current Situation and Recent Trends.* Brussels: CECODHAS European Social Housing Observatory, 2012.

SCANLON, K., ARRIGOITIA, M. F., WHITEHEAD, CH. *Social housing in Europe* [online]. Stockholm: Swedish Institure for European Policy Studies, 2015. [cit. 15.1.2019] < http://eprints.lse.ac.uk/62938/ >.

SENDI, R. *Housing accessibility versus housing affordability: Introducing universal housing care* [online]. Enhr Conference 2011, Toulouse, 5–8 July, 2011. [cit. 5.1.2019] <https://www.enhr.net/documents/2011%20France/WS14/Paper-Richard%20Sendi-WS14.pdf.>.

STONE, M. E. What is Housing Affordability? The Case for the Residual Income Approach. *Housing Policy Debate*, 2006, 17(1), pp. 151–184. DOI: 10.1080/10511482.2006.9521564.

STONE, M. E., BURKE, T., RALSTON, L. *The Residual Income Approach to Housing Affordability: The Theory and the Practice* [online]. AHURI Positioning Paper No. 139, 2011. [cit. 22.2.2019] <http://works.bepress.com/michael_stone/7>.

UN. *Convention on the Elimination of All Forms of Discrimination Against Women* [online]. New York: United Nations, 1979. [cit. 8.1.2019] <http://www.un.org/womenwatch/daw/cedaw/text/econvention.htm>.

UN. *Convention on the Rights of the Child* [online]. New York: United Nations, 1989. [cit. 8.1.2019] <https://www.ohchr.org/en/professionalinterest/pages/crc.aspx>.

UN. *International Covenant on Economic, Social and Cultural Rights* [online]. New York: United Nations, 1966. [cit. 8.1.2019] <https://treaties.un.org/doc/treaties/1976/01/19760103%2009-7%20pm/ch_iv_03.pdf.>.

UN. *Universal Declaration of Human Rights* (217 (III) A) [online]. Paris, 1948. [cit. 8.1.2019] <http://www.un.org/en/udhrbook/pdf/udhr_booklet_en_web.pdf.>.

WHITEHEAD, C. From need to affordability: An analysis of UK housing objectives [online]. *Urban Studies*, 1991, 28(6), pp. 871–887. <https://doi.org/10.1080/00420989120081101>.

WOETZEL, J. et al. *A blueprint for addressing the global affordable housing challenge* [online]. McKinsey & Company, 2014. [cit. 25.8.2019] <https://www.mckinsey.com/~/media/McKinsey/Featured%20Insights/Urbanization/Tackling%20the%20worlds%20affordable%20housing%20challenge/MGI_Affordable_housing_Executive%20summary_October%202014.ashx.>

WONG, F. K. W. et al. *Measuring affordability and factors affecting affordability of elderly in Hong Kong. W110-Special Track 18th CIB World Building Congress May 2010 Salford United Kingdom* [online]. UK: CIB Publication, 2010, pp. 1–19. [cit. 21.1.2019] <http://www.irbnet.de/daten/iconda/CIB18947.pdf.>.

Spatial Autocorrelation of a Demographic Phenomenon: a Case of One-Family Households and One-Person Households

Jaroslav Kraus[1] | *Charles University in Prague, Prague, Czech Republic*

Abstract

The paper aims to examine spatial distribution and to subsequently analyse a demographic phenomenon; share of one-family households and one-person households of all households by municipality in the Czech Republic (CR), because these two types of households are in a way contradictory. Although the Czech Republic is rather small and homogeneous with respect to demographic processes, it is questionable whether this also applies to the spatial distribution of households. Methods of local and global spatial autocorrelation represented by Moran's index were used; however, the challenge of normalisation of both variables using Box-Cox transformation had to be addressed before. For comparison classical measures of association on NUTS3 level were used with respect to the type of data being examined – e.g. the Pearson chi-square, the uncertainty coefficient, the lambda coefficient, as well as the Gini index (Gini ratio), which is a measure of statistical dispersion and the most commonly used measurement of inequality.

From the results a small statistically significant global autocorrelation was ascertained. Local results show that shares of both types of households are not a complementary phenomenon in terms of space. From the relationship of local and global Moran's index, it can be concluded that the global rate does not sufficiently depict detected local differences.

Keywords

Population and housing census, households, spatial autocorrelation, Czech Republic

JEL code

C21

INTRODUCTION – SOURCES AND REFERENCES

Spatial information related to demographic processes is an integral part thereof and becomes thus a subject of a demographic analysis. It can be presented clearly in population censuses – the largest demographic survey – in which specifically the analysis for small territorial units such as e.g. municipalities is one of the reasons why the censuses are carried out.

[1] Faculty of Science, Department of Demography and Geodemography, Albertov 6, 128 00 Prague, Czech Republic. E-mail: kraus@natur.cuni.cz. Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: jaroslav.kraus@czso.cz.

Households belong to the modern times' phenomena and are a long-term interest of demographers. The changing structure of households is one of the fundamental attributes of the second demographic transition, as mentioned in the standard contribution of demographic literature (van Kaa, 1987, p. 32). There, it is stated that the changes in the propensity to marry, divorce, separate, remarry, or cohabit, changes in fertility behavior and in the age at which children leave home, along with mortality trends and differentials, have had a marked impact on household patterns in Europe. Therefore, one-family households and one-person households were used as an example of possible spatial changes in this paper.

The second demographic transition, as one of the key elements of demographic development of contemporary societies, is closely related to spatial distribution of data. It is precisely described in (Howell et al., 2016, Chapter 6), where the local character of demographic data is mentioned and universal principles by means of a spatial analysis are being sought for distribution of the data. For example, retrospectively, a fertility decline in Europe is a classic example of spatial autocorrelation. In general, demographic transition – in all its components – is said to have an impact on all aspects of life in society.

Spatial analysis of (not only demographic) data can be defined as quantitative data analysis, in which the explanation is based on explicit spatial variables or prediction of the phenomenon observed based on spatial autocorrelation. Two elements are related to spatial distribution of data: spatial dependence and spatial heterogeneity. Spatial dependence is connected with Tobler's first law of geography – everything is related to everything else, but near things are more related than distant things (Tobler, 1970). Demographic data are governed by the same statistical principles as any other data of stochastic character. Spatial analysis of demographic data thus is the very essence of geodemography (unlike social geography) and is related to all components of demographic development (Howell et al., 2016, p. 102 and other) in their mutual relationships.

When abstracted from spatial autocorrelation, a variety of statistical methods can be applied. Variability rates are often used to quantify regional differences and to develop regional differentiation (NUTS3 level). Gini's concentration coefficient is used in geographic surveys, because it overcomes the deficiencies of the coefficient of variation depending on the average and is therefore more appropriate for affecting the variability of asymmetric distributions typical of socio-geographical phenomena (Netrdová, 2012, p. 270). In the case of measuring the intensity of statistical dependence, it is possible to use a classical chi-square test (non-parametric) for a nominal variable (NUTS3 level).

The aforementioned principles can be illustrated using examples of data for households, surveying and a subsequent analysis, which is an integral part of population censuses. Spatial analysis of households is a relatively new topic, especially a geostatistical approach, however, several interesting contributions with a focus on the Czech Republic have been published on this subject (see Bleha, 2019; or Netrdová, 2009). Changes occurred after 1981 when demographers started to be more aware of the issue of households. The paper is based on work with data on two types of households: one-family households and one-person households. Development and structure of families as one of the types of households, become a key element of demographic trend in many countries; the methodological definition is standardized and coordinated well – see, for example (UNECE, 2011, p. 10 and other). According to this, one-family households are comprised of one couple without children, a couple with one or more children, or a lone parent with one or more children, independently of their de jure marital status. A one-person household is made by a person who lives alone in a separate housing unit or who occupies, as a lodger, a separate room (or rooms) of a housing unit but does not join with any of the other occupants of the housing unit to form part of a multi-person household. In compliance with international recommendations, households are determined based on 'place of usual residence'.

One-person households belong to modern-day phenomena; they are constantly increasing in number (both in absolute and relative values) and their observation provides basic characteristics of population development in many countries including the Czech Republic. The subject of formation and dissolution

of (multi-member) households is rather broad. This paper is devoted only to their current situation, i.e. regardless of their formation and possible dissolution due to the death of a household member or divorce of the married couple. Besides structure of households, it is also possible to pay attention to variables that are connected to households, e.g. income, housing prices, but also changes in gender roles (van Imhoff et al., 1995, p. 91 and other). Further, regarding these variables, it is possible, as with households, to expect spatial variability.

As stated previously, population censuses belong to basic sources of data on households (and characteristics of their members), as was the case in 2011. During the 2011 Census, both dwelling households and private households were surveyed in detailed structures. *The aim of the paper is to detect – on the distribution of the share of one-person households and of the share of households consisting of one complete family by municipality – whether there is a spatial autocorrelation and, further, whether the spatial autocorrelation is the same for both types of households or whether it is different (e.g. complementary).*

1 POPULATION AND HOUSING CENSUS IN THE CZECH REPUBLIC 2011

In the 2011 Population and Housing Census, over 4 375 thousand households were counted. The total amount of private households has been increasing over a long period of time. In the last ten years, it increased by 4% and since 1970 it has increased by over a million, expressed in an absolute value (see Table 1).

Table 1 Number of households in census years 1970–2011

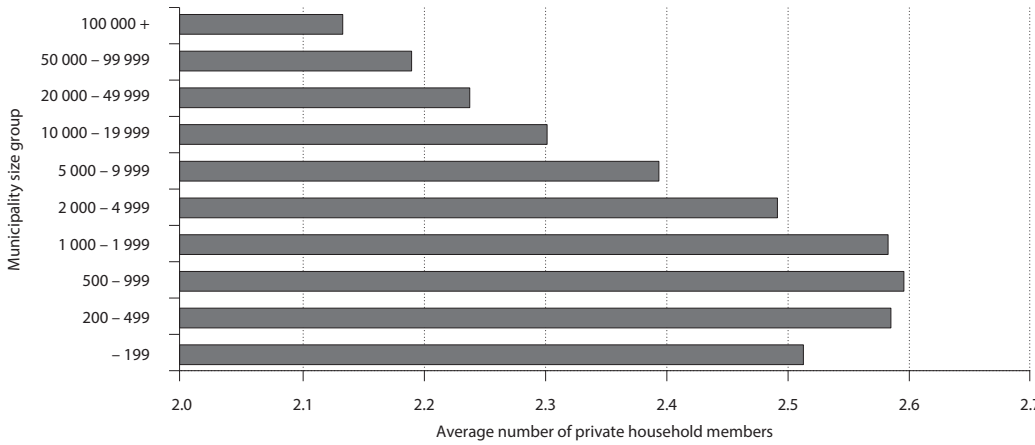
Households	1970	1981	1991	2001	2011	index 2011/1970	index 2011/2001
Total number of households	3 365 407	379 097	3 983 858	4 216 085	4 375 122	130	104
One-family households	2 526 778	2 760 247	2 856 608	2 803 340	2 667 867	106	95
One-person households	668 859	897 447	1 047 221	1 276 176	1 422 147	213	111

Source: 2011 Population and Housing Census (CZSO, 2013)

The basic characteristics of the development of households include the relative decline in the share of family households at the expense of uncompleted family households, and the absolute and relative growth of one-person households. While in 1970, private households consisting of one complete family made up two thirds of all private households, four decades later they made up hardly half. A decrease in their proportion was caused mainly by a marked absolute increase in one-person households. In 2011, one-person households comprised already a third of all households. The average size of a household has also been constantly decreasing for a long period of time. In 1970, on average 2.89 persons lived in a private household, whereas in 2011 it was only 2.34 persons. This, together with a change in the structure of private households, is a result of a long-term demographic trend, especially due to declines in fertility rate and a long-term high levels of divorce rate and an increasing availability of independent living (i.e. by frequent and simpler decomposition of complete families to singles and incomplete families) (CZSO, 2013, pp. 37–39).

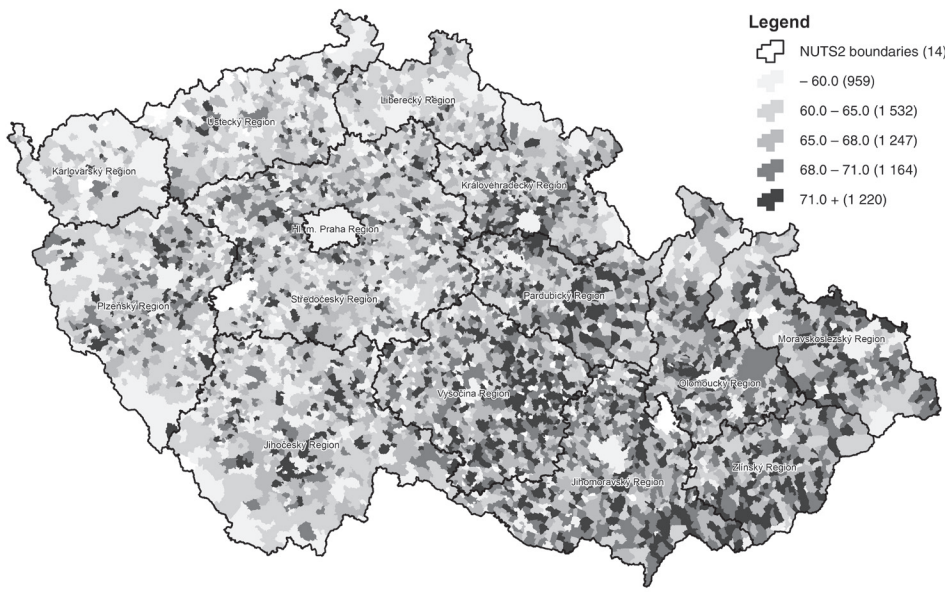
A household structure can be also viewed from the perspective of the number of households in a municipality (Figure 1). The highest average number of household members has been recorded in municipalities with a (usually resident) population of 200–1 999 (2.58–2.60 persons in a household); it is also valid that the bigger size of a municipality, the smaller the average size of a private household tends to be (CZSO, 2013, pp. 37–39).

Figure 1 The average number of private household members by municipality size group



Source: 2011 Population and Housing Census (CZSO, 2013)

Figure 2 Share of one-family households in all households by municipality in the CR

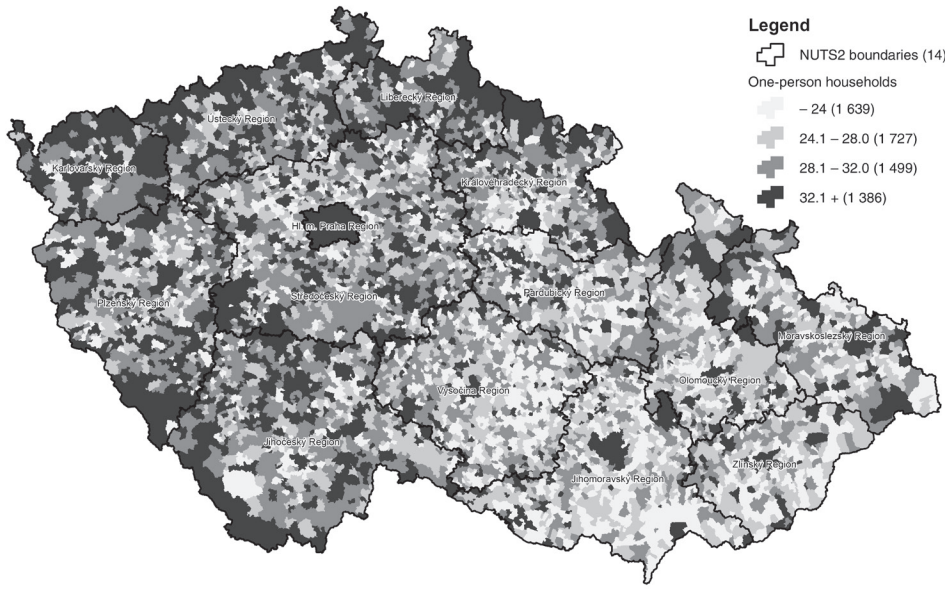


Note: For coloured map see the online version of Statistika journal No. 4/2019.
Source: Own calculation based on data from the 2011 Population and Housing Census (CZSO, 2013)

From Figures 2 and 3 it is obvious that the distribution of the two aforementioned household types is not homogeneous in the territory of the Czech Republic.[2] It appears that in general, one-person

[2] Processing thereof is based on results for municipalities. As at the Census date, 6 251 municipalities were in the Czech Republic; they are grouped together (with a hierarchy) to higher territorial units (NUTS).

Figure 3 Share of one-person households in all households by municipality in the CR



Note: For coloured map see the online version of Statistika journal No. 4/2019.
Source: Own calculation based on data from the 2011 Population and Housing Census (CZSO, 2013)

households are more frequent in the western part of the Czech Republic, while households consisting of one complete family are more common in the eastern part, while households consisting of one complete family are more common in the eastern part, relative to households of a given type. However, there are frequent exceptions from that distribution. The mentioned results raise the question *to what extent shares of one-person households and of one-family, households are influenced by their spatial distribution, i.e. whether they are spatially autocorrelated.* From a methodological point of view, variables can be seen as a continuous variable; a figure for a given territorial unit (e.g. a municipality) is a result of settlement of the entire municipality and, similarly, the settlement and structure of the entire CR results from data for all municipalities.

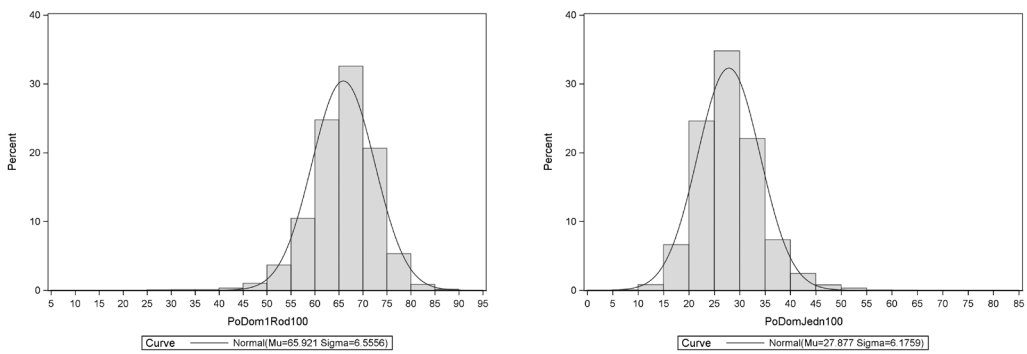
2 METHODS AND METHODOLOGY

Spatial autocorrelation may be a result of unobserved or hard-to-quantify processes, combined in various places, consequently causing spatial structuring of a given phenomenon. In the context of specification of econometric models (i.e. searching for explanatory variables), measuring of spatial autocorrelation can be considered to be a diagnostic tool. If there is a spatial autocorrelation, it is determined by examining whether the variable value for a given (e.g. geolocalized) observation is associated with values of the same variable for neighboring observations (INSEE, 2018, p. 67). Spatial autocorrelation may be positive, negative, or there is no spatial autocorrelation among given data. Spatial autocorrelation can be measured globally or locally; both ways assess the same, i.e. whether there is a spatial correlation of a given phenomenon – however, it is not the same. There are different ways of measuring spatial autocorrelation; Moran's I is often used.

The use of Moran's I requires data normality and stationarity (that is, the same data mean and data variance at any location). Moran's I, however, is rarely used in geostatistics in which data stationarity

is the main assumption and data normality is a desirable feature. (Krivoruchko, 2011, p. 61). Although it is not entirely clear from Figure 4 (especially after addition of a bell-shaped curve), this condition has not been met; the results of statistical tests that are used for tests for normality have not confirmed the hypothesis that the distribution of both the variables meets the condition. Many tests were elaborated to test the normality; in this case Kolmogorov-Smirnov test has been chosen followed by Cramér-von Mises test and Anderson-Darling test, respectively, for verification (SAS Base).

Figure 4 Histogram of frequency distribution of the share of one-family households (PoDom1Rod100) and the share of one-person households (PoDomJedn100)



Source: Own calculation

Based on a detailed data analysis the primary cause of why the condition of normality has not been met, was determined to be due to extreme values for some observations. However, because spatial autocorrelation is based on the concept of continuity of a variable, it was necessary to minimize the amount of excluded (i.e. extreme) observations so that continuity of data is disrupted as little as possible. The Czech Republic comprises 6 251 municipalities; after exclusion of extremely low and extremely high values of one-person households and one-family households we got 6 122 municipalities with the population of 10 413 thousand, equating to 99.8% of the total number of the usually resident population (10 437 thousand). However, that was still not enough to ensure normality; it was necessary to transform data in order to achieve the required normality. A commonly applied method of Box-Cox transformation (SAS Stat) was used, which generally is in the following form:

$$\begin{aligned} &(y^{\lambda} - 1) / \lambda, \quad \lambda \neq 0, \\ &\log (y), \qquad \lambda \neq 0, \end{aligned} \tag{1}$$

where we work with one lambda parameter. When lambda is approaching zero, it is basically a logarithmic transformation. An advantage of this type of transformation is that it is selected from a solution set based on individual values of the lambda parameter from a fixed interval so that the plausibility function (its logarithm) is maximized. For the computation itself we chose the software environment of the product (SAS Stat), namely the TRANSREG procedure, which enables data transformation without using an explicative variable or, to put it more precisely, with using a fictitious (constant) variable. The result of data transformation is satisfactory and the result of testing is that the transformed variable meets the condition of normality – see Table 2.

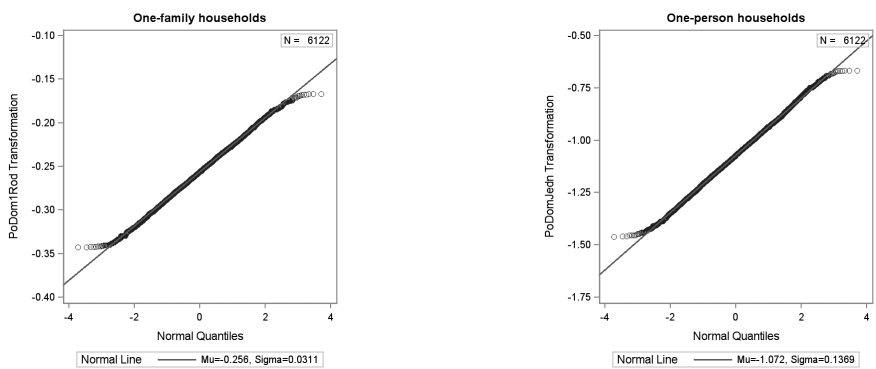
Table 2 Fitted Normal Distribution for transformed variables

One-family Household					Single Household			
Test	**Statistic**		**p Value**		**Statistic**		**p Value**	
Kolmogorov-Smirnov	D	0.011	Pr > D	0.067	D	0.008	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.058	Pr > W-Sq	>0.250	W-Sq	0.033	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.465	Pr > A-Sq	>0.250	A-Sq	0.354	Pr > A-Sq	>0.250

Note: If p Value > 0.05, the hypothesis on data normality is not declined.
Source: Own calculation

Similarly, for the result of a test, a shift of both variables to normality can be documented in a histogram of frequency distribution of both variables or by other graphic procedures such as a QQ plot – see Figure 5.

Figure 5 Quantile distribution of empirical data of a given type of household compared to quantiles of normal distribution



Source: Own calculation

Now we can start to investigate a spatial autocorrelation for both types of households on a newly defined set of 6 122 municipalities. Basic information can be found e.g. in Shekhar and Xiong (2008, p. 360, 644). In this paper we used Moran's I. The principle of computation is that it takes into account the difference between the value of the variable (i.e. the share of households) and the average of values of that variable for a given area (neighborhood). Moran's index is used preferably (in comparison to other ones), because it is more stable against extreme values, further it can be used in two ways (see below). The index can be written in several ways, it is frequently written as follows:

$$I_W = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \quad i \neq j\,. \tag{2}$$

Null hypothesis H_0, states that there is no spatial correlation in the given territory. Vice versa, if $I_w > 0$, then there is a positive autocorrelation, which means that high values are neighboring high ones and low values are neighboring low ones. In the case of a negative autocorrelation it would be the contrary.

Depending on the distribution of a spatial variable, the calculation of a median value:

$$E(I_W) = E(c_W) = -\frac{1}{n-1}, \tag{3}$$

and testing statistics:

$$\frac{I_W - E(I_W)}{\sqrt{\mathrm{Var}(I_W)}} \sim \frac{c_W - E(c_W)}{\sqrt{\mathrm{Var}(c_W)}} \sim \mathrm{N}(0,1) \tag{4}$$

can be defined.

A key element for calculation of indices of spatial autocorrelation is to determine the neighborhood, i.e. to select spatial entities that are neighbors *in definition*. The defining of the neighborhood is a rather complex issue, which should always be based on knowledge of the examined issue (i.e. determination of a working hypothesis on why given spatial elements are selected to be neighbors) and that has a major influence on the result of calculation of a spatial autocorrelation. Based on a definition of neighborhood, it is then possible to start calculation of the so-called spatially weighted matrix; in the I_w calculation formula, the elements of the matrix are denominated as w_{ij}. For more details about the issue see INSEE (2018, p. 57). Modelling of spatial correlations is also described clearly in (ArcGISPro), in which computations of spatial indices for this paper were also made.

Two working hypotheses of spatial autocorrelation were stated: the *influence of immediate neighborhood* and the *influence of a local center*. In the ArcGIS environment it meant that, the Moran's I index was calculated by the contiguity-edges-corners method and by the fixed-distance method. The first hypothesis results from an assumption that if two municipalities are neighboring (in terms of topology by their edge or a corner), then there is the biggest interaction between them. The second hypothesis is calculated as follows: neighbors are those whose centroids were less than 25.8 km apart from each other. This distance has been determined in such a way that each municipality has at least one municipality with more than 1000 inhabitants as a neighbor – thus it was defined as a local center.

Transformed variables of the shares (Box-Cox transformation, see above) of both one-family households and one-person households were worked with.

Table 3 Spatial autocorrelation (Moran's I) for the share of one-family households and the share of one-person households by chosen neighbourhood method

Test	One-family households		One-person households	
	Contiguity edges corners	Fixed distance	Contiguity edges corners	Fixed distance
Moran's Index	0.193**	0.117**	0.204**	0.112**
Expected Index	0.000	0.000	0.000	0.000
Variance	0.000	0.000	0.000	0.000

Note: ** p Value<0.05.
Source: Own calculation

The result (see Table 3) can be interpreted as follows: It is clear that in the data (i.e. in the share of one-family households or one-person households) there is a positive autocorrelation, which is not very high given the interval in which the theoretical values of Moran I are located. However, especially

in the case of the fixed-distance method and which in both cases is statistically significant. The value of the Expected Index indicator means that the index is calculated from randomly generated data streaming from a normal distribution, which is the case with this household data.

The Moran's index is a global statistic, which provides no information about the extent of local variation in spatial variability. For that there are tools that enable us to assess the local level of spatial autocorrelation (LISA) and to measure the intensity and importance of autocorrelation between the value of the variable in a spatial unit and the value of the same variable in neighboring spatial units. These indicators examine the following two features:

- for each observation they show the intensity of clustering of similar/opposite values around that observation;
- the sum of local indices at all observations is proportional to the corresponding global index, e.g. to global Moran's I.

In the case of Moran's I, its local value can be written as follows:

$$I_i = (y_i - \bar{y}) \sum_j w_{ij} (y_j - \bar{y}) , \tag{5}$$

and the value of the global index is as follows:

$$I_w = \text{konst} \cdot \sum_i I_i , \tag{6}$$

where:

$I_i > 0$ indicate clustering of similar values (higher or lower than the average for a given neighborhood),

$I_i < 0$ indicate clustering of different values.

Spatial clustering of similar or different values is observed *as follows: as High-High values (HH), Low-Low values (LL), High-Low values (HL), or Low-High (LH) values*. If we mean high value surrounded by another high values, resp. low value surrounded by another low value then they are referred to as hot spots, resp. cold spots. If we mean a high value surrounded by low values or a low value surrounded by high values, then these are spatial outliers (Anselin, 1995).

The significance of each local indicator is based on spatial distribution of data and statistics that is asymptomatically approaching the normal distribution:

$$z(I_i) = \frac{I - E(I_i)}{\sqrt{\text{Var}(I_i)}} \sim \text{N}(0,1) . \tag{7}$$

Since the global rate of spatial autocorrelation (Moran's I) proved to be distinctively higher in the case of usage of the neighboring municipalities method, local rates of Moran's I were further computed only for this method of neighborhood determination. Further, numbers of households that live in each of them were determined.

Another possibility of exploring spatial variability is to use the classical measures of association with respect to the type of data being examined. One test statistic for the hypothesis of no general association is the Pearson chi-square. This statistic is defined for i is from 1 to s and the summation for j is from 1 to r:

$$Q_P = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(n_{ij} - m_{ij})^2}{m_{ij}} , \text{ where:} \tag{8}$$

$$m_{ij} = E\{n_{ij} \mid H_0\} = \frac{n_{i+} n_{+j}}{n} \tag{9}$$

is the expected value of the frequencies in the *i* th row and *j* th column.

Measures of association when one or both variables are nominally scaled are more difficult to define, since you cannot think of association in these circumstances as negative or positive in any sense. However, indices of association in the nominal case have been constructed, and most are based on mimicking R-squared in some fashion. One such measure is the uncertainty coefficient, and another is the lambda coefficient (Stokes, 2012, p. 129).

Asymmetric lambda λ (*Colums* | *Rows*), is interpreted as the probable improvement in predicting the column variable *Y* given knowledge of the row variable *X*. The range of asymmetric lambda is $0 \le \lambda(C \mid R) \le 1$. Asymmetric lambda (C|R) is computed as:

$$\lambda\,(C|\,R) = \frac{\sum_i r_i - r}{n - r}, \tag{10}$$

and its asymptotic variance is:

$$\mathrm{Var}(\lambda\,(C|\,R)) = \frac{n - \sum_i r_i}{(n-r)^3}\left(\sum_i r_i + r - 2\sum_i (r_i \mid l_i = l)\right). \tag{11}$$

The nondirectional lambda (symmetric) is the average of the two asymmetric lambdas, ($\lambda(C \mid R)$ and ($\lambda(R \mid C)$. Its range is $0 \le \lambda \le 1$. Lambda symmetric is computed as:

$$\lambda = \frac{\sum_i r_i + \sum_j c_j - r - c}{2n - r - c} = \frac{w - v}{w}, \tag{12}$$

and its asymptotic variance is computed as:

$$\mathrm{Var}(\lambda) = \frac{1}{w^4}\left(wvy - 2w^2\left(n - \sum_i\sum_j (n_{ij} \mid j = l_i, i = k_j)\right) - 2v^2\,(n - n_{kl})\right). \tag{13}$$

The uncertainty coefficient U is the symmetric version of the two asymmetric uncertainty coefficients. Its range is $0 \le U \le 1$. The uncertainty coefficient is computed as:

$$U = 2(H(X) + H(Y) - H(XY)) \,/\, (H(X) + H(Y)), \tag{14}$$

and its asymptotic variance is:

$$\mathrm{Var}(U) = 4\sum_i\sum_i \frac{n_{ij}\left(H\,(XY)\ln\left(\frac{n_i n_j}{n^2}\right) - (H\,(X) + (H\,(Y))\ln\left(\frac{n_{ij}}{n}\right)\right)^2}{n^2\,(H\,(X) + (H\,(Y))^4}, \tag{15}$$

where *H*(*X*), *H*(*Y*), and *H*(*XY*) are defined in the previous section. See (SAS Stat) for completed description.

Gini index (or Gini ratio), is a measure of statistical dispersion and it is the most commonly used measurement of inequality preferably used in economics. It measures the inequality among values of a frequency distribution. Index of zero expresses perfect equality, where all values are the same, index of 1 (or 100%) expresses maximal inequality among values. The sample Gini coefficient was calculated using the formula:

$$G = \frac{1}{2\overline{X}n(n-1)} \sum_{i=1}^{n} (2i - n - 1)X_i, \tag{16}$$

where X_i are the sizes sorted from smallest to largest, $X_1 \leq X_2 \leq X_n$ (Dixon, 1987).

3 RESULTS

It is clear from the Table 4 that the highest number and the highest share of both one-family households and one-person households is made by households in municipalities, in which the value of the local index of spatial autocorrelation is not statistically significant (the *Not Significant* line). In the case of complete households consisting of one family it is 4 805 municipalities, in the case of one-person households it is 4 766 municipalities from the total number of 6 121 municipalities, which entered the computation after data transformation.

Table 4 Absolute and relative frequencies of individual types of households by municipality of the CR

Statistic	One-family households		One-person households	
	Frequency	Percent	Frequency	Percent
Not Significant	1 747 947	65.7	715 930	50.5
HH	118 119	4.4	117 824	8.3
HL	22 167	0.8	540 553	38.1
LH	542 743	20.4	9 078	0.6
LL	231 240	8.7	36 017	2.5
Total	2 662 216	100.0	1 419 402	100.0

Source: Own calculation

From map outputs (see the online version of Statistika journal No. 4/2019) it is possible to find that for both variables there are territories in which the spatial autocorrelation is higher than in the remaining territory of the Czech Republic. In the case one-family households it applies to many municipalities in the *Karlovarský* Region, the *Ústecký* Region, the *Liberecký* Region, and partially also the *Plzeňský* Region, and the *Jihočeský* Region, where there are mainly Low-Low clusters. In the case of one-person households, the situation in those municipalities is – quite logically – the opposite, i.e. High-High clusters. An interesting situation is observed in the eastern part of the Czech Republic. For the share of the one-family households variable, High-High clusters (hot spots) are quite frequent (a high share of complete households consisting of one family), meaning that this share is significant in many municipalities. But also Low-High cases (spatial outliers), i.e. the low values of this share accompanied by a high share in neighbouring municipalities are significant. Similarly, for one-person households in the eastern part of the Czech Republic, there are Low-Low clusters (cold spots), meaning there is a lower share of one-person households spread in a significant way. At the same time a High-Low type (spatial outliers), i.e. the high value of the share of one-person households in the given municipality is observed in the neighbourhood of municipalities with the low share. This is the case for the cities of *Brno*, *Ostrava*, *Pardubice*, *Hradec Králové*, but also for many smaller towns in the eastern part of the country. The same situation is seen for the share of one-person households in the capital city of Prague, while in the neighbouring municipalities

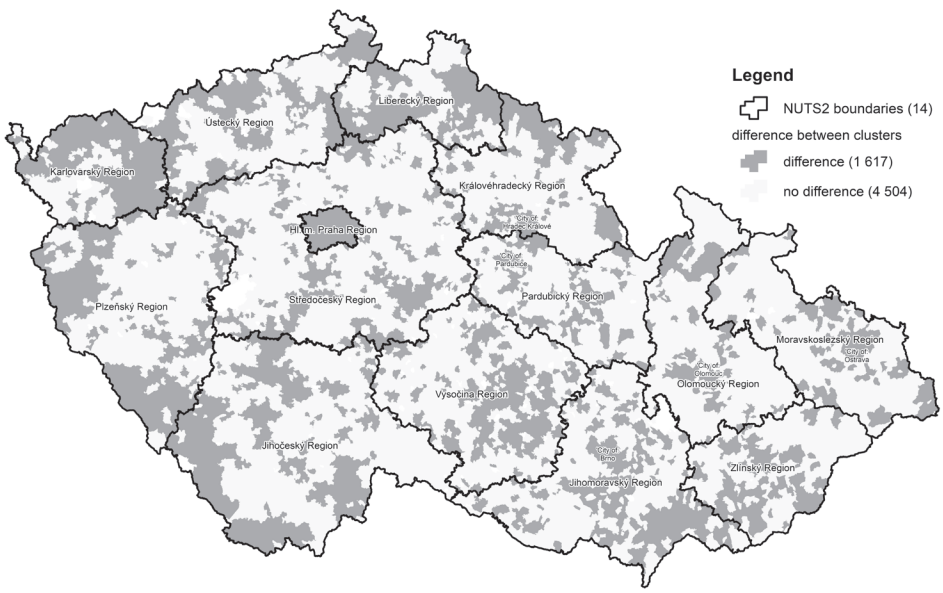
of the *Středočeský* Region this phenomenon is less frequent. Interpretation of hot spots and cold spots for both variables (one-person households, one-family households) is interesting and will be addressed in the upcoming article on spatial regression analysis.

It is interesting to compare agreement of both types of clusters, i.e. one-family households and one-person households by municipality (see Figure 6). It turns out that the agreement or disagreement is not the same in all regions and that it differs even within the regions. For example, virtually the entire *Středočeský* Region contains the not significant result for the clusters of both types of households and therefore the agreement is high there (light colour is prevailing). In contrast, the *Capital City of Prague* is, in the case of one-person households, the High-Low type, i.e. a high share of one-person households in Prague surrounded by a low share of one-person households in neighbouring municipalities (on average). In the case of one-family households, it is the *Not Significant* type and therefore the value for Prague is marked with a dark colour on the map.

As in the case of Prague and other big cities, also in the remaining (i.e. smaller) municipalities of the Czech Republic, the situation is differentiated and *it cannot be said that shares of one-family households and one-person households are a complementary phenomenon: i.e. where there is a high share of one-person households the opposite is true for one-family households*.

Frequencies of significant clusters are different for both types of households. As established, the High-Low type refers to the high value on the given territory (municipality) surrounded by low values in the neighbourhood (on average) and a Low-High type means that the low value on the given territory is surrounded by high values in its neighbourhood (on average). For the case of one-person households it means that the High-Low types occur in big cities (e.g. *Brno*, *Ostrava*, *Olomouc*, *Hradec Králové*, *Pardubice*),

Figure 6 Cluster and outlier analysis according to the agreement of individual clusters of municipalities in the Czech Republic (comparison between one-family households and one-person households)

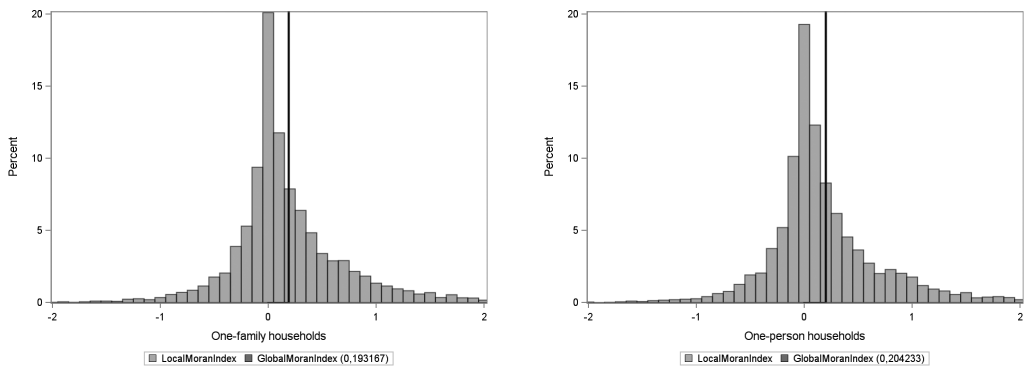


Note: For coloured map see the online version of Statistika journal No. 4/2019.
Source: Own calculation

while households consisting of one family (of Low-High type) are more frequently observed in smaller municipalities and are therefore more likely to occur in the western part of the country.

Local indicators of spatial autocorrelation enable to identify areas in which similar values are clustered in a statistically significant way. In general, if the global spatial autocorrelation is strong or at least observable (as it is in the case of households) then local indicators indicate those areas that have a special impact on the global process (local autocorrelation is higher than the total autocorrelation) or, vice versa, where an obvious autocorrelation exists although the global autocorrelation is not significant.

Figure 7 Indexes of global and local rates of spatial autocorrelation (Moran's I)

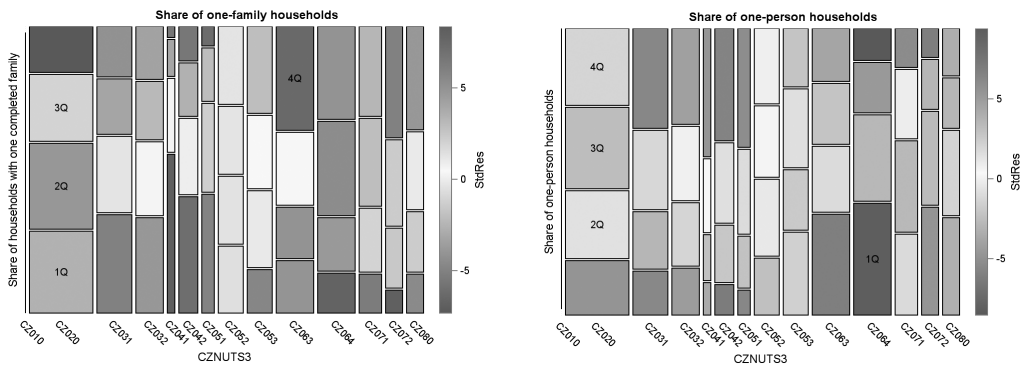


Note: For coloured figure see the online version of Statistika journal No. 4/2019.
Source: Own calculation

The relationship can be observed in the histogram of frequencies showing local and global rates of autocorrelation of one-family households or one-person households as computed by nearest neighbour method.

From mutual comparison of the global rate and local rates (see Figure 7) it is apparent in both cases (i.e. in the case of one-family households as well as in the case of one-person households) the global rate in the territory of the Czech Republic does not capture the observed phenomenon in full. Nevertheless, especially the big share of non-significant values of spatial autocorrelation should be kept in mind.

Figure 8 Frequency plot by type of households and NUTS3



Note: For coloured figure see the online version of Statistika journal No. 4/2019.
Source: Own calculation

The previous results can be compared with the results of spatial variability measurement, i.e. by measures of association at the NUTS3 level.

The mosaic plot (Figure 8) is a graphical depiction of the frequency table. It shows the distribution of the weight categories by dividing the x axis into 14 intervals (NUT3 level). The length of each interval is proportional to the percentage of the share of households, which are divided into four quartiles by type. Within each quartile category, the share of households is further subdivided by NUTS3 level. In order to make the residuals comparable across cells, the standardized residuals were added on the right side of the both graphs. The width of the column indicates the frequency of the phenomenon monitored.

The results show that the distribution is neither identical nor complementary. Both types of households create separate spatial patterns and show a relatively large variability of the observed phenomenon.

Table 5 Statistics for share of households by NUTS3

Statistic	Share of one-family households			Share of one-person households		
	DF	Value	Prob	DF	Value	Prob
Chi-Square	39	634.432	<0.0001	39	580.641	<0.0001
Likelihood Ratio Chi-Square	39	655.759	<0.0001	39	594.397	<0.0001
MH Chi-Square (Rank Scores)	1	17.470	<0.0001	1	322.076	<0.0001
Phi Coefficient		0.322			0.308	
Contingency Coefficient		0.306			0.294	
Cramer's V		0.186			0.178	
Gini index		0.050			0.113	

Source: Own calculation

Output in Table 5 displays the Chi-Square statistics $QP = 634.4328$ with 39 df and $p<0.0001$ for variable share of one-family households and $QP = 580.641$ with 39 df and $p<0.0001$ for variable share of one-person households. Both results are statistically significant on *0.05* level. Other statistics, such as the Mantel-Haenszel Chi-Square statistic with the result that is also statistically significant at *0.05* and that also shows a statistically significant dependence of both variables, was calculated too.

Interesting is the comparison with the result of the GINI index calculation, which is a measure of statistical dispersion and the most commonly used measurement of inequality. Multiple approaches can be used to estimate the Gini coefficient. One of the frequently used estimates is the so-called Somers' d statistics, but in this case the GINI index was calculated directly according to the procedure described in (Dixon, 1987). The results show that the value of the Gini index (especially in the case of share of one-family households) is approaching zero, therefore *there is no significant diversity in the data at NUTS3 level*.

Previous results of the Gini index confirm the association rates contained in Table 6. The entry was the share of one-family households resp. share of one-person households by NUTS3 (nominal variable) and the output by Lambda Statistics and Uncertainty Coefficients. Again, the results contained in Table 6 show that the association rates do not deviate significantly from zero and thus confirm the lack of diversity of the shares of both types of households at the NUTS3 level.

Table 6 Measures of association for share of households by NUTS3

Statistic	Share of one-family households		Share of one-person households	
	Value	ASE	Value	ASE
Lambda Asymmetric C|R	0.012	0.004	0.009	0.004
Lambda Asymmetric R|C	0.128	0.010	0.123	0.011
Lambda Symmetric	0.068	0.006	0.064	0.007
Uncertainty Coefficient C|R	0.022	0.002	0.020	0.002
Uncertainty Coefficient R|C	0.039	0.003	0.035	0.003
Uncertainty Coefficient Symmetric	0.028	0.002	0.025	0.002

Source: Own calculation

The relationship can be observed in the histogram of frequencies showing local and global rates of autocorrelation of one-family households or one-person households as computed by nearest neighbour method.

From mutual comparison of the global rate and local rates (see Figure 7) it is apparent in both cases (i.e. in the case of one-family households as well as in the case of one-person households) the global rate in the territory of the Czech Republic does not capture the observed phenomenon in full. Nevertheless, especially the big share of non-significant values of spatial autocorrelation should be kept in mind.

CONCLUSION – DISCUSSION

The aim of the paper was to study the issue of population trend, represented by households in a slightly different way, preferably by means of spatial autocorrelation. It is indisputable that the Czech Republic is a rather homogeneous territory and that changes resulting from the development of the population take place over time (population ageing, change in the structure of households). However, this does not imply that differences among individual parts of the country cannot be observed and that subsequent impacts of these changes cannot be investigated on the level of education or economic characteristics.

The highest number and highest share of both one-family households and one-person households are made up of households in municipalities, in which the value of the local index of spatial autocorrelation is not statistically significant. From map outputs it is, however, possible to conclude that for both variables there are territories in which the spatial autocorrelation is higher than in the remaining territory of the CR.

Frequencies of significant clusters are different for both types of households. In the case of one-person households it means that the High-Low types occurs in big cities, while households consisting of one family are more frequently cases of smaller municipalities and are more likely to occur in the western part of the country.

From mutual comparison of the global rate and local rates it is obvious in both cases (i.e. in the case of one-family households as well as one-person households) that the global rate on the territory of the Czech Republic does not capture the observed phenomenon in full. Nevertheless, especially the big share of non-significant values of spatial autocorrelation should be kept in mind.

Comparison with statistics calculated on NUTS3 level (again) show, that the distribution is neither identical nor complementary. Both types of households create separate spatial patterns and show a relatively large variability of the observed phenomenon. The value of the Gini index (especially in the

case of share of one-family households) is approaching zero and therefore there is no significant diversity in the data at NUTS3 level.

The basic fact that has been ascertained is that shares of one-family households and one-person households are not a complementary phenomenon in the territory of the Czech Republic; i.e. in municipalities (that are defined as neighbouring), in which there is a higher share of one-person households there is, in general, a lower share of one-family households and vice-versa. In a given territory, other factors also exert an influence (e.g. size group of the municipality of the place of residence); they are modelling the situation and deserve further attention.

Comparing the agreement of both types of clusters, i.e. one-family households and one-person households by municipality, shows that the agreement or disagreement is not the same in all regions and that it differs even within regions.

Obviously, there is a problem in the interpretation of the results with respect to the above calculations. The Gini coefficient implies low variability between NUTS3 regions, which is in part contradictory to the results of the spatial autocorrelation. Does it therefore make sense to focus on regional differentiation of the share of different household types?

There could be two explanations. The first computes the so-called MAUP, which is a source of statistical bias that can significantly impact the results of statistical hypothesis tests (Openshaw, 2000). The results of the Gini coefficient, as well as the association rates were calculated at the NUTS3 level and only then interpreted, while the results of spatial autocorrelation were calculated directly at the municipal level.

The second explanation is essentially related to data. If we respect the spatial autocorrelation in the data, then it is necessary to choose such methods that allow spatial autocorrelation – *by definition*, such as Moran's Iindex. This explanation is more realistic according to the author of the paper.

Searching for causes of existence or non-existence of spatial autocorrelation on the level of municipalities is a different challenge; it deserves more intensive attention and an explanation based on age, education or economic variables may be present a first possible option.

ACKNOWLEDGMENT

This paper is supported by grant GAČR, 2018–2020 No. 18-12166S.

References

ANSELIN, L. *Local Spatial Autocorrelation Clusters.* The Center for Spatial Data Science, 2016.

ANSELIN, L. Local Indicators of Spatial Association – LISA. Geographical Analysis, 1995, 27, pp. 93–115. DOI: 10.1111/j.1538-4632.1995.tb00338.x.

ARCGIS PRO [online]. ESRI. [cit. 3.11.2019] <http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/modeling-spatial-relationships.htm>.

BIVAND, R., PEBESMA, E., GÓMEZ-RUBIO, V. Applied Spatial Data Analysis with R. 2nd Ed. *Modelling Areal Data*, Springer, Science + Business Media, 2013.

BLEHA, B. AND ĎURČEK, P. An interpretation of the changes in demographic behaviour at a sub-national level using spatial measures in post-socialist countries: A case study of the Czech Republic and Slovakia. *Papers in Regional Science*, February 2019, Vol. 98, No. 1, pp. 331–352.

CZSO. *Atlas sčítání 2011* (in Czech). Prague: Czech Statistical Office, 2013.

CRESSIE, N. *Statistics for spatial data.* Rev. Ed. New York: Wiley, 1993.

DE SMITH, M. J., GOODCHILD, M., LONGLEY, P. A. *Geospatial Analysis.* 6th Ed. Global spatial autocorrelation, 2018.

DIXON P. et al. Bootstrapping the Gini Coefficient of Inequality. *Ecology*, Oct. 1987, Vol. 68, No. 5, pp. 1548–1551.

FISCHER, M. AND GETIS, A. eds. *Handbook of Applied Spatial Analysis, Software Tools, Methods and Applications.* Springer, 2010.

GRIFFITH, D. A., CHUN, Y., DEAN, D. J. *Advances in Geocomputation, Geocomputation 2015.* The 13th International Conference, Springer.

HOWELL, F. M. et al. Recapturing Space New Middle-Range Theory in Spatial Demography. *Demography Is an Inherently Spatial Science*, Spatial Demography, Springer, 2016.
CHING-LAN, CH., YI-CHI, CH., TZU-MING, L., YEA-HUEI, K. Y. *Using spatial analysis to demonstrate the heterogeneity of the cardiovascular drug-prescribing pattern in Taiwan.* BMC Public Health, 2011.
INSEE-EFGS-EUROSTAT. *Handbook of Spatial Analysis.* 2018.
KRIVORUCHKO, K. Spatial Statistical Data Analysis for GIS Users. *Statistical Approach to GIS Data Analysis, Principles of Modeling Regional Data.* ESRI Press, 2011.
NETRDOVÁ, P. AND BLAŽEK, J. Aktuální tendence lokální diferenciace vybraných socioekonomických subjektů v Česku: směřuje vývoj k větší mozaikovitosti prostorového uspořádání? (in Czech), *Geografie*, 2012, 3, pp. 266-288
NETRDOVÁ, P. AND NOSEK, V. Vývojové pravidelnosti a specifika geografické diferenciace obyvatelstva a jeho struktury na úrovni obcí v Česku v transformačním období (in Czech). *Geografie*, 2018, 123, 2, pp. 225–251.
OPENSHAW, S. AND ABRAHART, R. J. eds. *Geocomputation. The modifiable areal unit problem.* Taylor & Francis, 2000, pp. 36–38. ISBN 0-203-30580-9.
VAN IMHOFF, E. et al. *Household Demography. Theories of Household Formation: Progress and Challenges.* Springer, Science + Business Media, 1995.
VAN KAA, D. *Europe's Second Demographic Transition, Population Bulletin.* March 1987, Vol. 42, No. 1.
SAS. *Base SAS(R) 9.4 Procedures Guide: Statistical Procedures* [online]. 3rd Ed. [cit. 3.11.2019] <http://support.sas.com/documentation/cdl/en/procstat/67528/HTML/default/viewer.htm#procstat_univariate_details52.htm>.
SAS. *Stat SAS(R) 9.2 User's Guide* [online]. 2nd Ed. [cit. 3.11.2019] <https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_transreg_sect015.htm>.
SHEKHAR, X. *Encyclopedia of GIS.* Springer Science + Business Media, 2008.
STOKES, E. et al. *Categorical Data Analysis.* 3rd Ed. SAS Institute Inc., 2012.
TOBLER, W. A Computer Movie Simulating Urban Growth in the Detroit Region [online]. *Economic Geography. Supplement: Proceedings,* International Geographical Union, Clark University, 1970, Vol. 46, pp. 234–240. [cit. 3.11.2019] <http://www.jstor.org/stable/143141>.
UNECE. *Measurement of emerging forms of families and households.* NY and Geneva, 2011.

ANNEX

Figure A1 Cluster and outlier analysis of share of one-family households by municipality in the Czech Republic

Note: See the online version of *Statistika journal* No. 4/2019.
Source: Own calculation

Figure A2 Cluster and outlier analysis of share of one-person households by municipality in the Czech Republic

Note: See the online version of *Statistika journal* No. 4/2019.
Source: Own calculation

Determinants Affecting Health of Slovak Population and their Quantification

Mária Vojtková[1] | *University of Economics in Bratislava, Bratislava, Slovakia*
Eva Kotlebová[2] | *University of Economics in Bratislava, Bratislava, Slovakia*
Daniela Sivašová[3] | *University of Economics in Bratislava, Bratislava, Slovakia*

Abstract

The state of health of the population is the result of various determinants, but also a barometer of the conditions that affect the formation of the individual's health. In a healthy society, a healthy individual can develop, and healthy individuals can develop from a healthy society. This article deals with the analysis of the impact of selected factors on the health status of the Slovak population. This is based on data from the latest EHIS (The European Health Interview Survey). We worked with the respondent's answer to the question whether he / she suffers from a long-term health problem (variable with variations yes-no). From the variables surveyed, we chose the ones we thought they could have effect on the selected indicator. With respect to the binary dependent variable, we used logistic regression for the analysis, where all calculations have been carried out in the SAS Enterprise Guide statistical program. The results are findings that have to some extent confirmed our assumptions about the impact of selected factors on health, although some of them have not been shown to the extent we expected.

Keywords

Health, health determinants, EHIS, logistic regression

JEL code

I10, C31

INTRODUCTION

Health is an important attribute of quality of life and well-being. Not only does it represent functional and instrumental value, but it also has importance for one's own identity as it determines who one is (Blaxter, 2010). In order to determine health, it is essential to define precisely when a person is healthy and when we can consider him / her sick. There is no reliable and accurate definition of health, not even that of transition from a healthy person to a sick person, because several exogenous and endogenous factors constantly cause gradual or sudden changes in human health. However, there are many definitions

[1] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Statistics, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia. E-mail: maria.vojtkova@euba.sk, phone: (+421)267295724.
[2] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Statistics, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia. E-mail: eva.kotlebova@euba.sk, phone: (+421)267295718.
[3] University of Economics in Bratislava, Faculty of Economic Informatics, Department of Statistics, Dolnozemská cesta 1, 852 35 Bratislava, Slovakia. E-mail: daniela.sivasova@euba.sk, phone: (+421)267295731.

of health. The most commonly used definition under the Constitution of the World Health Organization (WHO), which entered into force on 7 April 1948, is that "the health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity".[4] An important attribute is subjective feeling of the examined person, but still based on it we can only estimate the health condition.

Health and well-being are a very important part of human life, but they are influenced by many factors (Evans, Barer, Marmor, 2017; Marmot, 2005). Those related to poor health, disability, disease or deaths are known as risk factors. The risk factor is an individual's behavior or condition that increases the probability of disease or injury. They are often presented in isolation, but experience has shown that they interact. Most factors, such as lifestyle and, in part, social factors are largely up to the individual´s decision whether and to what extent they will incorporate them into their lives. The aim of this article is to identify the factors that affect the health of the population in Slovakia.

According to a survey of Citizens' Views on the Future of Slovakia (Bunčák et al., 2009), health and long life ranked second in the list of preferred life goals of the Slovak population. When it comes to future concerns, responses such as illness, deterioration of health, as well as a lack of funding for medicines and health care, came first.

In general, in the UN Human Rights Declaration, health, medical care and sickness are considered fundamental human rights. In November 2017, the Health Profile of the country was published, based on the collective work of the OECD and the European Observatory on Health Systems and Policies in cooperation with the European Commission (OECD, 2017). It is an overview of the state of health of the population and health care of the individual countries of the European Union (the EU). Based on these profiles of the 28 EU countries, it is evident that Slovakia's health has improved compared to previous years, but Slovakia still lags behind the EU average. This is evident, for example, by the average life expectancy at birth, which is one of the main synthetic indicators of population living conditions and mortality rates (indirect indicators relating to health). In 2017 it reached 77.3 years in Slovakia, which is shorter by 3.6 years compared to the EU28 average. There is a big difference between female and male sex – women live on average 7 years longer (80.7) than men (73.8). This gender gap is greater than the EU28 average (5.9 years). On the contrary, the interesting fact is that the healthy life years in Slovakia in 2017 are the same for men and women, 55.6 years, while in the EU28 there is a slight difference, 64 years for women and 63.5 years for men.

1 DETERMINANTS OF HEALTH

Differences in morbidity or mortality between countries are not only dependent on the quality of healthcare. It is true that in some countries (including Slovakia), as compared to more advanced countries, not so much money was invested to health care, either in more advanced technologies or medicines, but other important factors also affect the health status of the population.

Almost all diseases are largely initiated by risk factors, and their presence decides whether or not the disease will break out. Risk factors, in turn, are strongly influenced by the environment, which may encourage or even eliminate their occurrence. Therefore, we consider the environment as a significant determinant of health. Each risk factor has its own specifics – for some diagnoses it has a high initiation potential, for other diagnoses it can eliminate their occurrence.

We categorize health determinants into certain groups, which are:

a) lifestyle,
b) genetic basis,
c) socio-economic,
d) health care.

[4] *<https://www.who.int/about/who-we-are/constitution>.*

Among these factors, lifestyle has the highest impact on health – its impact is up to 50–60%, other factors contribute significantly lower: genetic basis 10–15%, socioeconomic and natural environment 20–25% and health care 10–15%. (Čeledová and Čevala, 2010).

The aim of this article is to identify the factors that influence the health of Slovaks, based on the European Health Survey (hereinafter referred to as "EHIS"), which was carried out in Slovakia in the second wave by the end of 2015. The number of respondents was 5 490. As we are interested in what determines our health, we decided to choose as the target variable the expression of the respondent, whether he has a certain disease or a long-term health problem for more than 6 months. A complementary goal is to quantify the impact of significant determinants on the target variable. Selected determinants (factors), whose influence we decided to investigate include:[5]

- Age,
- Gender,
- Legal marital status,
- Highest level of educational attainment,
- Respondents' employment status,
- Net monthly equivalent household income,
- General health condition perceived by the respondent,
- Hospitalization in hospital over the last 12 months,
- Last visit to a general practitioner or family doctor,
- The respondent could not afford prescribed drugs in the last 12 months,
- Body mass index (BMI),
- Physical effort in performing duties – including paid and unpaid work activities,
- Frequency of fruit consumption, excluding fruit juices made from concentrate,
- Frequency of consumption of vegetables or salads, excluding potatoes and vegetable juices made from concentrates,
- Frequency of alcoholic beverages of any kind in the last 12 months,
- Smoking habits.

Since the target variable is categorical, we decided to use the logistic regression method to achieve the designed goal. Its aim is to find the most suitable model for describing the relationship between a binary dependent variable and a set of selected explanatory variables, which can be both continuous and categorical. The analysis itself was performed using the SAS Enterprise Guide statistical tool (Dhand, 2010).

2 METHODOLOGY

To assess the statistical significance of the impact of the considered factors on probability that a person will suffer from a long-term health problem, we have decided to use the logistic regress model with logit link function (Hilbe, 2016; Hosmer and Lemeshow, 2013; Bagley et al., 2001):

$$\text{logit}(p_i) = \ln \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}, \tag{1}$$

where p_i is the probability that a person will suffer from the long-term health problem. $\beta_0, \beta_1, \ldots, \beta_k$ are parameters of logit model and $x_{i1}, x_{i2}, \ldots, x_{ik}$, where $i = 1, 2, \ldots, n$ are the values of the explanatory variables $X_1, X_2, \ldots, X_k$ observed for i-th statistical unit (in this case a person). To estimate the parameters of the logistic regression model, we used the standardly applied maximum likelihood method that maximizes the likelihood function L. To obtain maximum likelihood estimates is generally used the Newton-Raphson algorithm.

[5] A detailed description of the variables with each character category is given in the Annex.

The significance of the logistic model is verified by testing the null hypothesis, according to which holds $\boldsymbol{\beta}^{\mathrm{T}} = (\beta_1 \quad \beta_2 \quad \ldots \quad \beta_k) = \mathbf{0}^{\mathrm{T}}$, against an alternative hypothesis which is claiming that at least one regression coefficient is non-zero. We used Chi-square tests (Likelihood ratio, Score statistics, Wald statistics) in our analysis. It is well known (Allison, 2012) that for large samples all tests generally give comparable results. To verify the significance of the impact of individual explanatory variables on probability *p*, we applied the Wald test in SAS Enterprise Guide. For each of the listed factors above we tested the null hypothesis according to which the explanatory variable does not affect the probability of the investigated event occurrence. To verify the null hypotheses, we used Wald´s test statistics:

$$Wald = \hat{\boldsymbol{\beta}}^{\mathrm{T}} \cdot \mathbf{S}_{\mathbf{b}}^{-1} \cdot \hat{\boldsymbol{\beta}}\,, \tag{2}$$

where $\hat{\boldsymbol{\beta}}$ is vector of estimates of regression coefficients that stand at dummy variables for the respective factor - a categorical explanatory variable and $\mathbf{S}_{\mathbf{b}}$ is a variance-covariance matrix of a vector $\hat{\boldsymbol{\beta}}$. Wald's test statistic has an asymptotic χ^2 distribution with a number of degrees of freedom equal to the number of estimated vector parameters $\boldsymbol{\beta}$. A special case of above test is the Wald test, which verifies the statistical significance of one regression coefficient. In this case Wald statistics has an asymptotic χ^2 distribution with 1 degree of freedom and it is as follows:

$$z^2_{Wald}) = \left(\frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}\right)^2, \tag{3}$$

where $s_{\hat{\beta}_j}$ is an estimated standard error of the j^{th} estimated coefficient.

In logistic regression, the effect of the explanatory variable X_j on explanatory variable Y is quantified by the odds ratio (OR), which is estimated as follows:

$$OR_j = e^{\hat{\beta}_j}\,, \tag{4}$$

where $\hat{\beta}_j$ is an estimate of the relevant regression coefficient. The odds ratio in binary logistic regression represents the change of the chance that Y = 1 (in our – case that a person will suffer from a long-term health problem) versus the chance that Y = 0 (in our case a person will not suffer from a long-term health problem), influenced by unit increase of the explanatory variable X_j under the condition of ceteris paribus. If the explanatory variable is an artificial variable, the odds ratio compares the odds at two different levels of the predictor.

The quality of the logistics model can be evaluated according to various measures. One group consists of penalty models of quality, namely AIC – Akaike Information Criterion and SC – Schwarz-Criterion, which are based on the logarithmic transformation of the likelihood function. The second group consists of the measures of association between predicted and original values of the dependent variable, including Somers D, Goodman-Kruskal gamma, Kendall tau-a and c-statistics[6] (Katamuri, 2017).

3 ASSESSMENT OF SELECTED DETERMINANTS INFLUENCING HEALTH

In this section, we focused on assessment contingency and creating a model of logistic regression, where the modeled variable is a "Long Term Health Problem", specifically whether or not the respondent suffers from any disease or health problem that persists for at least 6 months. At the same time, the dependent variable is the main subject of the study, with two variations 1 – yes, 2 – no.

[6] Note that the concordance index, c, also gives an estimate of the area under the receiver operating characteristic (ROC) curve when the response is binary.

We confirmed, that the fact, whether respondent suffered from a long-term health problem, was in 2015 significantly influenced by almost all selected determinants, by the analysis of association or contingency (Šoltés, 2008) using Chi-square tests shown in table 1. In case of significant determinants, the *p*-value is lower than the commonly used significance level 0.05. Surprisingly, only factors related to the lifestyle of the respondent, namely the frequency of fruit and vegetable consumption, proved to be insignificant determinants. Due to its nature, we have omitted the numeric variable age.

To measure the intensity of this dependence, we constructed different measures. To interpret the results, we decided to use Cramer V, which is based on the average square contingency and is a useful measure when comparing the degree of association for contingency tables of different dimensions. This degree of association has shown that the risk of a respondent's suffering from a long-term health problem lasting at least 6 months is mostly affected by the respondent's General health condition, the Ability to buy prescribed medication, and the Status of the job. A moderate significant relationship between the modeled variable and the factor can be observed with the Last doctor visit and Marital status factors.[7] The lowest degree of significant dependence can be observed between the dependent variable and Physical effort in the performance of duties, Smoking habits and Gender.

By analyzing the contingency, we assessed the relationship between the dependent variable and the analyzed determinants individually, but it should also be taken into consideration that there may also

Table 1 Assessment of contingency between analyzed determinants and risk of long-term health problem of Slovak population

Statistic	SUB STATUS			DRUGS			EMPLOYMENT			VISIT		
	DF	Value	Prob	DF	Value	Prob	DF	Value	Prob	DF	Value	Prob
Chi-Square	2	2 178.0161	<.0001	2	1 389.4363	<.0001	3	1 318.0307	<.0001	2	735.0518	<.0001
Likelihood Ratio Chi-Square	2	2 590.8150	<.0001	2	1 437.7054	<.0001	3	1 491.2656	<.0001	2	741.2896	<.0001
Cramer's V		0.6299			0.5031			0.4900			0.3659	
Statistic	**MARITAL STATUS**			**BMI**			**HOSPITAL**			**ALCOHOL**		
	DF	Value	Prob	DF	Value	Prob	DF	Value	Prob	DF	Value	Prob
Chi-Square	3	719.5360	<.0001	2	303.7197	<.0001	1	236.6876	<.0001	8	214.1047	<.0001
Likelihood Ratio Chi-Square	3	805.8298	<.0001	2	304.6617	<.0001	1	263.7490	<.0001	8	224.9365	<.0001
Cramer's V		0.3620			0.2383			0.2076			0.1976	
Statistic	**INCOME**			**EDUCATION**			**PHYSICAL EFFORT**			**SMOKING**		
	DF	Value	Prob	DF	Value	Prob	DF	Value	Prob	DF	Value	Prob
Chi-Square	5	178.7282	<.0001	5	145.3824	<.0001	3	92.3897	<.0001	3	77.1706	<.0001
Likelihood Ratio Chi-Square	5	179.1298	<.0001	5	147.2321	<.0001	3	94.9327	<.0001	3	78.6991	<.0001
Cramer's V		0.1804			0.1627			0.1297			0.1186	

[7] *<http://www.acastat.com/statbook/chisqassoc.htm>.*

Table 1 (continuation)

Statistic	SEX			FRUITS			VEGETABLES		
	DF	Value	Prob	DF	Value	Prob	DF	Value	Prob
Chi-Square	1	63.6577	<.0001	4	11.89	0.0239	4	5.0003	0.2873
Likelihood Ratio Chi-Square	1	63.6358	<.0001	4	11.45	0.0231	4	5.0063	0.2867
Cramer's V		–0.1077			0.0453			0.0302	

Source: EHIS 2015, created in SAS Enterprise Guide

be certain relationships between some factors. For example, a group of factors where contingency analysis has shown dependence on the analyzed variable (education, income, smoking, and alcohol consumption) can be determined by subjective perception of the health status of respondents in each category. Therefore, we will also assess the impact of individual factors through logistic regression, in which the impact of other relevant variables included in the model will be fixed.

We first considered the impact of all selected variables using the full regression model (see Table 2). We can see from the results of Table 2 that not all variables have a statistically significant effect on the dependent variable, so we decided to modify the model and gradually eliminate insignificant factors from the model by a stepwise regression method.

Table 2 Estimation of regression model expressing dependence of long-term health problem of person on selected factors (the full model)

Testing Global Null Hypothesis: BETA = 0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	3 151.8170	48	<.0001
Score	2 524.2487	48	<.0001
Wald	1 271.4345	48	<.0001
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Sub. Status	2	345.1817	<.0001
Drugs	2	251.9051	<.0001
Visit Doctor	2	55.9095	<.0001
Age	1	44.4643	<.0001
BMI	2	17.9255	0.0001
Employment	3	12.8506	0.0005
Physical Effort	3	6.8247	0.0777
Marital Status	3	8.5354	0.0362

Table 2 (continuation)

Effect	DF	Wald Chi-Square	Pr > ChiSq
Sex	1	4.1589	0.0414
Hospitalization	1	6.4816	0.0109
Education	5	4.6868	0.4553
Alcohol	8	9.8140	0.2783
Vegetables	4	7.6045	0.1072
Fruits	4	2.2136	0.6965
Income	5	2.8937	0.7164
Smoking	2	0.6401	0.7261

Source: EHIS 2015, created in SAS Enterprise Guide

The resulting adjusted model (see Table 3) contains ten statistically significant factors. The degree of influence of individual explanatory variables can be seen by the value of chi-square statistics. The existence of a long-term health problem for Slovak population in 2015 is mainly influenced by the subjective perception of the subject's difficulties, the possibility of affording prescribed drugs over the last 12 months, and age. To some extent, it is surprising for us to find out that the variables related to the lifestyle of the population (alcohol consumption and smoking) have been excluded from the model.

Table 3 Estimation of regression model expressing dependence of long-term health problem of person on selected factors (reduced model)

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	**Sub. Status**	2	1	1 794.5969	<.0001
2	**Drugs**	2	2	586.6429	<.0001
3	**Age**	1	3	250.8074	<.0001
4	**Doctor's Visit**	2	4	66.6645	<.0001
5	**Employment**	3	5	13.5674	0.0036
6	**BMI**	2	6	11.2212	0.0037
7	**Sex**	1	7	9.3152	0.0023
8	**Hospitalization**	1	8	5.9978	0.0143
9	**Physical Effort**	3	9	8.4544	0.0375
10	**Marital Status**	3	10	8.0824	0.0443

Source: EHIS 2015, created in SAS Enterprise Guide

To verify the significance of the model of dependence of long-term health problem of the Slovak population on selected factors, the plausibility test, score test and Wald test were used (see Table 4). For all three tests, the p-value was shown to be less than the commonly used significance level (0.05), so we can reject the hypothesis according to which all model parameters are zero. However, this result does not exclude the possibility of a zero value for any of the model parameters. In the second part of the output, there are three measures of model quality (Akaik's information criterion, Schwartz-Bayes criterion and logarithmic transformation the likelihood function), separately for the model with an intercept only and separately for the specially estimated logistic model (intercept and covariates). Since all of the above measures are lower in the logistics model, we consider it to be better than the model with an intercept only.

Table 4 Quality assessment of the reduced logistic regression model

Testing Global Null Hypothesis: BETA = 0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	3 123.7083	20	<.0001
Score	2 508.4425	20	<.0001
Wald	1 271.5884	20	<.0001

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	7 233.722	4 150.014
SC	7 240.305	4 288.268
–2 Log L	7 231.722	4 108.014

Source: EHIS 2015, created in SAS Enterprise Guide

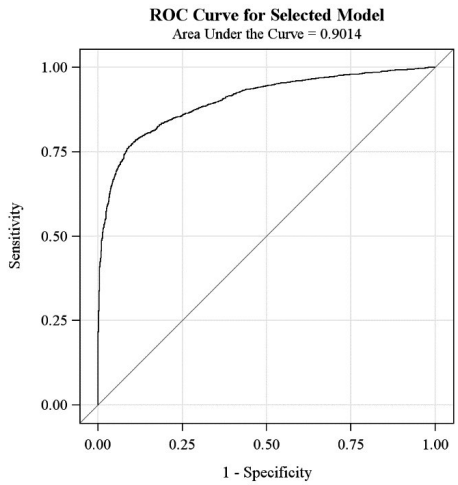
Table 5 Association between predicted probabilities obtained from the model of logistic regression of long-term health problem of Slovak population and observed values

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	90.1	**Somers' D**	0.803
Percent Discordant	9.9	**Gamma**	0.803
Percent Tied	0.0	**Tau-a**	0.388
Pairs	6 904 106	**c**	0.901

Source: EHIS 2015, created in SAS Enterprise Guide

In evaluating the model quality, we also used association measures (see Table 5) to assess the association between the predicted probabilities for the modeled variation of the dependent variable and the actual values of the dependent variable. As can be stated from column 2 of the Table 5, the proportion of matched pairs of observations is significantly higher than the proportion of opposite pairs, which indicates

Figure 1 ROC curve of logistic model of long-term health problem of Slovak population



Source: EHIS 2015, created in SAS Enterprise Guide

the quality of the model. The last column of the table contains the association measures values (Somers D, Goodman-Kruskal gamma, Kendall tau and c-statistics), which, with the exception of Kendall tau, are high, which is another argument in favor of the model accuracy.

The value of statistics c (0.901) can be represented graphically by using the ROC (see Figure 1) curve – its value is the area under the curve. As we can see, the curve is placed high above the diagonal of the square, so the quality of the model is confirmed.

In the next step, we used the unconditional maximum likelihood method to estimate model parameters. The results of the estimated parameters for each model category, point and interval estimates of the odds ratio for 2015, which we will use for the interpretation, are shown in Table 6. We will mainly focus on statistically significant variations of the variables compared to the reference

Table 6 Estimates of logistic model coefficients and odds ratios of long-term health problem of Slovak population

Analysis of Maximum Likehood Estimates		Coefficient		Odds Ratio Estimates		
Parameter	Effect	Estimate	Pr > ChiSq	Point Estimate	95% Wald Confidence Limits	
Intercept		0.1272	0.6975			
Age		0.0319	<.0001	1.032	1.023	1.042
Sex	Male	–0.1091	0.0106	0.804	0.680	0.951
	Female					
Marital Status	Single	–0.0958	0.3067	1.108	0.905	1.355
	Widower	0.3061	0.0436	1.655	1.120	2.447
	Divorced	–0.0123	0.9089	1.204	0.921	1.575
	Married					
Hospitalization	Yes	0.1784	0.0141	1.429	1.075	1.899
	No					
Visit Doctor	More than 12 months ago	0.0371	0.7466	0.542	0.453	0.648
	Less than 12 months ago	–0.6872	0.0009	0.263	0.143	0.484
	Never					

Table 6 (continuation)

Analysis of Maximum Likehood Estimates		Coefficient		Odds Ratio Estimates		
Parameter	**Effect**	**Estimate**	**Pr > ChiSq**	**Point Estimate**	**95% Wald Confidence Limits**	
Physical Effort	**Stand or sit mainly**	0.2084	0.0076	1.113	0.941	1.318
	Manual labour	0.1238	0.2681	1.023	0.779	1.344
	No work done	–0.4334	0.0086	0.586	0.382	0.900
	Moderate activity/ Walking					
BMI	**Underweight**	–0.1487	0.4149	0.953	0.557	1.633
	Overweight / obese	0.2497	0.0152	1.420	1.200	1.680
	Normal weight					
Drugs	**Yes**	0.7456	0.0001	6.177	3.426	11.135
	No	0.3296	0.0018	4.075	3.419	4.855
	Not needed					
Sub Status	**Neither good nor bad**	–0.1655	0.3628	6.876	5.524	8.558
	Bad / very bad	252 051	<.0001	77.674	28.486	211.797
	Good / very good					
Employment	**Unemployed**	–0.0407	0.6880	1.196	0.917	1.560
	Other	0.0379	0.7258	1.294	1.003	1.669
	Retired	0.2225	0.0846	1.556	1.141	2.123
	Employed					

Source: EHIS 2015, created in SAS Enterprise Guide

categories, where the *p*-value is less than the significance level of 0.1. The reference variations of each category are listed for each variation in the last empty line. All parameter interpretations are given under the ceteris paribus condition, and this will not be repeated for each individual interpretation given the scope of the article.

As we have already stated on the basis of the values in Table 1, the greatest impact on the long-term health problem suffered by the Slovak population is the variable General health state perceived by the respondent (Sub status). Overall, we can say, that this variable is statistically significant for one variation, with the category of good or very good general health being chosen as the reference category, given its most frequent occurrence. The probability of a long-term health problem of a person who perceives his general health condition as bad or very bad is up to 77.674 times higher than that in group of a persons with a good or very good general health condition. A generally perceived state of health, neither good nor bad, appears to be a statistically insignificant category.

Another statistically significant variable by both criteria is Respondent's ability to afford prescribed medication (drugs) over the last 12 months. Persons who cannot afford them are 6.177 times more likely to suffer from a long-term health problem compared to those who do not need medicines. Somewhat lower the probability of a long-term health problem was also observed for persons who, on the other hand, can afford medicines, 4.075 times higher than in those who do not need medicines. The two odds ratios presented are in line with our expectations: The absence of a prescribed medication is a strong indication of good health, and it is logical that in the case of prescribing drugs, those who can afford it, are in a better contition comparing with those, who cannot afford it.

The model results also confirmed the well-known fact that increasing age has a negative impact on health. If a person's age increases by a year, the probability of a risk of suffering from a long-term health problem is 1.032 times higher.

Our expectations were also confirmed by the variable Status of employment. In comparison with the reference category, there was a statistically significant difference (at the significance level of 0.1) only for the pensioner category – compared to the employed person, the chance of a long-term health problem is 1.556 times higher.

Analysis of variable a Doctor visit showed that those who visited a doctor less than 12 months ago had a 3.8-fold (1 / 0.263) lower statistically significant probability of risk of a long-term health problem than those who had never visited a doctor. This finding shows the importance of a doctor's visit also in terms of disease prevention.

The importance of a healthy lifestyle is confirmed by the influence of the body mass index. A person with a high weight or obesity is likely to suffer from a long-term health problem by up to 42% higher than a person with a normal weight.

For us, an interesting fact has been shown in the comparison of sexes. Although women's life expectancy is higher than that of men, this does not necessarily mean that they are generally healthier – we have found that the men´s risk of a long-term health problem is 1.244 times lower than women´s.

The group of people who have been hospitalized in the hospital for the last 12 months also proved to be a risk category. The probability of risk of suffering from a long-term health problem is 1.429 times higher than that of those who have not been hospitalized.

A person who does not perform any work tasks has a 1.7-fold lower risk of a long-term health problem than a person who usually walks or performs tasks with moderate physical exertion. This leads to the idea that physical exertion, both during and outside work, has a negative impact on health. On the other hand, it should be noted that the group of people who do not perform work tasks is formed predominantly by students or retired people living a healthy lifestyle. In assessing the impact of physical exertion in the performance of duties, there was also a statistically significant difference between those who are at work and those who are moderately physically stressed – such people are 1.1 times more likely to have a long-term health problem.

We chose a married person as the reference category of the marital status variable due to the ever-increasing importance of the harmonious family in Slovakia. The statistically significant difference was only in one category: widowed persons are likely to suffer from a long-term health problem 1.655 times higher. However, this situation needs to be seen in a broader context: the worst position of widowed persons is probably also related to the fact that they are often elderly.

DISCUSSION AND CONCLUSION

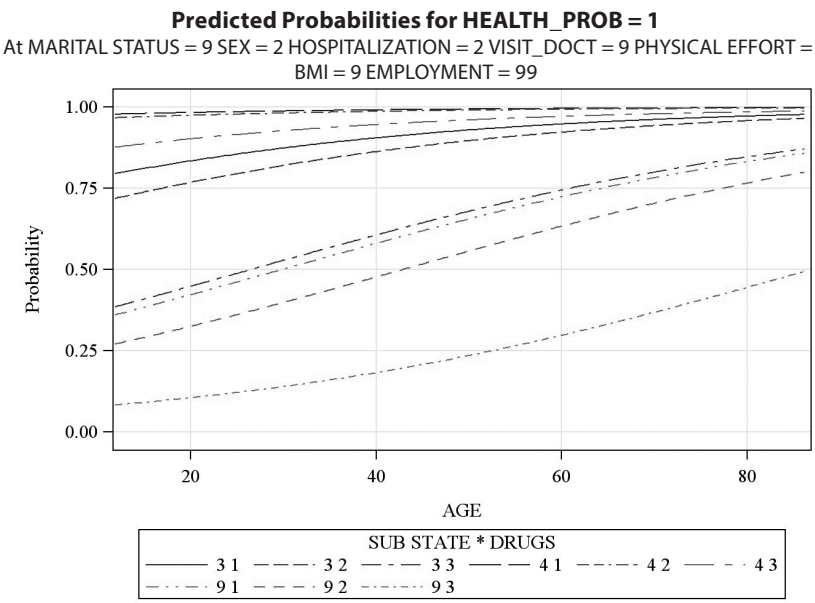
A new health policy for Europe from the WHO Regional Office for Europe Health 2020 underlines that its main objective is "to significantly improve the health and prosperity of the population, reduce the extent of health inequalities, strengthen public health and secure universal, fair, sustainable and high-quality health systems".

Concerning long-term health problems or diseases, in 2017 about one third (36.9%) of the EU28 population reported having suffered from these problems.[8] Up to 30.5% of people in the EU28 with the highest income (above fourth quintile) reported having a long-term illness or health problem, the equivalent share for people with a lower income threshold (first quintile) was up to 44.0%. While some researches suggest that health problems are more common for people with lower incomes, according to our results in Slovakia, income does not play such an important role. The results of the analysis showed that the most significant factor is the subjective perception of the subject's difficulties. In 2015, up to 60.7% of respondents perceived their health as very good or good, while only 14.2% of respondents perceived their health as bad or very bad.

In Figure 2 shows the simultaneous action of the three most important factors (general health status perceived by the respondent, the possibility to afford prescribed drugs and the person's age), while the other factors have been fixed at the reference levels. Under these conditions, it has been shown that with increasing age the probability of a person suffering from a long-term health problem increases. The results also confirmed the general fact that as the population is aging older, the risk of disease increases. The riskiest category consists of people who perceive their health as very bad or bad (SUB_STATE 4) and at the same time had (DRUGS 1) or had not (DRUGS 2) the ability to afford prescribed drugs over the last 12 months, combinations of variations appear to be not very significant. On the other hand, the least risk of a person suffering from a long-term health problem is among young people who perceive their health as very good or good (SUB_STATE 9) without needing health care (DRUGS 3).

Figure 3 highlights the importance of prevention in healthcare. The probability of a person suffering from a long-term health problem is the lowest if the person has been visited a doctor less than 12 months

Figure 2 Estimates of the risk of a person suffering from a long-term health problem depending on age, general perception and the possibility of affording prescribed drugs



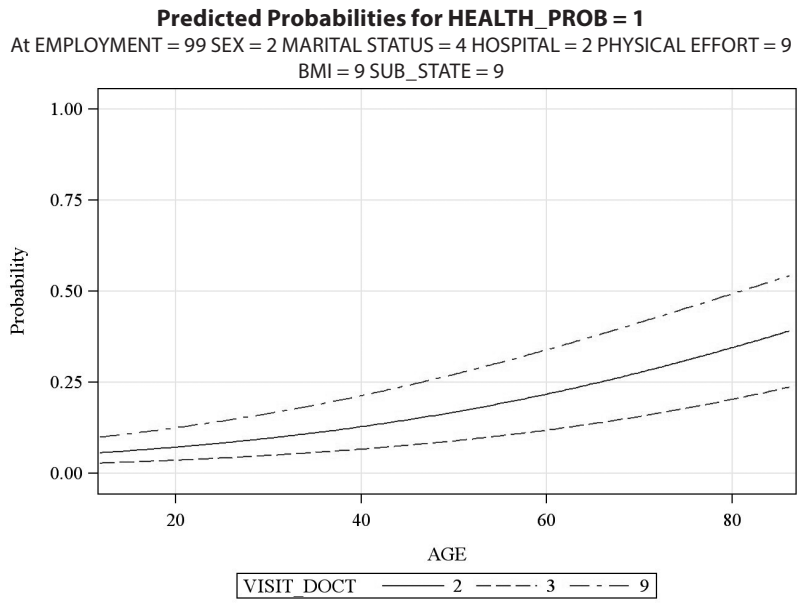
Note: For coloured figure see the online version of Statistika journal No. 4/2019.
Source: EHIS 2015, created in SAS Enterprise Guide

[8] *<https://ec.europa.eu/eurostat/statistics-explained/index.php/Quality_of_life_indicators_-_health>.*

ago. On the contrary, it is highest if the person has not visited the doctor at all, and if the person is over 80, the probability value is 50% or more.

The riskiest category in terms of all the factors analyzed can be considered an elderly man, a widower or pensioner who has perceived his condition as bad or very bad and has been hospitalized for the last 12 months, never visited a doctor, usually sits or stands, while he couldn't afford prescribed drugs for the last 12 months.

Figure 3 Estimates of the risk that a person will suffer from a long-term health problem depending on the doctor's visit



Note: For coloured figure see the online version of Statistika journal No. 4/2019.
Source: EHIS 2015, created in SAS Enterprise Guide

Analysis of association confirmed by using of Chi-square tests showed that the long-term health problem from which the Slovak population in 2015 suffered was significantly influenced by almost all selected categorical variables. The p-value of the tests is in all cases lower than the commonly used significance level.

Lifestyle (Beblavá, 2003) is a frequently discussed determinant that affects the health of the population, but according to our findings, variables such as fruit and vegetable consumption in the contingency analysis and also in the logistic regression model were proved to be insignificant. Moreover, in fixing the impact of other relevant variables included in our model, they have shown to be insignificant to alcohol consumption or smoking. Only the influence of the BMI factor and the Physical Exertion in the fulfillment of duties were significant. These findings are surprising to us; to some extent, they can be explained by the fact that the issues of consumption of vegetables and fruits about smoking were present, while the respondent's health problem lasts for at least half a year and it is not clear what the problem is and what causes it.

The strategic role of Slovak health care is to strengthen citizens' interest and responsibility for their own health, which can be achieved by informing them about the determinants affecting them. This paper provides, through the results of the present analysis, a list of potential factors that may affect health, while

quantifying their impact on the expression of the Slovak population, whether it suffers from a long-term health problem.

Especially nowadays it is important to realize that other factors affecting the health of the population (which go beyond the scope of the present analysis) are environmental. Understanding and assessing the impact of environmental factors on human health (both physical and mental) is a multidisciplinary approach. It depends mainly on the knowledge of the quality of the environment, from the internal environment (working and non-working), through the outdoor environment in urbanized units to the natural environment. Good environmental quality of man, which significantly effects his health, is a sum of good quality of air, water and food.

The World Health Organization is actively monitoring the impact of environmental factors on the occurrence of various types of diseases[9] and is actively seeking effective measures to improve the situation. However, it is necessary, especially now that global warming is objectively proven, to carry out relevant research on its impact on the health of the population in individual (and developed) countries and to take appropriate measures based on the findings.

ACKNOWLEDGMENT

The paper was supported by a grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic VEGA. Project: VEGA No. 1/0770/17 Availability and affordability of housing in Slovakia.

References

ALLISON, P. D. *Logistic Regression using SAS. Theory and Application.* 2nd Ed. North Carolina, USA: SAS Institute, 2012.

BAGLEY, SC., WHITE, H., GOLOMB, BA. Logistic Regression in the Medical Literature: Standards for Use and Reporting, with Particular Attention to One Medical Domain. *Journal of Clinical Epidemiology*, 2001, Vol 54, pp. 979–985.

BEBLAVÁ, E. et al. *Národná správa o ľudskom rozvoji SR 2001–2002* (in Slovak). Bratislava: PR 1, Centrum pre hospodársky rozvoj, 2003.

BLAXTER, M. *Health*. Cambridge: Polity Press, 2010.

BUNČÁK, M. et al. *Názory občanov na budúcnosť Slovenska* [online]. Bratislava: Ekonomický ústav Slovenskej akadémie vied, 2009, 96 p. [cit. 25.3.2019] <https://archiv.vlada.gov.sk/old.uv/data/files/5615.pdf>.

ČELEDOVÁ, L. AND ČEVELA, R. *Výchova ke zdraví: Vybrané kapitoly* (in Czech). Prague: Grada Publishing, a. s., 2010.

DHAND, N. K. UniLogistic: A SAS Macro for Descriptive and Univariable Logistic Regression Analyses [Code Snippet 1] [online]. *Journal of Statistical Softwar*e, 2010, Vol 35. <https://www.jstatsoft.org/article/view/v035c01>.

EVANS, R. G., BARER, M. L., MARMOR, T. R. eds. *Why are some people healthy and others not? The determinants of the health of populations.* New York: Routledge. 2017.

HILBE, J. M. *Practical Guide to Logistic Regression.* Boca Raton: Chapman & Hall/CRC, 2016.

HOSMER, D. W. AND LEMESHOW, S. *Applied Logistic Regression.* New York: John Wiley & Sons, 2013.

KATAMURI, S. S. *Predictive modeling with SAS Enterprise Miner. Practical Solutions for Business Applications.* 3rd Ed. North Carolina, USA: SAS Institute Inc., 2017.

MARMOT, M. Social determinants of health inequalities. *The lancet*, 2005, Vol 365, Iss. 9464, pp. 1099–1104.

ŠOLTÉS, E. *Regresná a korelačná analýza s aplikáciami* (in Slovak). Bratislava: Iura Edition, 2008.

OECD/EUROPEAN OBSERVATORY ON HEALTH SYSTEMS AND POLICIES. *Slovensko: Zdravotný Profil Krajiny 2017* [online]. State of Health in the EU, OECD Publishing, Paris/European Observatory on Health Systems and Policies, Brussels, 2017. [cit. 20.8.2019] <https://www.oecd-ilibrary.org/social-issues-migration-health/slovensko-zdravotny-profil-krajiny-2017_9789264285408-sk;jsessionid=j9rRgLzQUXFVuXb72icZzaqX.ip-10-240-5-163>.

WHO. *Zdravie 2020: a European policy framework and strategy for the 21st century.* Copenhagen, WHO Regional Office for Europe, 2013.

[9] *<https://apps.who.int/iris/bitstream/handle/10665/204585/9789241565196_eng.pdf?sequence=1&isAllowed=y>.*

ANNEX

Table A1 Description of input variables from EHIS database

Name of the artificial variable	Original variables in EHIS	Variations		Position in EHIS
HEALTH PROB	Long-term health problem: Suffers any illness or heath problem	1	Yes	ST02
		2	No	
DRUGS	The respondent could not afford any prescription medication within past 12 month	1	Yes	CR24
		2	No	
		3	No healthcare need	
AGE	Respondent's age (number of completed years)	15–80	Persons age of 80 and over are listed as 80	HH04
MARITAL STATUS	Legal marital status	1	Single	RE03
		9*	Married	
		3	Widowed	
		4	Divorced	
EDUCATION	Highest level of education	1	Primary education	RE05
		2	Secondary education	
		3	Secondary diploma	
		4	Post-secondary education	
		5	Undergraduate education	
		6	Graduate and post graduate education	
SUB STATUS	Respondent's general heath status: How person perceives his/her own health	9*	Very good or good	ST01
		3	Neither good nor bad	
		4	Very bad or bad	
HOSPITALIZATION	Hospital stay within past 12 months	1	Yes	CR01
		2	No	
EMPLOYMENT	Respondent's employment status	99*	Employed or self-employed	RE06
		20	Unemployed	
		31	Others	
		32	Retired	

Table A1 (continuation)

Name of the artificial variable	Original variables in EHIS	Variations		Position in EHIS
INCOME	Net monthly equivalent household income	1	Under 1st quintile	HH06
		2	Between 1st quintile and 2nd quintile	
		3	Between 2nd quintile and 3rd quintile	
		4	Between 3rd quintile and 4th quintile	
		5	Above 4th	
VISIT DOCTOR	Last visit to general practice or family doctor	9*	Never	CR06
		2	Over 12 months	
		3	Less than 12 moths	
SMOKING	Habits, in terms of smoking (Are you a smoker?)	1	Yes, daily	DT15
		2	Yes, occasionally	
		3	No	
PHYSICAL EFFORT	Physical activity while accomplishing tasks	1	Mostly sitting or standing	DT03
		9*	Mostly walking or doing tasks with moderate physical activity	
		3	Mostly hard manual labour	
		4	No work done	
SEX	Sex of respondent	1	Male	HH03
		2	Female	
BMI	Body Mass Index BMI = weight (kg) / height $(m)^2$; calculated only in adults (18 years and over)	1	Underweight BMI < 18,5	
		9*	Normal weight 18,5 ≤ BMI < 25	
		3	Overweight 25 ≤ BMI < 30 obesity BMI ≥ 30	
FRUITS	Frequency of fruit consumption	1	Once or twice per day	DT11
		2	4 to 6 times per week	
		3	1 to 3 times per week	
		4	Less than once a week	
		5	Never	

Table A1 (continuation)

Name of the artificial variable	Original variables in EHIS	Variations		Position in EHIS
VEGETABLES	Frequency of vegetables or salads consumptions	1	Once or twice per day	DT13
		2	4 to 6 times per week	
		3	1 to 3 times per week	
		4	Less than once a week	
		5	Never	
ALCOHOL	Frequency of alcohol consumption	1	Every day or almost every day	DT19
		2	5 to 6 days per week	
		3	3 to 4 days per week	
		4	1 to 2 days per week	
		5	2 to 3 days per month	
		6	Once a month	
		7	Less than once a month	
		8	No alcohol within past 12 months, because I quit drinking alcohol	
		9	Never, or only few drinks throughout the life	

Note: * the highest variation number is always selected as the reference category.
Source: EHIS 2015, created in SAS Enterprise Guide

Risk Premium Prediction of Motor Hull Insurance Using Generalized Linear Models

Marek Strežo[1] | *University of Economics in Bratislava, Bratislava, Slovakia*
Vladimír Mucha[2] | *University of Economics in Bratislava, Bratislava, Slovakia*
Erik Šoltés[3] | *University of Economics in Bratislava, Bratislava, Slovakia*
Michal Páleš[4] | *University of Economics in Bratislava, Bratislava, Slovakia*

Abstract

Pricing is a quite complex endeavour, understood as a process with beginning and end where several different tasks have to be executed in a certain order. Set the price for some individual policy can be considered an art, taking into consideration various features of policyholder or the insured object. Actually, approach performed by insurance companies, is necessary to apply different premiums depending on the degree of risk because of presence of heterogeneity within insurance portfolio, which could lead to the appearance of asymmetric information.

The aim of this paper is to present the methodology of segmented pricing model with generalized linear models, known as GLMs, for setting the risk premium. Nowadays, the GLMs are widely recognized as the industry standard method for pricing motor, the other personal lines and the retail insurance in the European Union.

Keywords

GLMs, Poisson regression, Gamma regression, risk premium, motor hull insurance

JEL code

C21, G22

INTRODUCTION

Actuaries use many statistical methods to measure risk in process of setting the risk premium. Practically the most widely method used in practise is the regression analysis. Linear regression had been applied until the 1980s using various transformations of predicted variable. Nowadays, generalized linear models or GLMs for short are preferably applied. Restrictions in linear regression are discussed by (Anderson et al., 2007). The comprehensive reference for GLMs in actuarial field is (McCullagh and Nelder, 1989; Fahrmair and Tutz, 1996; Mildenhall, 1999; Kaas et al., 2001). Valecký (2017, p. 451) states that more

[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: marek.strezo@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: vladimir.mucha@euba.sk.
[3] Department of Statistics, Faculty of Economic Informatics, University of Economics in Bratislava, Dolnozemská cesta 1/b, 852 35 Bratislava 5, Slovakia. E-mail: erik.soltes@euba.sk.
[4] 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: michal.pales@euba.sk.

applications of the GLMs occurred mostly after the 1990s when the insurance market was being deregulated in many countries and the models were used to perform tariff analysis. Even though the GLMs are used mainly in the non-life insurance practise, Haberman and Renshaw (1996) referred to their wide use in the actuarial practice, including life insurance (survival data analysis – SDA, health insurance modelling and mortality modelling). As David (2015) indicates, the GLMs allow modelling of non-linear behaviour and non-Gaussian distribution of residuals, which is very useful non-life insurance analysis. A random component (error term) in an ordinary linear model is assumed to be normally distributed. However, when the claim frequency (count of the claims per exposure) and the claim severity (average cost per claim) are modelled this condition is not fulfilled. For that reason, the GLMs are suitable for analysis with non-normal data, i.e. insurance data because the error term can follow the number of different distributions from the exponential dispersion family – EDF, which generalizes normal distribution used in the linear models. The Poisson distribution belongs to this family and represents the main tool for the claim frequency modelling meanwhile Gamma distribution allows econometric modelling of the claim costs (Ewald and Wang, 2015) and (Duan et al., 2018). It might be considered using a Tweedie model to analyze the risk premium directly (see Xacur and Garrido, 2015; Frees et al., 2016; Jörgensen and Souza, 1994).

In general, two approaches are commonly used to calculate the risk premium in the non-life insurance. In the first case, the risk premium is modelled directly. The second case describe the standard GLMs analysis with separated analysis for the claim frequency and severity. Goldburd et al. (2016) point out the reason for this separation where the claim frequency is more stable than the claim severity and much more predictive factors are associated with the claim frequency. Such a separate analysis represents greater accuracy and offers deeper insights to the risk w.r.t regression coefficients.

Here, both the claims count, and the claims amount are assumed to be independent in case of the separate claim frequency and claim severity analysis. When this fundamental assumption is not fulfilled, authors Shi et al. (2015) or Garrido et al. (2016) discuss about this problem. Charpentier and Denuit (2005) also prefer separate analyses for claim frequency and claim severity as the benefit of such approach is visible in fact that both models (frequency and severity) can be affected by different various factors. Mentioned facts give us the reason why to choose separate analysis in the GLMs for calculating the risk premium in motor hull insurance in Slovakia.

The GLMs are an efficient and reliable tool used in various fields of predictive modelling. According to (Xie and Lawniczak, 2018, p. 2) the main reason for the prevalence of GLMs is that it enables a simultaneous modelling of all possible risk factors as well as the determination of the retention of risk factors in the model.

The main effort of this paper is not only to estimate the claim frequency and claim severity and then set price of transfer risk from the insured to an insurer, but also to identify relevant risk factors as well as to quantify their impact in the claim frequency, claim severity and also on the expected loss per exposure.

Data on which the research was based are real and comes from an unnamed insurance company operating in the Slovak insurance market. All calculations in this paper have been realized in R environment (R Core Team, 2019) using *glm()* function and packages *data.table* (Dowle et al., 2015) and *MASS* (Venables et. al., 2002).

1 METHODS OF ANALYSIS

The expected loss (also known as a risk premium) consists of the claim frequency and claim severity that are in the multiplicative relation:

$$Risk\ Premium = Frequency \cdot Severity \tag{1}$$

The frequency refers to the number of claims that an insurer anticipates will occur for a specific risk over a given time period. The severity represents the average cost of claims for specific risk. This article focuses on separate modelling of the claim frequency and claim severity using generalized linear models and determining the risk premium. This part of the paper provides a brief description of the methodology of sophisticated mathematical and statistical methods associated with the GLMs.

1.1 Theoretical framework of Generalized linear models

Generalized linear models include a wide set of statistical models consisting of three keystone elements – random component, linear predictor and the link function.

A *random component* refers to the conditional distribution of the response variable Y given the values of the explanatory variables in the model. Nelder and Wedderburn (1972) present the basics of the GLMs theory and declare that distribution of Y with independent observations y_i ($i = 1, 2, \ldots, n$) is a member of an exponential dispersion family. Exponential dispersion family, shortly EDF, has the probability density function in the following form:

$$f(y_i) = c(y_i, \phi) \exp\left\{\frac{y_i\theta_i - a(\theta_i)}{\phi}\right\}, \tag{2}$$

where θ_i and ϕ are the parameters, θ_i is called canonical or natural parameter and ϕ is a dispersion parameter (Agresti, 2015; Kafková and Křivánková, 2014). So called cumulant function $a(\theta_i)$ is assumed twice differentiable, where the first derivative is invertible. EDF includes the univariate Bernoulli, binomial, Poisson, geometric, Gamma, normal, inverse Gaussian, lognormal, Rayleigh, and von Mises distributions (Forbes et al., 2011).

The claim severity is modelled by two commonly used distributions the Gamma and inverse Gaussian distribution. Both these distributions are right-skewed with a lower bound at zero. According to Goldburd et al. (2016) inverse Gaussian compared to the Gamma distribution has a sharper peak and a wider tail and is therefore appropriate for the situations where the skewness of the severity curve is expected to be more extreme.

The claim frequency is modelled by the GLMs with Poisson noise. Some members of EDF such as Poisson and Bernoulli distribution have the distribution determined by the mean. When fitting models to data with binary or count dependent variables, it is common to observe that variance exceeds and anticipated by the fit of the mean parameters. This phenomenon is known as overdispersion (Edward, 2010). One way to check for and deal with it is to run negative binomial distribution or overdispersed Poisson distribution (Valecký, 2016; Ohlsson and Johansson, 2010). There are also several probabilistic models available to explain this phenomenon, depending on the application on hand. For a more detailed inventory see McCullagh and Nelder (1989).

A *linear predictor* is a linear function of the regressors:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \text{ or } \eta_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \tag{3}$$

where:

$\boldsymbol{\beta}$ is $p \times 1$ vector of model parameters ($p = k + 1$) including intercept β_0 and the regression coefficients $\beta_j (j = 1, 2, \ldots, k)$,

$\mathbf{X}$ is $n \times p$ matrix of the regressors (known from the classical regression) and x_{ij} is i-th observation of j-th regressor X_j.

The regressor can be expressed as quantitative explanatory variable, transformation of quantitative explanatory variable, e.g. polynomial regressor, dummy variable (coding the particular categorical variable), interaction, etc. (see Wooldridge, 2013).

The *link function* $g(\cdot)$ is strictly monotone and twice differentiable. This fundamental object links the mean of the response variable to the linear predictor through:

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta} \text{ or } g(\mu_i) = \eta_i\,, \tag{4}$$

where:
$\boldsymbol{\mu} = E(\mathbf{y})$ or $\mu_i = E(y_i)$,
$\mathbf{y}$ is $n \times 1$ vector of observations of target variable Y (called also response variable, explained variable or dependent variable),
$\boldsymbol{\mu}$ is $n \times 1$ vector of expected values of the elements of $\mathbf{y}$.

The link function that transforms μ_i to the natural parameter θ_i of distribution from exponential family is called canonical (or natural) link function (Agresti, 2015; Fox, 2015; Littell et al., 2010).

A maximum likelihood method is used to estimate the regression parameters $\boldsymbol{\beta}$ in Formula (4) (De Jong and Heller, 2008; Littell et al., 2010). As a result of this method is system of equations:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{X})\boldsymbol{\beta} = \mathbf{X}^{\mathrm{T}}\mathbf{W}\mathbf{y}^{*}, \tag{5}$$

where:
$\mathbf{W} = \mathbf{D}\mathbf{V}^{-1}\mathbf{D}$, whereby $\mathbf{V} = diag[\phi \cdot Var(\boldsymbol{\mu})]$ and $\mathbf{D} = diag\left[\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}\right]$ is $n \times n$ diagonal matrix whose elements are derivates of the elements of $\boldsymbol{\eta}$ with respect to $\boldsymbol{\mu}$ and $Var(\boldsymbol{\mu})$ is a covariance matrix of $\boldsymbol{\mu}$.
$\mathbf{y}^{*} = \hat{\boldsymbol{\eta}} + \mathbf{D}^{-1}\,(\mathrm{y} - \boldsymbol{\mu})$ [more detailed in (Littell et al., 2010; McCullagh and Nelder, 1989)].

We note that for the normal error model is $\mathbf{V} = \sigma_{\varepsilon}^{2}\mathbf{I}$ where $\mathbf{W}$ is the unit matrix and system (5) is reduced to the well-known system of normal equations, that we can estimate parameters of classical linear regression model (Agresti, 2015; Littell et al., 2010). In general, the system of equations from (5) is nonlinear in $\hat{\boldsymbol{\beta}}$, therefore the iterative methods are used for solving nonlinear equations such as Newton-Raphson method using a Hessian matrix itself and Fisher scoring method which uses expected values of Hessian matrix (Allison, 2012; Agresti, 2015).

1.2 Assessment of impact of explanatory variables on target variable and model selection

After estimating the generalized linear model, it is important to verify its statistical significance and verify if influence of the individual explanatory variables on probability target variable is significant. The significance of model is revealed by zero-hypothesis test $\boldsymbol{\beta} = (\beta_1\ \beta_2\ \ldots\ \beta_k) = \mathbf{0}^{\mathrm{T}}$ against an alternative hypothesis – at least one regression coefficient should not be zero, while three different chi-square statistics are prevalently used (Likelihood ratio, Score statistics, Wald statistics). Allison (2012) discusses differences between mentioned statistical methods and notes that in the large samples, there is no reason to prefer any of these statistics and they will be quite close in value.

In order to validate the significance of the explanatory variable influence, a Wald test is used. It tests the zero-hypothesis showing that the respective explanatory variable does not affect the probability of occurrence of explored event. To verify hypothesis, Wald statistic

$$Wald = \hat{\boldsymbol{\beta}}^{\mathrm{T}} \cdot \hat{\mathbf{S}}_{\mathrm{b}}^{-1} \cdot \hat{\boldsymbol{\beta}} \tag{6}$$

is used, where $\hat{\boldsymbol{\beta}}$ is the vector of regression coefficients estimates that stand at dummy variables for the respective factor (categorical explanatory variable) and $\hat{\mathbf{S}}_{\mathbf{b}}$ is the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$. Wald statistic has asymptotically χ^2 distribution with degrees of freedom equal to the number of parameters estimated for a given effect. A special case of the test above is the Wald test, which verifies the statistical significance of one regression coefficient. In this case Wald statistics is asymptotically distributed as χ^2 with 1 degree of freedom. The test statistic has an equation:

$$Wald = \left(\frac{\hat{\beta}_j}{s_{\hat{\beta}_j}}\right)^2, \tag{7}$$

where $s_{\hat{\beta}_j}$ is an estimated standard error of the j-th estimated coefficient.

When the process of the model building starts, there is a wide set of potential regressors and not all of them have significant impact on the dependent variable. It is obvious to use methods for variable selection, namely, the stepwise regression (see Draper and Smith, 1981; Hebák et al., 2005). In the stepwise regression, the selection procedure is automatically performed by statistical packages. In the practical part, it is used one of the main approaches of the stepwise selection known as backward elimination (see Agresti, 2015).

To evaluate how well the model fits the experience criteria AIC (Akaike information criterion) and BIC (Bayesian information criterion) are used. These measures are based on logarithmic transformation of the likelihood function (see Kim and Timm, 2006; Agresti, 2015). Preferred model is considered have with the lowest AIC and BIC, respectively. As state (De Jong and Heller, 2008, p. 63) BIC applies a greater penalty for the number of the parameters. When number of observations is large, as it is in most of cases of insurance data sets, the BIC tends to select the model which most of analysts consider too simple. In this case the AIC is preferable.

2 DATA PROCESSING AND MODEL BUILDING

In this part, we demonstrate practical usage of GLMs in actuarial practice which have been described in previous sections of this paper. We will try to set price of a non-life insurance policy, taking into consideration various properties of the insured object and policyholder as well. In this empirical study, we will go through models for short-term insurance schemes based on the Slovak market´s conditions. The study in this paper works with a very basic feature of the portfolio of risks – heterogeneity, which means that risks generate different values of claims. Consequently, charging each policy with the same premium (flat rate) is both unjust and uncompetitive. Therefore, we will try to classify each risk into the homogeneous risk groups where the *i*th risk has the same risk premium. Basic assumption that will give foundation to our statistical models is policy independence. This means that independence between random variables $Y_1, \ldots, Y_n$ is made in modelling the value of single claims and in the number of claims as well. Presented frequency-severity models will decompose the aggregate claim amount for a single risk into two parts, where the frequency part examines the number of claims by Poisson regression, the severity part by the GLMs Gamma regression. The R software will be used to calculate and analyse the results of these different multiplicative models.

2.1 Motor hull insurance data and descriptive analysis

Before the modelling it is useful to provide certain preliminary analyses, such as data checks, identification of observations with negative claim counts, zero or negative exposures, etc. The portfolio D consists of $n = 91\,685$ car insurance policies for which we have features information $\mathbf{x}_i \in \mathbf{X}$ and exposure - years at risk information, denotes as $v_i \in [0; 1]$, for $i = 1, 2, \ldots, n$. Nature of the data comprises a Slovak motor

hull insurance with corresponding claim sizes and counts for calendar year 2018. Now, we briefly describe the list of variables in our dataset D:

- *ID profile*: represents unique identifier; policy number;
- *Claim.No*: number of claims which occurred on each policy;
- *Claims*: total claim cost per every policy;
- *Policyholder_Age*: the owners age in years, between 0 and 91, non-linear continuous feature portioned as nominal categorical variables;
- *Vehicle_Age*: age of cars in years, narrowly defined categorical factor;
- *Policy_Exposure*: the exposure is widely applied in non-life pricing. In order to illustrate this concept, we take GLM for the frequency claims. Policies that begin in a given calendar last year until the end of the coverage period. This period is longer for annual contracts than for short-term policies, which results in a higher number of expected claims for longer contracts. Therefore, it is necessary to include this effect in the model as exposure with the use of weights;
- *Region*: regional divisions of Slovakia according to the company´s internal policy, categorical feature with 11 labels;
- *B-M Class*: bonus class, taking values for bonus from 0 to 7 and for malus from 1 to 2, with the reference level 0;
- *Vehicle_Engine_Volume*: represents engine volume of car, continuous feature;
- *Total Sum_Insured (TSI)*: specified car value which represents the upper limit of what would be pay out for the claim;
- *Power*: power of car, non-linear continuous feature split as categorical variables;
- *Payment_Frequency*: expresses the frequency of premium payments (payment option is 1,2,4 and 12);
- *Vehicle_Weight*: weight of car, non-linear continuous feature portioned as nominal categorical variable;
- *Policyholder_entity*: categorical variable which can obtain 2 values;
- *Mileage_per_Year*: total length in miles per given period (calendar year);
- *Deductible_group*: policyholders can choose the excess at level that exploits reduction in premium, categorical variable.

In the next step, we provide a short summary of the data D. Since the policy number is not considered to be an explanatory variable, we drop this feature from all our next considerations.

Table 1 Split of portfolio w.r.t. number of claims and the severity claims

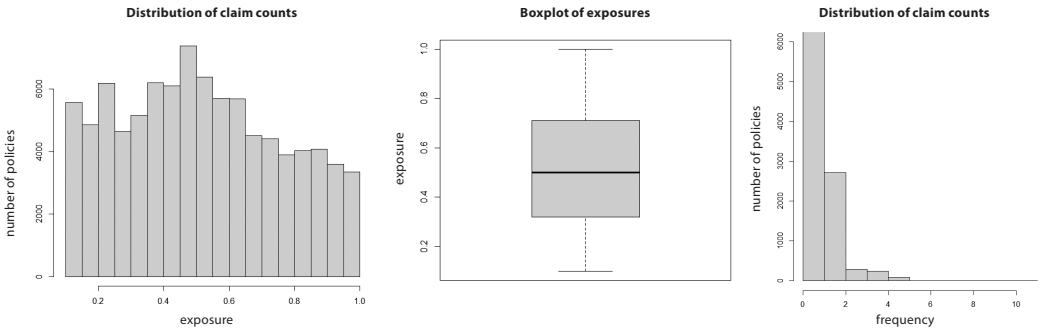
# of claims	0	1	2	3	4	5	6	7	8	9	10	11
# of policies	79 667	8 667	2 715	285	244	93	7	2	2	1	1	1
# of policies in %	86.89	9.45	2.96	0.31	0.27	0.10	0.01	0	0	0	0	0
Total exposures	40 243	5 256	1 741	182	171	65	4	1.60	1.63	0.96	0.98	0.75

Source: Own construction

In Table 1 you can be see the distribution of the observed claims $(N_i)_{1\leq i\leq n}$ across the whole portfolio of our dataset D with the attributable policy exposure. We note that 86.89% of the policies don't have a claim. In practice, this claim imbalance can often causes difficulties in the model calibration. Next, we provide helpful preliminary analysis to determine distribution of the key data items to investigate any problems or unfamiliar features prior the modelling. This concerns the distributions for claim counts and for the claim severity. Typical claim distribution is shown in Figure 1 (lhs) and in Figure 2.

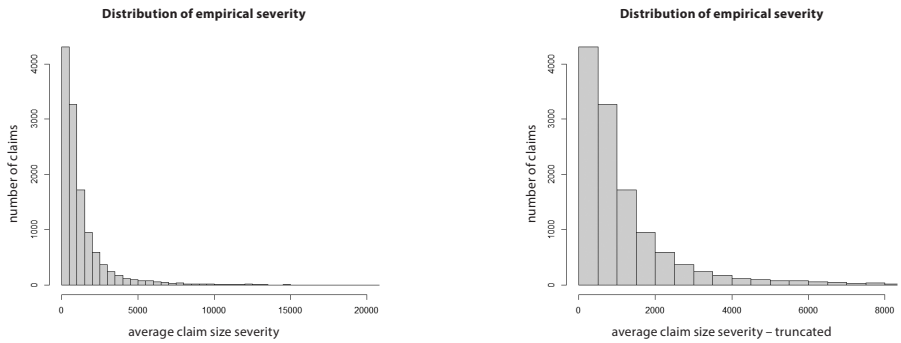
The years at risk *Policy_Exposure* is illustrated in Figure 1 (lhs and middle). For this feature, we have following properties $\min_i v_i = 0.1$ and $\max_i v_i = 1$, that is, minimal time insured in our portfolio is 36.5 days and the maximal insurance time is 1 year. The average insured time is represented as $\sum_i v_i/n = 0.3461$, which corresponds to 126 days. Median time insured is 183 days and only 35.33% of policies are in force the whole year.

Figure 1 Histogram (lhs) and (middle) boxplot of the years at risk, (rhs) histogram of frequencies of whole portfolio dataset *D*



Source: Own construction, customized in R

Figure 2 Histogram of empirical severity (lhs) and histogram of truncated empirical severity over the interval (0; 8 000]



Source: Own construction, customized in R

The heavy tail of the severity distribution is obvious. The average claim size of whole portfolio is 1 300.87 EUR. In practise, when the severity is modelled, it is often useful to provide a large loss threshold to certain claims. This helps to assess the possible thresholds. Presented study does not work with large claims in the dataset *D*.

Before modelling, it is necessary to investigate if and how the explanatory variables should be categorized, and if some of variables should be modelled as the continuous component. Some features used in our models are (highly) non-linear which does not support the log-linear assumption. This is certainly true for the components like policyholder age, vehicle age, vehicle power and volume, etc. Our approach, for these continuous feature components is to group values into intervals, where treated values in the same interval are identical. This approach is based purely on the expert judgement. Next Table 2 shows final predictors with chosen categorization used in presented risk frequency-severity model. In GLMs, it is advised to select the level with maximum exposure as reference for each predictor, because

it minimizes the standard errors of parameter estimates. The sign ® in Table 2 refers to the reference level of the particular predictor.

Table 2 The predictors used in the final step of the frequency and the severity modelling

	Categorical Predictors	*# of Class*	*Multi-level factors*
FREQUENCY MODEL	*Payment_Frequency*	4	1, 2, 4®, 12
	B-M Class	7	*B*0®, *B*1–*B*3, *B*4, *B*5, *B*6, *B*7, *M*1–*M*2
	Region	5	*R*01–*R*04–*R*06–*R*09–*R*11, *R*02–*R*05–*R*10, *R*03®, *R*07, *R*08
	Policyholder_Age	9	18–23, 24–27, 28–31, 32–37®, 38–44, 45–53, 54–61, 62+, *LE*
	Vehicle_Age	9	0, 1, 2, 3, 4®, 5, 6, 7, 8+
	Vehicle_Power	3	0–76®, 77–112, 133+
	TSI	6	0–5 000, 5 001–10 000®, 10 001–15 000, 15 001–25 000, 25 001–35 000, 35 001+
	Vehicle_Engine_Volume	5	0–1 354, 1 355–1 397®, 1 398–1480, 1 481–1 750, 1 751+
	Mileage_per_Year	3	0–15 000®, 15 001–30 000, 30 001+
	Deductible	4	*No Deductible*, ≤ 1%®, ≤ 2%, > 2%
SEVERITY MODEL	*B-M Class*	6	*B*0®, *B*1–*B*2–*B*3, *B*4, *B*5, *B*6–*B*7, *M*1–*M*2
	Region	4	*R*01–*R*07–*R*08, *R*03–*R*04®, *R*02–*R*05–*R*10, *R*06–*R*09–*R*11
	Policyholder_Age	8	18–26®, 27–32, 33–37, 38–45, 46–55, 56–61, 62+, *LE*
	Vehicle_Age	6	0, 1, 2, 3, 4, 5+®
	Vehicle_Power	4	0–80®, 81–95, 96–124, 125+
	TSI	5	0–5 000, 5 001–10 000®, 10 001–15 000, 15 001– 25 000, 25 001+
	Mileage_per_Year	4	0–5 000, 5 001–10 000, 10 001–13 000®, 13 001+

Source: Own construction

2.2 Model building and validation

In previous chapter we started with descriptive statistics on the motor hull portfolio and explanatory data to gain insight on behaviour of the dataset with respect to the number of claims and its subsets with respect to the explanatory variables. As already stated in the last chapter, we will only use 10 predictors in our tarification model; an intercept will be included.

The most frequently used is the backward elimination process, where on intends to reduce the saturated model to a complete model, meaning a model with the best explanatory terms. To begin, all possible variables are included in the model and then the stepwise terms are excluded, every time the term which p-value is bigger than a 5% significance level. The other option is to use the Wald test to check the statistical significance of predictors.

Following Table 3 shows performed Wald test to check relevance of the explanatory variables for final proposed risk models to explain the response variable. It was tested based on the relation (6). Variables *Vehicle_Weight* and *Policyholder_entity* were excluded from both frequency and severity models. Moreover, variables Deductible and *Payment_Frequency* were also removed from the severity model because a p-value of it is lower than a predefined level 5%. These variables do not improve significantly the quality of this model.

Table 3 Wald test of significance of explanatory variables for risk models

Predictors	FREQUENCY_MODEL			SEVERITY_MODEL		
	df	Chisq	Pr(>Chisq)	df	Chisq	Pr(>Chisq)
Intercept	1	702.712	< 2.2e-16	1	17 802.802	< 2.2e-16
Payment Frequency	3	56.432	3.397e-12	-	-	-
B-M Class	6	577.505	< 2.2e-16	5	57.727	3.580e-11
Region	4	327.139	< 2.2e-16	3	16.361	0.0009
Policyholder Age	8	208.724	< 2.2e-16	7	64.065	2.317e-11
Vehicle Age	8	205.724	< 2.2e-16	5	129.799	< 2.2e-16
Vehicle Power	2	24.171	5.64e-06	3	37.222	4.130e08
TSI	5	106.849	< 2.2e-16	4	195.684	< 2.2e-16
Vehicle Engine Vol.	4	23.757	8.934e-05	-	-	-
Mileage per Year	2	60.868	6.062e-14	3	40.405	8.744e-09
Deductible	3	1 621.859	6.854e-12	-	-	-

Source: Own construction, customized in R

When the models were constructed and parameters were estimated (column *Estimate* in Table 4), their significance was tested by Wald test (column p-value in Table 4) defined by (7).

The estimated regression models in Table 4 will be discussed in section 3 but let us first consider the degree of multicollinearity. In our observational study we have many explanatory variables where some relations among them may imply perfect linear combinations with other predictors. In practise, presence of the multicollinearity, regression estimates are unstable and have high standard errors. Variable has a little partial effect because it is predicted well by others. Excluding a nearly redundant predictor can help to reduce standard errors of other estimated effects. To identify potential problem of the collinearity among the explanatory variables we chose according to (Agresti, 2015) variance inflation factors (*VIF*) which measure the inflation in the variances of parameter estimates due to collinearities in the model. A VIF_j of 1 means that there is no correlation among the j-th predictor and remaining predictor variables, and hence the variance of β_j is not inflated at all. These calculations are straightforward and easily comprehensible; if the value of *VIF* is higher than 5 there is a problematic multicollinearity.

In case of this empirical study, the backward selection of variables could produce inconsistent results, variance partitioning analyses may be unable to identify unique sources of the variation, or the parameter estimates may include substantial amounts of uncertainty. In our proposed risk models, we didn't find any *VIF* value higher than 5, that is, no issue with this task.

3 RESULTS AND DISCUSSION

In this part, we present the process results of establishing the risk premium. We follow the standard process in GLMs analysis by separate analyses for the claim frequency and the claim severity. The authors (Ohlsson and Johansson, 2010) state some logical reasons for this separation. In our dataset D, we have an information about the number of claims and the claim costs on policy level with the duration of policy in force measured in years. In the Table 3 are presented the estimated regression coefficients (designated as Estimate) for each category of both proposed risk models, that includes all effects that explain the variation of the claim frequency and costs.

To illustrate, we give an interpretation of the value denoted as $e^{Estimate}$ shown in Table 4, for example, within the *Policyholder* age variable for the *Frequency model*. From the data in this table, we find that the age of the vehicle owner is a significant factor affecting the frequency or the expected number of claims during the year, and as the age of the owner decreases this frequency. Based on the relations (3) and (4) it is possible to formulate the following statements. The most risk category in the policyholder age is between the ages of 18 and 23 ($e^{Estimate} = e^{0.2959} = 1.3443$. For which the expected (average) number of claims during the year is 34.43% greater than in the reference category of 32 to 37 years, and up to 68.16% (1.3443 / 0.7994) greater than in the least risk category 62+. The above statements are based on the assumption that the other factors incorporated in the regression frequency model are at the same level (ceteris paribus). If the owner of the vehicle is a legal entity (LE), the expected number of claims during the year is approximately at category of 28 to 31 years, more precisely 8.2% higher than in the reference category.

Table 4 Analysis of parameter estimates in the risk models

	FREQUENCY MODEL					SEVERITY MODEL				
Predictor	Categories	Estimate	Std. Error	p-value	$e^{Estimate}$	Categories	Estimate	Std. Error	p-value	$e^{Estimate}$
	Intercept	–1.0413	0.0393	0.0000	0.3530	**Intercept**	7.6252	0.0571	0.0000	2 049.1903
Policyholder Age	**18–23**	0.2959	0.0880	0.0008	1.3443	**18–26**	**0.0000**	-	-	**1.0000**
	24–27	0.1791	0.0436	0.0000	1.1961	**27–32**	–0.1674	0.0539	0.0019	0.8459
	28–31	0.0609	0.0326	0.0413	1.0628	**33–37**	–0.2645	0.0539	0.0000	0.7676
	32–37	**0.0000**	-	-	**1.0000**	**38–45**	–0.3526	0.0553	0.0000	0.7029
	38–44	–0.1745	0.0305	0.0000	0.8399	**46–55**	–0.2965	0.0544	0.0000	0.7434
	45–53	–0.1881	0.0312	0.0000	0.8285	**56–61**	–0.2761	0.0594	0.0000	0.7587
	54–61	–0.1948	0.0316	0.0000	0.8230	**62+**	–0.3645	0.0631	0.0000	0.6945
	62+	–0.2239	0.0405	0.0000	0.7994	**LE**	–0.3072	0.0542	0.0000	0.7355
	LE	0.0788	0.0289	0.0064	1.0820					

Table 4 (continuation)

	FREQUENCY MODEL					SEVERITY MODEL				
Predictor	**Categories**	**Estimate**	**Std. Error**	**p-value**	$e^{Estimate}$	**Categories**	**Estimate**	**Std. Error**	**p-value**	$e^{Estimate}$
Vehicle Age	**0**	–0.5923	0.0674	0.0000	0.5531	**0**	–0.5263	0.0764	0.0000	0.5908
	1	–0.3148	0.0380	0.0000	0.7299	**1**	–0.4359	0.0445	0.0000	0.6467
	2	–0.1184	0.0339	0.0005	0.8883	**2**	–0.2655	0.0398	0.0000	0.7668
	3	–0.0890	0.0285	0.0018	0.9148	**3**	–0.0955	0.0336	0.0045	0.9089
	4	**0.0000**	-	-	**1.0000**	**4**	–0.0721	0.0293	0.0137	0.9304
	5	0.0692	0.0276	0.0122	1.0717	**5+**	**0.0000**	-	-	**1.0000**
	6	0.1832	0.0332	0.0000	1.2011					
	7	0.2757	0.0415	0.0000	1.3175					
	8+	0.1879	0.0492	0.0001	1.2067					
Payment Frequency	**1**	–0.1271	0.0216	0.0000	0.8806	n. s.				
	2	–0.0743	0.0277	0.0073	0.9284					
	4	**0.0000**	-	-	**1.0000**					
	12	0.1429	0.0406	0.0004	1.1536					
Vehicle Power	**0–76**	**0.0000**	-	-	**1.0000**	**0–80**	**0.0000**	-	-	**1.0000**
	77–112	0.1065	0.0245	0.0000	1.1124	**81–95**	0.0862	0.0303	0.0045	1.0900
	113+	0.1974	0.0484	0.0000	1.2182	**96–124**	0.1581	0.0407	0.0001	1.1713
						125+	0.3687	0.0647	0.0000	1,4459
TSI	**0–5000**	–0.2373	0.0314	0.0000	0.7888	**0–5 000**	–0.3070	0.0322	0.0000	0.7357
	5001–10 000	**0.0000**	-	-	**1.0000**	5001–10 000	**0.0000**	-	-	**1.0000**
	10 001–15 000	0.2017	0.0266	0.0000	1.2235	**10 001–15 000**	0.1964	0.0300	0.0000	1.2170
	15 001–25 000	0.2083	0.0407	0.0000	1.2316	**15 001–25 000**	0.3630	0.0484	0.0000	1.4376
	25 001–35 000	0.3328	0.0756	0.0000	1.3949	**25 001+**	0.8070	0.0881	0.0000	2.2412
	35 001+	0.3576	0.1014	0.0004	1.4299					
Engine Volume	**0–1354**	0.0835	0,0292	0.0043	1.0871	n.s.				
	1 355–1 397	**0.0000**	-	-	**1.0000**					

Table 4 (continuation)

	FREQUENCY MODEL					SEVERITY MODEL				
Predictor	Categories	Estimate	Std. Error	p-value	$e^{Estimate}$	Categories	Estimate	Std. Error	p-value	$e^{Estimate}$
Engine Volume	**1 398–1 480**	0.0744	0.0328	0.0234	1.0772					
	1 481–1 750	0.0476	0.0287	0.0467	1.0488			n.s.		
	1 751+	0.1437	0.0337	0.0000	1.1545					
Mileage per Year	**0–15 000**	**0.0000**	-	-	**1.0000**	**0–5 000**	–0.1688	0.0533	0,0015	0.8447
	15 000–30 000	0.1141	0.0198	0.0000	1.1209	**5 001–10 000**	–0.1523	0.0332	0,0000	0.8587
	30 000–inf	0.5379	0.0898	0.0000	1.7124	**10 001–13 000**	**0.0000**	-	-	**1.0000**
						13 001+	–0.1276	0.0244	0.0000	0.8802
B-M Class	**B0**	**0.0000**	-	-	**1.0000**	**B0**	**0.0000**	-	-	**1.0000**
	B1-B3	–0.2408	0.0216	0.0000	0.7860	**B1-B2-B3**	–0.1618	0.0265	0.0000	0.8506
	B4	–0.3893	0.0283	0.0000	0.6775	**B4**	–0.1988	0.0346	0.0000	0.8197
	B5	–0.5957	0.0345	0.0000	0.5512	**B5**	–0.1848	0.0414	0.0000	0.8313
	B6	–0.6677	0.0485	0.0000	0.5129	**B6-B7**	–0.2349	0.0556	0.0000	0.7906
	B7	–0.9646	0.1316	0.0000	0.3811	**M1-M2**	–0.1374	0.0531	0.0096	0.8716
	M1-M2	0.1419	0.0439	0.0012	1.1525					
Deductible	**No Deductible**	0.7058	0.0209	0.0000	2.0255					
	<=1%	**0.0000**	-	-	**1.0000**			n.s.		
	<=2%	–0.2518	0.0291	0.0000	0.7774					
	>2%	–0.8400	0.1099	0.0000	0.4317					
Region	**R_A**	–0.4376	0.0247	0.0000	0.6456	**R_C**	–0.0972	0.0249	0.0001	0.9074
	R_B	–0.1266	0.0260	0,0000	0.8811	**R_D**	**0.0000**	-	-	**1.0000**
	R03	**0.0000**	-	-	**1.0000**	**R_E**	–0.0710	0.0319	0.0258	0.9315
	R07	–0.0718	0.0271	0.0080	0.9307	**R_F**	–0.0623	0.0309	0.0438	0.9396
	R08	–0.1905	0.0265	0.0000	0.8265					

Legend: *R_A – R01-R04-R06-R09-R11, R_B – R02-R05-R10, R_C – R01-R07-R08, R_D – R03-R04, R_E – R02-R05-R10, R_F – R06-R09-R11, n. s. – non-significant.*

Source: Own construction, customized in R

Similarly, we can analyse and interpret the expected (average) severity in the context of individual variables. As an example, let's take a situation for the variable vehicle power (*Vehicle_Power*), which is given in the kilowatts (kW). The most risk category in terms of vehicle power consists of vehicles with an engine power of more than 125kW ($e^{Estimate} = e^{0.3687} = 1.4459$). For which the expected (average) severity per year and per policy is 44.59% greater than in the reference category with engine power up to 80kW, provided that the other factors incorporated in the severity regression model are at the same level (ceteris paribus).

The both final risk models introduced in the Table 4 represent the best choice among the other proposed ones. Determining appropriate model is crucial in the regression modelling and the emphasis is on simplicity. In this section, the models with different risk factors are compared based on the analysis of deviance and AIC and BIC, see Table 5.

The several predictive models for frequency and severity has been proposed and tested to find suitable subset of variables in the data set resulting for the best performing model. All predictors in the frequency and severity in MODEL 1 (full model) were processed as categorical variables. Using the stepwise regression with the backward selection strategy the variables *Vehicle_Weight* and *Policyholder_entity* were iteratively removed as least contributive predictors. Afterwards it was tested MODEL 2 for the frequency and severity without these two insignificant variables. In case of the severity MODEL 2 it has been excluded also the variable *Deductible*. According to the results of the analysis of deviance, AIC and BIC, the best model for the claim frequency and severity was chose as MODEL 2 in the both cases.

Table 5 The analysis of deviance, AIC and BIC

Criterion	FREQUENCY		SEVERITY	
	MODEL 1	MODEL 2	MODEL 1	MODEL 2
Deviance	53 517.93	53 518.26	11 877	11 734
AIC	71 380.00	71 375.00	199 914	199 850
BIC	71 842.14	71 808.19	200 284	199 984

Source: Own construction

Regarding to the descriptive data analysis provided in the section 2.1 the real data is not normal distributed, that is, we cannot use ordinary linear regression model. The linear regression model assumes that the outcome of response variable can be expressed by a weight sum of the selected variables with an individual error that follows a normal distribution. Simple weight sum is too restrictive for many real prediction problems. The outcome given the features might have a non-Gaussian distribution, the features might interact and the relationship between the features and the outcome might be nonlinear. This paper deals with estimation of the annual claim frequency and severity in the motor hull insurance based on generalized linear models.

We try to achieve better understanding the relation of the frequency and severity on the presented risk factors. The empirical study results are represented in the Table 4. This particular case study shows that the variables *Vehicle_Weight* and *Policyholder_entity* and *Deductible* have no statistical significance for the annual claim analysis. Based on the principle of simplicity we used the analysis of deviance to choose suitable model. In fact, this model is quite simple, what is very important and useful in the actuarial practice.

To better demonstration of achieved results from the Table 4, it is computed random policyholder profile to set the risk premium, see Table 6.

Table 6 Motor hull insurance: the model results for random selected potential customer profile

Policyholder's properties		Frequency			
	Risk profile	Reg. coeff	$e^{\hat{\beta}_{freq_j}}$	Reg. coeff	$e^{\hat{\beta}_{freq_j}}$
Intercept	1	–1.0413	0.3530	7.6252	2 049.1903
Payment Frequency	12	0.1429	1.1536	0.0000	1.0000
Policyholder Age	28	0.0609	1.0628	–0.1674	0.8459
Vehicle Age	0	–0.5923	0.5531	–0.5263	0.5908
B-M Class	B0	0.0000	1.0000	0.0000	1.0000
Region	R2	–0.1266	0.8811	–0.0710	0.9315
Vehicle Engine Volume	1 420	0.0744	1.0772	0.0000	1.0000
Vehicle Power	78.6	0.1065	1.1124	0.0000	1.0000
TSI	17 300	0.2083	1.2316	0.3630	1.4376
Deductible	<=1%	0.0000	1.0000	0.0000	1.0000
Mileage per Year	7800	0.0000	1.0000	–0.1523	0.8587
$\Pi e^{\hat{\beta}_j}$	×	×	**0.3112**	×	**1 177.6**

Source: Own construction

The frequency model predicts the number of claims for the different categories of the policyholders. General form of this model (see Table 4) is given by:

$$\hat{y}_f = e^{-1.0413} \cdot (e^{0.2959})^{ph\_age\ 18\text{–}23} \cdot (e^{0.1791})^{ph\_age\ 24\text{–}27} \cdot \ldots \cdot (e^{-0.0718})^{region\mathrm{R07}} \cdot (e^{-0.1905})^{region\mathrm{R08}}.$$

The expected claim frequency (the average number of the claims during the year) is then determined for some client with the properties listed in the Table 6 according to the formula:

$$\hat{y}_f = 0.3530 \cdot 1.1536 \cdot 1.0628 \cdot 0.5531 \cdot 1 \cdot 0.8811 \cdot 1.0772 \cdot 1.1124 \cdot 1.2316 \cdot 1 \cdot 1 = 0.3112.$$

The similar form can be expressed for the severity model which predicts the claim costs per policy where the various properties of the policyholder are taken into consideration:

$$\hat{y}_s = e^{7.6252} \cdot (e^{-0.1674})^{ph\_age\ 27\text{–}32} \cdot \ldots \cdot (e^{-0.0623})^{region R\_F}.$$

The expected severity during the year per policy, is then determined for the client with the properties listed in the Table 6 according to the formula:

$$\hat{y}_s = 2\,049.1903 \cdot 1 \cdot 0.8459 \cdot 0.5908 \cdot 1 \cdot 0.9315 \cdot 1 \cdot 1 \cdot 1.4376 \cdot 1 \cdot 0.8587 = 1\,177.6.$$

According to the Formula (1) we can calculate the risk premium for some client as:

RiskPremium = 0.3112 · 1177.6 = 366.5005.

To sum it up, it is proposed GLMs approach to investigate the risks connected with non-life policy. Based on the risk models from section 3.2, estimated premium for the specific risk profile of policyholder is EUR.

CONCLUSION

Motor hull insurance is one of the most widespread insurance in many countries and lots of data is disponible. Process of the setting the price is often difficult exercise since there are many different explanatory variables available. It is also very important that the rating system for set the risk premiums is treated carefully by company. Policyholders may leave when they are overcharged or in the contrary very low price may attract bad risks.

We have discussed in the paper the use of generalized linear models in actuarial practise which represent a suitable tool to predict key ratios, like the claim frequency, claim severity and the risk premium. GLMs are very effective because they are fairly accurate and are easy to explain to the layman in terms of the effect of each rating factor. Classification of the observed losses according to the appropriate risk factors is very important in determining how accurate the rating system is, the risk factors tells us exactly which level of which risk factor causes the biggest loss – should be charged the highest risk premium and which causes the smallest loss should be the lowest premium. The core concept of GLMs is to keep the weighted sum of features but allow non-Gaussian outcome distributions and connect the expected mean and the weighted sum through a possibly non-linear function.

At the first stage, the frequency of claims is estimated using the Poisson regression. In the next stage, the severity is determined by Gamma model where the log-link function is defined in both cases. The risk premium can be then expressed as the product of the expected claim counts and average cost per claim. Since all the weights are in the exponential function, the effect interpretation is not additive, but multiplicative. The regression coefficients as resulting from the frequency-severity model presented in the Table 4 can be also not continuous or their progress is not smooth enough which can be caused by inadequate accuracy, but also the data that does not have the behave how we would be expected. In practice this happen very often, when some factors really reflect an increasing or decreasing risk.

Apart from the general risk factors as *Policyholder age*, *Vehicle age*, *TSI*, etc…, we tend to classify the observed losses according to the Bonus-Malus system variable. This system leads to a discount – bonus in risk premium. When the claims have occurred the premium increases as the consequence of it – malus, see Table 5.

We processed a dataset with n = 91 685 policies. According to descriptive analyses provided in the initial section of the empirical study we see, that histogram of the claim frequency and claim severity is strongly right-skew, see the Figure 1 and Figure 2. It follows from this that ordinary linear regression is not fully suitable. The policyholders are divided into the groups based on the risk factors, see Table 2. According to these 10 risk factors, we get 192 000 groups. Exposure, total number of claims and total claim amount is known for each group. The variables *Vehicle_Weight* and *Policyholder_entity* are statistically insignificant and rejected at significance level of 0.05 in the risk model. The next variable Deductible is rejected just for claim severity model.

The actuaries should be aware of the so-called "one-dimensional analysis" and should not be tempted to stop the analysis in finding the averages of responses caused by each risk factor in our portfolio. The reason is very justified, these risk factors are very likely to be correlated.

We try to find the suitable GLMs for the claim frequency and claim severity in term of the risk factors. The models with different risk factors are constructed and compared each other using the analysis of deviance AIC and BIC criterion. The best risk models are those that have the lowest decision criterions compared to others that is MODEL 2 in both cases, see Table 5.

ACKNOWLEDGMENT

The paper was supported by a grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic VEGA No. 1/0120/18 – *Modern risk management tools in the internal models of insurance companies in 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. 1/0647/19 – *Modern tools for managing and modelling of risks in non-life insurance*.

The paper was supported by the internal grant project I-19-102-00 of the University of Economics in Bratislava for young pedagogical staff, scientific and PhD students entitled *Modern stochastic methods applied in tourism in Slovak Republic*.

References

AGRESTI, A. *Foundations of linear and generalized linear models.* New York: John Wiley & Sons, 2015.

ALLISON, P. D. *Logistic regression using SAS: Theory and application.* 2nd Ed. North Carolina: SAS Institute, 2012.

ANDERSON, D., FELDBLUM, S., MODLIN, C., SCHIRMACHER, D., SCHIRMACHER, E., THANDI, N. *A Practitioner's Guide to Generalized Linear Models: A foundation for theory, interpretation and application.* 3rd Ed. Towers Watson, 2007.

CHARPENTIER, A. AND DENUIT, M. *Mathématiques de l'Assurance Non-Vie, Tome II: Tarification et provisionnement.* Paris: Economica, 2005.

DAVID, M. Auto insurance premium calculation using generalized linear models. *Procedia Economics and Finance*, 2015, 20(1), pp. 147–156.

DE JONG, P. AND HELLER, G. Z. *Generalized linear models for insurance data.* Cambridge: Cambridge University Press, 2008.

DOWLE, M., SRINIVASAN, A, SHORT, T., LIANOGLOU, S, SAPORTA, R., ANTONYAN, E. *data.table: Extension of Data. frame.* R package, 2015.

DRAPER, N. AND SMITH, H. *Applied Regression Analysis.* 2nd Ed. New York: Wiley, 1981.

DUAN, Z., CHANG, Y., WANG, Q., CHEN, T., ZHAO, Q. A Logistic Regression Based Auto Insurance Rate-Making Model Designed for the Insurance Rate Reform. *International Journal of Financial Studies*, 2018, 6(1), p. 18.

EDWARD, W. F. *Regression Modeling with Actuarial and Financial Applications.* Cambridge University Press, 2010.

EWALD, M. AND WANG, Q. *Predictive Modeling: A Modeler's Introspection – A Paper Describing How to Model and How to Think Like a Modeler* [online]. Schaumburg: Society of Actuaries, 2015. [cit. 9.9.2019] <*https://www.soa.org/globalassets/assets/files/research/projects/2015-predictive-modeling.pdf*>.

FORBES, C., EVANS, M., HASTINGS, N., PEACOCK, B. *Statistical distributions.* New York: John Wiley & Sons, 2011.

FOX, J. *Applied regression analysis and generalized linear models.* New York: Sage Publications, 2015.

FREES, E., LEE, G., YANG, L. Multivariate frequency-severity regression models in insurance. *Risks*, 2016, 4(1), p. 4.

GARRIDO, J., GENEST, C., SCHULZ, J. Generalized linear models for dependent frequency and severity of insurance claims. *Insurance: Mathematics and Economics*, 2016, 70(1), pp. 205–215.

GOLDBURD, M., KHARE, A., TEVET, D. *Generalized linear models for insurance rating.* Casualty Actuarial Society, CAS Monographs Series 5, 2016.

HABERMAN, S. AND RENSHAW, A. E. Generalized linear models and actuarial science. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 1996, 45(4), pp. 407–436.

HEBÁK, P., HUSTOPECKÝ, J., MALÁ, I. *Vícerozměrné statistické metody (2).* Prague: Informatorium, 2005.

KAFKOVÁ, S. AND KŘIVÁNKOVÁ, L. Generalized Linear Models in Vehicle Insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 2014, 62(2), pp. 383–388.

KIM, K. AND TIMM, N. *Univariate and multivariate general linear models: theory and applications with SAS.* Boca Raton: Chapman and Hall/CRC, 2006.

LITTELL, C. L., STROUP, W. W., FREUND, R. J. *SAS for Linear Models.* 4th Ed. North Carolina: SAS Institute, 2010.

MCCULLAGH, P. AND NELDER, J. A. *Generalized linear models.* 2nd Ed. London: Chapman and Hall, 1989.

NELDER, J. A. AND WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 1972, 135(3), pp. 370–384.

OHLSSON, E. AND JOHANSSON, B. *Non-Life Insurance Pricing with Generalized Linear Models. EAA Lecture Notes.* Berlin: Springer, 2010.

R CORE TEAM. R: *A language and environment for statistical computing* [online]. R Foundation for Statistical Computing, Vienna, Austria, 2019. <http://www.R-project.org/>.

SHI, P., FENG, X., IVANTSOVA, A. Dependent frequency – severity modeling of insurance claims. *Insurance: Mathematics and Economics*, 2015, 64(1), pp. 417–428.

VALECKÝ, J. Modelling Claim Frequency in Vehicle Insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 2016, 64(2), pp. 683–689.

VALECKÝ, J. Calculation of solvency capital requirements for non-life underwriting risk using generalized linear models. *Prague Economic Papers*, 2017, 26(4), pp. 450–466.

VENABLES, W. N. AND RIPLEY, B. D. *Modern Applied Statistics with S.* 4th Ed. New York: Springer, 2002.

WOOLDRIDGE, J. M. *Introductory econometrics: a modern approach.* 5th Ed. Mason: South-Western, 2013.

XACUR, O. A. Q. AND GARRIDO, J. Generalised linear models for aggregate claims: to tweedie or not? *European Actuarial Journal*, 2015, 5(1), pp. 181–202.

XIE, S. AND LAWNICZAK, A. Estimating Major Risk Factor Relativities in Rate Filings Using Generalized Linear Models. *International Journal of Financial Studies*, 2018, 6(4), p. 84.

Recent Developments and Challenges in Energy Statistics in the Czech Republic

Miluše Kavěnová[1] | *Czech Statistical Office, Prague, Czech Republic*

Abstract

Energy statistics in the Czech Republic are responding to national and EU policy-makers' growing needs and requests for more detailed and more recent data, going beyond energy balances, as a way to support decision-making processes. While energy balances remain the most important and most used output of energy statistics, increasing emphasis in the field of energy statistics is now also being placed on energy efficiency indicators, physical energy flows and on the development of new common IT tools. This trend is being implemented within the existing European, international and national organizational set-up of institutions involved in energy statistics, in which various stakeholders at various levels have interconnected roles. This includes several institutions at the national level, and this, in turn, has necessitated increased communication and coordination between stakeholders. After presenting a synopsis of the current functioning of energy statistics, this article aims to provide an overview of the main recent developments and challenges in this field, including information about ongoing discussions regarding further developments and expected challenges in the near future, from both an international as well as domestic point of view.

Keywords

Energy statistics, energy balance, energy efficiency, Sankey diagrams, energy statistical surveys

JEL code

Q40, Q49

INTRODUCTION

Energy statistics in the Czech Republic, just as in the European Union, are changing to better respond to the growing needs of domestic and international users (including the EU and international organizations) for more detailed and more recent data. This has meant that energy statistics outputs have needed to expand beyond their traditional core (which has always been, and continues to be, the energy balance) to now include, for example, data on energy efficiency. At the same time, the format for presenting these outputs has also needed to evolve to go beyond the traditional table matrix and to now also reflect more modern data visualization approaches.

Notwithstanding these developments, the importance of maintaining a high quality of the traditional core output of energy statistics, i.e. the energy balance, remains unchanged. This is because the quality of energy statistics is also reflected in foreign trade statistics and national accounts, as this data is used for the construction of supply and use tables (see Sixta, 2013).

[1] Czech Statistical Office, Director of External Trade Statistics Department, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: miluse.kavenova@czso.cz, phone: (+420)274054176.

In order to maintain this high quality, but also in order to jointly coordinate and implement new developments, maintaining and further developing cooperation between the main domestic and international institutions and organizations involved in energy statistics is of key importance.

The present article first provides a synopsis of the current situation and approaches to energy statistics at the international level, keeping in mind that the international approach is the main frame of reference for domestic approaches, including the approach taken in the Czech Republic. The information that is used, as well as relevant tools that employed in the process of collection and evaluation of energy statistics data, along with a description of the role of individual institutions is then described. Then, an overview of the most important recent challenges and developments in energy statistics is provided, as well as an overview of recent discussions regarding future developments in this field. Finally, this article concludes with a description of Czech domestic specificities, including organizational aspects, which are relevant to the field of energy statistics in the Czech Republic.

1 THE TRADITIONAL ENERGY BALANCE AND THE INSTITUTIONS INVOLVED IN ENERGY STATISTICS

Before turning to more recent developments, it is worth noting that the making of an energy balance continues to be the basis of energy statistics and energy balances continue to be the most used output in this field. An energy balance is based on information about where fuels and energies come from, on the one hand, and about how fuels and energies are being consumed, on the other hand. In this way, an energy balance gives an overall picture of the energy situation in a given country. Moreover, it allows users to understand the energy security situation and the effective functioning of energy markets and serves as a tool to ensure comparability of statistical information between different reference periods and between different countries. It also provides data for the calculation of greenhouse gas emissions from fuel combustion, provides the essential basis for calculating various indicators of each energy product's role in the country's economy (energy efficiency, share of renewable energy, energy savings, consumption of energy by sector and others), see Eurostat (2018a).

The traditional form for presenting an energy balance is the table matrix. For this purpose, the table matrix is vertically divided into 3 main parts (rows pertaining to items of supply, transformation and final consumption). The columns in the table matrix then show the commodity balances for individual products (coal, oil, natural gas, renewables, electricity and heat, nuclear energy etc.). All data in an energy balance is comparable, thanks to the use of a common energy unit (and this is why an energy balance can define "Total energy", despite this total being based upon various different products). A common energy unit can be the Terra joule (TJ), Peta joule (PJ), tonnes of oil equivalent (toe) etc. To convert physical units (i.e. how much of a certain product) to such common energy units, calorific values of the products (fuels) need to be calculated and assigned, keeping in mind that the calorific value of a certain fuel has some variability, as it depends upon the quality of the fuel. These calorific values tell us how much energy is produced when burning the given fuel (see Gigoux, 2018). Further information on the structure of the energy balance is available, for example, on the Eurostat website, where the diagram of this matrix is also displayed according to the relevant methodology (Eurostat, 2018a).

Currently, energy balances for individual EU member states are calculated and published by Eurostat based on its methodology. Additionally, energy balances for individual states, such as the Czech Republic, are also calculated and published by several other international organizations (such as the International Energy Agency, as part of the OECD, and also by the United Nations) based upon the same data, but each with its own specific methodology. In addition, some states, such as Germany, also compile and publish their national version of the energy balance, which is based on their domestic methodology.

In order to build these energy balances at the EU and international level, structured data (information) is needed. The necessary data is gathered from national (country) responses to six annual questionnaires

jointly developed and issued by the international organizations collecting energy statistics and publishing energy balances. These six joint annual questionnaires are the Coal questionnaire, Natural gas questionnaire, Electricity and heat questionnaire, Oil questionnaire, Renewables questionnaire and Nuclear questionnaire. The goal of these international organizations is to agree on a joint version of each questionnaire and to harmonize and connect the various concepts in the questionnaires. The challenge for the international organizations is to develop joint questionnaires which are suitable for the various needs of the organizations in question while ensuring that the national statistical offices can fill in only a particular joint questionnaire on Coal, for example, instead of having to transmit various different versions of a Coal questionnaire to each international organization.

However, these six annual questionnaires do not only serve to build the energy balance, but are also used for environmental statistics. Eurostat has developed a method for converting energy statistics collected via the annual questionnaires into the Physical energy flow accounts (PEFA) framework, which records the flows of energy (in terajoules) from the environment to the economy (natural inputs), within the economy (products), and from the economy back to the environment (residuals).

However, this joint international approach requires some level of simplification and thus, a certain national level of detail can be missing. From the Czech domestic perspective, we could say that some national specificities are not captured in the international questionnaires, as we might otherwise like them to be. These common international questionnaires are continually being developed and refined and their development follows developments in the energy sector. In practice, the scope of the questionnaires is constantly increasing and the questionnaires are becoming more detailed. Thus, the continual task of ensuring a joint approach to the questionnaires means that cooperation between international organizations collecting energy statistics is also continually being strengthened.

2 RECENT DEVELOPMENTS AND CHALLENGES FOR ENERGY STATISTICS

For end-users with little experience reading energy balances, it can be a challenge to read the traditional table matrix to get relevant information pertaining to energy statistics. The goal of contemporary energy statistics experts is to make energy statistics more easily understandable to the public and to non-statistics professionals. At the international level, in addition to the traditional energy balance in the form of a table matrix, energy statistics data is now also being visualized using infographics and delivered to end-users through explanatory publications, such as articles and videos. The first digital publication on energy statistics in the EU has recently been published ("Shedding light on energy in the EU: A guided tour of energy statistics") and energy flow diagrams, called Sankey diagrams, have begun to be used. These diagrams are one of the important recent developments for visualizing energy statistics data.

Sankey diagrams visualize the flows of fuels and energies in a given country and in a given year. These flows would otherwise have been described by numbers in the energy balance table matrix, but the same information is better visualized in the diagrams representing the flows of production, imports, stocks, transformation, final consumption, etc. A significant advantage of the Sankey diagram is that the reader can quickly see the proportion of individual fuels and energies in the energy flows and their contribution to the economy.

Sankey diagrams are designed to be viewed in electronic format and thus, they are interactive. The user can click to make the transformation flows and processes more detailed and can then see what happens in such processes (in power stations or in refineries, for example) and what other energy transformation processes occur at the same time. The user can choose just one family of fuels, for example solid fuels and can click deeper to see which solid fuels are included in the flow of solid fuels (brown coal, bituminous coal, coking coal etc.). If a user wants to see the ratios of the flows, they can access the graphs for different parts of the flow. In these Sankey diagrams, data can also be compared between different EU countries and also over time (by selecting the appropriate years). Sankey diagrams have been recently complemented

with another interactive tool for energy prices. Sankey diagrams in energy statistics are being further developed and more information on this topic is available on the Eurostat and IEA webpages.[2]

Another challenge for energy statistics is that the traditional energy balance itself is now seen as providing insufficient data for policy-makers, analysts and other expert users. Thus, policy-makers in the field of energy statistics have begun to focus on providing an even more detailed breakdown of final energy consumption. For energy consumption in households, more detailed data is already available and the main interest of expert users of such data has been to see how much energy and what kind of energy households consume for space heating, space cooling, water heating, lighting, cooking and for appliances. For the industrial sector, there will be further legislative changes in 2019 to the *Regulation (EC) No. 1099/2008 of the European Parliament and of the Council of 22 October 2008 on energy statistics* (hereinafter "Regulation on energy statistics") and EU member states will begin reporting more detailed data about the final consumption of individual types of fuels and energies in the industrial sector, based on the – more or less detailed – codes of the *Statistical classification of economic activities* (NACE classification). A similar approach is being prepared for the transport and services sectors and the EU is currently evaluating what kind of additional data could be collected and provided in these sectors. The goals of EU policy makers appear to be rather ambitious in this area. Initial proposals include reporting data for the final consumption of fuels and energies in the services sector by type of building and by the NACE classification, classified according to whether the consumption occurs inside or outside of the building, whether the building is public or private, etc. The final consumption reporting according to the purpose of use (heating, cooling, lighting, etc.) is also being taken into account. In the field of transport, it is currently being proposed to divide the final consumption of fuels and energies by freight and passenger transport, by fuel type, by type of transport (road, rail, air, etc.) and to break down the data for urban and extra-urban transport.

An important focus of policy makers is now also being placed on energy efficiency indicators. Energy efficiency indicators tell us how much energy is needed to provide a certain service. The principle of energy efficiency is generally the energy consumption related to an activity. The most typical energy efficiency indicator is the Final consumption related to GDP. However, much more detailed information can be generated from energy efficiency data. For example, information on the energy costs of producing bricks, cellulose, cement etc. Energy efficiency indicators can also provide important information on reductions in emissions, which can then be used to better set targets and to better monitor the impacts of changes in energy policies. Moreover, energy efficiency indicators can help to identify how much of existing energy consumption is covered by existing energy efficiency regulations (only 30% of global energy consumption is a subject to mandatory efficiency targets, for example), see Silva (2018).

Fundamental energy efficiency indicators can be calculated from the energy balance. These indicators, among others, are collected and monitored by the European Commission, which sets and pursues energy efficiency targets. However, even more detailed data for energy efficiency indicators is collected by the International Energy Agency. In order to calculate these more detailed indicators, member states send data to the IEA for four main sectors (residential, services, industry and transport), including data about end uses and about the consumption of individual fuels and energies in these end uses. Examples of end uses by sector include: space heating, space cooling and lighting (residential sector, services sector); production of textiles, chemicals, paper, basic metals (industry sector); operation of passenger cars, buses, trucks (transport sector). Examples of energy efficiency indicators which are then calculated are: per capita energy intensity, per floor area energy intensity, fuel intensity and vehicle-kilometer energy intensity. In calculating these indicators, the IEA also uses macroeconomic data and data from social statistics.

[2] *<https://ec.europa.eu/eurostat/cache/sankey/sankey.html?geos=EU28&year=2017&unit=KTOE&fuels=TOTAL&highlight=_&nodeDisagg=0101000000000&flowDisagg=false&translateX=0&translateY=0&scale=1&language=EN>. <https://www.iea.org/Sankey>.*

The industry sector can be used to illustrate just how much data requests have increased in scope: the original data requirements pertaining to this sector which were then used to calculate the energy balance were for 5 items, but the data requirements pertaining to this sector which are used as a basis for calculating the energy efficiency indicators are for 23 more detailed items.

Just as with energy balances, data visualization is also used for energy efficiency indicators as a tool to make the numerical data better accessible to non-expert users. The IEA has established a new website[3] devoted to energy efficiency indicators this year which includes Sankey diagrams for energy efficiency indicators and for the various datasets for these indicators in the member countries.

The latest challenge for energy statistics that is currently being discussed is the fundamental transformation of European energy system, in the context of the establishment of the Energy Union and ambitious plans in climate policy. The European Union and its member states are currently discussing options for better aligning data collections with such EU policy developments. In particular, the EU's comprehensive update of its energy policy framework (which aims to facilitate the transition away from fossil fuels towards cleaner energy and which was introduced through new legal texts published as part of the "Clean Energy for All Europeans") brings regulatory certainty, in particular through the introduction of the first integrated national energy and climate plans. Ambitious regulatory targets for renewables, energy efficiency as well as electricity interconnection, aim to stimulate Europe's industrial competitiveness, boost growth and jobs, reduce energy bills, help tackle energy poverty and improve air quality. Using reliable high quality statistical data is necessary to monitor these new energy and climate policy goals. Official energy statistics thus need to contribute to this process in order to remain in tune with the needs of EU policy-makers. Recent energy statistics improvements agreed at the European level include the Crude oil import register, improved timeliness of monthly coal and electricity data, early estimates of energy balances and early estimates of indicators for Europe 2020/2030 targets. Moreover, to capture the development of new phenomena in energy statistics, such as the emergence of electric mobility, the expansion of the use of space cooling in southern Europe, the use of hydrogen as fuel, and so on.

3 SPECIFIC CHALLENGES FOR ENERGY STATISTICS IN THE CZECH REPUBLIC

In addition to the developments and challenges described above, which the Czech Republic faces as well, there are also some other challenges for the energy statistics in the Czech Republic.

Several domestic authorities are involved in the production of energy statistics data at the national level in the Czech Republic, with various roles assigned and with a defined scope for their inter-institutional relations. The primary responsibility for energy statistics is shared between the Czech Statistical Office (CZSO) and the Ministry of Industry and Trade (MIT). In addition to the CZSO and MIT, the other authorities involved are the Energy Regulatory Office (ERO), The Czech electricity and gas market operator (OTE), Czech Hydro-meteorological Institute (CHMI), the Administration of State Material Reserves (ASMR) and the General Directorate of Customs Administration of the Czech Republic. These authorities then also cooperate with private-sector associations, such as the Czech Association of the Petroleum Industry and Trade.

The CZSO acts as the main coordinator in energy statistics, which includes a primary methodological role. It is also the primary national contact for international organizations in this field. The CZSO is the main authority responsible for data reporting further to the Regulation on energy statistics and it is co-responsible for data reporting further to *the Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency* (hereinafter EED directive). Conversely, the MIT is the main authority responsible for data reporting further to the EED directive and is co-responsible for data reporting further to the Regulation on energy statistics. Furthermore, the CZSO is responsible

[3] *<https://www.iea.org/statistics/efficiency>.*

for monthly and annual oil statistics, monthly and annual natural gas statistics, annual solid fuels statistics, annual electricity and heat statistics, energy prices statistics and energy efficiency indicators. The MIT is responsible for monthly solid fuels statistics, monthly and annual electricity and heat statistics (for licensed entities), energy efficiency indicators, renewable energy sources, liquid biofuels, filling/gas stations statistics and nuclear energy statistics. Inter-institutional cooperation and coordination among the domestic authorities involved in the production of energy statistics data in the Czech Republic is thus both important and necessary and is becoming increasingly challenging.

The other main challenge for the Czech Republic in producing energy statistics is the same as in other EU member states. On the one hand, there is a growing demand for more energy data and also for more flexible, more detailed and earlier outputs. This means increased workload and necessitates seeking out new administrative sources of data and data from social statistics, macroeconomic statistics or statistics in other sectors (these sources are often new for energy experts). On the other hand, there is also pressure to reduce the administrative burden for respondents and the material resources available in the public sector to perform these tasks are not increasing (in terms of number of employees available to perform such tasks and in terms of the allocated budget).

Regarding the data sources that are currently used to produce energy statistics data in the Czech Republic, the most important source are statistical surveys. Statistical surveys for energy statistics are conducted by the CZSO, MIT and ERO and they are collected in the framework of the *Act No. 89/1995 Coll., on the state statistical service, as subsequently amended* and the *Act No. 458/2000 Coll., Energy Act, as subsequently amended*. The respondents for whom statistical surveys in the area of energy statistics are intended are then defined by the Decree on the Statistical Survey Program, which is issued for each given year and also by Decree No. 404/2016 Coll., on statistics.

The CZSO has two monthly statistical surveys on crude oil processing and petroleum products and five annual statistical surveys (one on fuels and energy sources, two on fuels transformation, two on fuels and energy consumption). Moreover, every five years, the ENERGO survey on energy consumption in households is held. The ENERGO survey is a unique type of survey in Europe and it has become an important source of data for all Czech institutions. For the future, while it is hoped that this survey should continue be held every five years, due the fact that it is rather demanding in terms of the capacity that is required for its development and processing, as well as due to the financial costs incurred, such a frequency is not yet guaranteed. The MIT has two monthly statistical surveys (one on solid fuels and one on biofuels) and three annual surveys (two on energy production, one on operation of service/filling stations, and one biannual survey on network of service/filling stations). The ERO has three monthly and one annual survey on electricity production, distribution, transmission and related licenses and also one quarterly survey for heat production and a further assortment of nine additional surveys for natural gas production, distribution transmission and related licenses.

In addition to statistical surveys, a variety of other sources are used in the production of energy statistics. These are, for example: the Natural Gas Balance and the Electricity Balance (produced by the ERO), Intrastat and Extrastat; business licenses in the various energy sectors; information pertaining to emissions (from the Register of Emissions and Air Pollution Sources administered by the CHMI), renewable energy sources (provided by the ERO and the State Environmental Fund), liquid biofuels (from State Agricultural Intervention Fund) and Oil and Petroleum Products (provided by the Czech Association of Petroleum Industry and Trade).

As regards any new requirements, in practice, it can be difficult to quickly include them in the scope of data collected in the surveys and the participation of all stakeholders, including respondents, is needed. Thus, one new trend in the field of energy statistics is a move away from expanding traditional statistical surveys, and instead using statistical modelling and seeking more existing administrative sources of data.

Lastly, one further challenge that needs to be addressed in the Czech Republic is the need to concentrate energy statistics outputs in one place, or at least to provide information in one centralized location, about where the various outputs can be obtained from the institutions involved, in order to make research easier for non-expert users (essentially, the creation of an energy statistics portal to seamlessly bridge the divide between the various institutions involved).

CONCLUSION

Demand for new energy statistics data is closely linked to developments in the energy sector, as new fuels and energy technologies are now ever more quickly being developed. The demand for new data is linked not only to developments within the energy sector itself (the ongoing fundamental transformation of the European energy system), but in the context of the Energy Union, it is linked to an ever-increasing demand for factually related statistics in related policy areas such as environmental statistics and climate policy (for example, the need to calculate CO_2 emissions). As new fuels, such as hydrogen, begin to be used and as the accumulation of electricity and electric mobility become the new reality, energy statistics will continue to face new challenges and a growing and changing demand. It will be the task of the domestic and international institutions involved and a task for energy statistics experts to address these new challenges and to find solutions which will allow more data to be produced and presented in a user-friendly manner, while at the same time reducing the administrative burden on respondents and making do with limited public resources. It is important to keep in mind, however, that the production of data is not the end-game – once data is produced, it falls especially upon law-makers and policy-makers in the public sector, as well as on private sector businesses, to make effective use of this data in their decision-making processes and in their policy and business planning.

References

Directive 2012/27/EU of the European Parliament and of the Council of 25 October 2012 on energy efficiency, amending Directives 2009/125/EC and 2010/30/EU and repealing Directives 2004/8/EC and 2006/32/EC.

EUROSTAT. *Energy balance – Statistics explained* [online]. Louxembourg: Eurostat, 2018a. [cit. 3.4.2019]. <https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Energy_balance#What_is_an_energy_balance.3F>.

EUROSTAT. *Shedding light on energy in the EU: a guided tour of energy statistics* [online]. Luxembourg: Eurostat, 2018b. [cit. 4.4.2019]. <https://ec.europa.eu/eurostat/cache/infographs/energy>.

GIGOUX, R. *Energy Balances and RD&D statistics.* Paris: International Energy Agency, 2018 (unpublished presentation).

Regulation (EC) No. 1099/2008 of the European Parliament and of the Council of 22 October 2008 on energy statistics.

SILVA, M. *Energy Efficiency Indicators.* Paris: International Energy Agency, 2018 (unpublished presentation).

SIXTA, J. Development of Input-Output Tables in the Czech Republic [online]. *Statistika: Statistics and Economy Journal.* Prague: Czech Statistical Office, 2013, 93(2), pp. 4–14. ISSN 1804-8765.

Official Statistics between Past and Future

Marek Rojíček[1] | *President, Czech Statistical Office, Prague, Czech Republic*

Abstract

In January 2019, the Czech Statistical Office representing its predecessors celebrated 100 years of existence. In the middle of 1990's, the modern legal framework anchored independence of the State Statistical Service as a basic condition for its professional and impartial work. Although statistics is in principle a very conservative discipline, it needs to reflect changes in the economy and the society. We can now observe a change in the basic paradigm of the official statistics consisting in movement from a stovepipe model of statistical surveys to a modern data hub linking all kinds of data sources and using sophisticated statistical models. Targeted and understandable communication is an integral part of the statistical production process and a necessary condition for a statistical office to compete on the information market. The Czech Statistical Office started 15 years ago to redesign its statistical information system and the basic principles are still valid. In the future, we are going to further reduce statistical surveys, intensify usage of administrative and private held databases, and modernise dissemination tools.

Keywords

Official statistics, consistency, statistical information system, communication

JEL code

C18

INTRODUCTION

The Czech Statistical Office (CZSO) is an institution with a long tradition. In January 2019, we celebrated the centenary (representing its predecessor State Statistical Office) and this is a good occasion to reflect on the general development of the official statistics. The CZSO is nowadays widely respected for its professionalism, which was always in the history at a relatively high level. The modern legal framework for the State Statistical Service was anchored in 1995 with a significant contribution of the President of the CZSO Edvard Outrata who transferred his professional experience from Canada into the Czech environment. The statistical law strengthened especially the independence of the State Statistical Service and this aspect turned out to be timeless. All successive Presidents of the CZSO could further build on these grounds and had a very ambitious aim to maintain high standards. The CZSO as well as all other statistical institutions are also facing many challenges resulting from the rapidly changing society. In this article, I would like to enumerate some of the challenges and outline the future of the official statistics.

1 THE MAIN CHALLENGES FOR THE OFFICIAL STATISTICS

The society as a whole is developing and official statistics is, too. National statistical institutions (NSIs) are facing increased number of users both on national and international level. Furthermore, a great

[1] President, Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic. E-mail: predseda@czso.cz.

challenge for NSIs are changing users' needs and calls for better quality of statistical information. This phenomenon is caused by progressing economic globalisation and rapid growth of information and communication technologies (ICT), namely spreading use of the Internet. At the same time, however, NSIs are requested to increase efficiency of statistical production and to reduce burden on statistical respondents. They use new data sources such as administrative or big data instead of traditional statistical surveys. The challenges for the statistical system lie also in changes in user needs influenced by the changes in the society. A hundred years ago, the main statistical domains were agriculture or industrial statistics and the population census, whereas today's statistics should reflect e.g. services, information society, sustainable growth, and many other domains. There are also changes in dissemination tools, statistical institutions addresses the users and media more directly using social networks, multimedia, etc. On the other hand, official statistics is one of the stabilizing elements in the rapidly changing society – the users can rely on its independence and objectivity.

If we summarize these challenges, they are connected with some basic questions, which are related to statistics. The first question is "What to measure?" There is a demand to statistically capture new phenomena such as social welfare, financial transactions, global value chains, etc. The second question for statisticians is "What data to use?" Except the traditional statistical surveys, more and more data are stored in various government registers and there is also a huge number of digital data held by internet platforms like Google, Airbnb, etc. Huge amounts of data are also processed by retailers or telecommunication operators. We call them generally "Big data". The third basic question is "How to communicate" data and related stories and how to find the most efficient way of transferring the information to individual users.

2 CHANGE OF PARADIGM IN OFFICIAL STATISTICS

Development of the official statistics could be basically described as a transformation from a traditional to a new generation model (or paradigm). These models could be in a simplified way described as follows:

Traditional official statistics is based on a "stovepipe" model of parallel statistical surveys, fit for purpose, output data are result of summarization of input data, and data are presented in the form of isolated tables and graphs. Low consistency of statistical indicators is determined by low level of consistency in methodology across individual statistical domains. It requires a relatively high number of routine job positions checking the quality, communicating with respondents and transferring data from questionnaires to statistical databases.

The stovepipe model is an outcome of a long historic process in which statistics in individual domains have developed independently from each other. It has a number of advantages: the production processes are best adapted to the corresponding products; it is flexible in that it can adapt quickly to relatively minor changes in the underlying phenomena that the data describe; it is under the control of the domain manager and it results in a low-risk business architecture, as a problem in one of the production processes should normally not affect the rest of the production.

The traditional model also has a number of disadvantages. Firstly, it imposes an unnecessarily heavy burden on respondents. Given that the collection of data in different domains is done in an independent and uncoordinated manner, respondents are regularly asked for the same information more than once. Secondly, the traditional model is not well adapted to collect data on phenomena that cover multiple dimensions, such as globalisation. Last but not least, this way of production is highly inefficient and costly, as it does not make use of standardisation across statistical domains.

The statistical office of the new generation can be on the input side described as a data hub – i.e. it is linked to various data sources available in public registers and corporate information systems (ERP,[2] cash

[2] ERP = Enterprise Resource Planning is a centralized system that provides tight integration with all major enterprise functions be it HR, planning, procurement, sales, customer relations, finance or analytics, as well as to other connected application functions.

registers). Statistical surveys are rather a supplement of administrative data then the basic source. Input data are usually not originally tailored for statistics, statistical indicators are compiled by a combination of various data sources, and more than a simple summarization the work of statisticians consists of data mining, linking and modelling. It also has an impact on the qualification of the staff who needs to be familiar with the sophisticated methods and tools combining statistical and IT skills. On the other hand, this model opens more space to a better consistency of statistical indicators, which can be ensured at the level of the data model construction. It also enables to present statistical data in consistent "stories" about the economy and the society supported by relevant figures.

Development of experimental calculations is also an important part of the new generation model. Traditional pace of the progress in official statistics is nowadays too slow compared to the changes in the society and the economy. It is necessary to maintain certain conservatism of the official statistics (to keep consistency in time and prevent "dead ends"), but higher courage to experiment seems to be inevitable. Using experimental procedures or data sources is a good way to protect the robustness of the official statistics as well as to reflect the new user needs. It is logical that the experimental procedures could eventually become a part of the official statistics and, on the other hand, some of the official methods become obsolete.

The two institutional models described above are, in a way, extreme cases and in reality most statistical offices represent a mix of them. The long-term ambition of the Czech Statistical Office is to move from the traditional institutional model to the new generation one.

3 CONSISTENCY AND COHERENCE OF STATISTICAL DOMAINS

One of the benefits of the new generation models is to ensure higher consistency and coherence across statistical domains. In addition to the combination of data sources and software tools, the change from the traditional model to a new one needs especially the change of thinking of the statisticians. This is not an issue at the national level only; it begins at the level of international organisations responsible for coordination of methodologies and standards. The experts in these institutions very often live in the "stovepipe" model. There are many examples of these inconsistencies across statistical domains, e.g. micro and macro indicators about income, consumption, and wealth or a consistency between structural business statistics and national accounts.

It is clear that it is unreal to ensure totally consistent data across domains. Very often, the differences in the methodology of similar indicators are justified. It is also necessary to distinguish between very experienced users and the lay ones. The solution consists in strict delineation between primary (input) and output indicators, using of transparent bridge tables, or an introduction of the system of satellite accounts.

A good example of the (in)consistency of statistical data is external trade statistics (in goods). Data about external trade are published as a part of national accounts, balance of payments, and international trade statistics. Originally, this statistics served for microeconomic as well as macroeconomic analysis and the definitions of exports and imports were clear: it is the change of ownership of goods between two countries. In practice, the trade was measured by custom statistics measuring crossing of the state borders and, in the past, it was a good approximation. In the globalized world, the national borders became less important for trading companies, especially in the custom unions and free trade areas. This problem is the most evident within the European Union, which is from the legal point of view a free trade zone, but from the point of view of statistics, the trade between the states is declared in the same way as if there were custom borders. The companies can trade in all EU countries in the same way as in the domestic country. It means that the movement of goods across the borders does not correspond to the change of ownership and international trade statistics (represented by systems of Extrastat and Intrastat) does not provide an objective picture of the external trade (especially in small open economies). The Czech Republic (where this problem became significant compared

to its GDP) is one of the few countries, which were very proactive in this field and introduced a sophisticated and consistent solution of this problem.

4 HOW TO COMMUNICATE STATISTICAL DATA?

The ways in which the data are collected and compiled are very important for statisticians. Nonetheless, for the users it is only something that is in the "black box" – they perceive only the outcome at the end of the "production line", i.e. the figures or even better the narrative behind them. In today's open market economies and modern society, official statistics have to compete with many private and public data sources freely available and of diverse quality. This challenges traditional thinking that users and consumers of available data first provide an assessment of the respective sources and second can differentiate between good quality "official statistics" and less good quality "statistics". To be able to compete on this market, it is absolutely necessary for the NSIs to be proactive and have the communication function as an integral part of the statistical production process. Communicating understandable and easy-to-use statistics not only supports (statistical and other) literacy, but it also contributes to enhancing the trust in official statistics and in the institutions responsible for producing statistics and, furthermore, it contributes to a knowledge-based society critically verifying and maintaining the accountability of policy decisions.

The new communication strategy means movement from a traditional "pull" concept, where statistics is released in databases for public use to a new "push" concept. This new push concept relates to a new function, whereby the statisticians segregate and provide tailored statistics to different users' groups by facilitating the understanding and simplifying the integration of statistics into the "non-statistical world". The professional expert users will continue to know and use the variety and granularity of public released statistics in databases and they have often a special need for statistics (for instance, being granted access to confidential data for research purposes). On the other hand, the lay users are rather confused, when they have to make many complicated choices. The important issue is to acknowledge that the "statisticians are best placed" as producers of statistics and with their statistics knowledge of the business and applied methodology to guide a layman to the most relevant set of statistics.

Statistics has to be understood before it can be used. Statisticians have a competitive advantage based on their long-term reputation of providing independent, factual, and credible statistics and they have the knowledge to understand the methodology, reporting guidelines, economic concepts, and estimation methods. This knowledge is a prerequisite for communicating statistics. The way in which statistics are presented is vital in facilitating the users' understanding of the statistics and in enhancing their usability: they must be presented according to the needs of the various user segments. There are many tools available on the market to assist statisticians in this regard, such as web-based movies, interactive tables, info-graphics, mobile platforms, etc.

The fundamental issue in communication is the ability to interpret narrative and statistics using common language that is tailored to the target audience. Each statistical domain includes methodology concepts that need to be converted into text, thereby building bridges between the language of statistics and the common language. It is important to realise that presenting statistics in a common language and using references to statistical definitions does not compromise the accuracy of statistics. Statisticians are not able to force professional users and policy-makers to adapt and use statistical terminology and statistical classifications. Statisticians need to engage externally and contribute with their wealth of knowledge to ensure that the statistics are used in the right context and are understood, as part of reflecting the structures and changes in our economy and the society. The use of the language of statistical classifications is a barrier to facilitate users' understanding and thereby frequent use.

5 MODERNISATION OF THE CZECH STATISTICAL OFFICE

During the accession process to the EU at the beginning of the 2000's, the CZSO was mainly driven by the needs to fulfil the EU requirements of the European statistical and other related legislations.

The statistical activities have been extensively developed as regards both statistical data collection and the amount of data available to the users. The result of the accession process was, in principle, an extended national framework of statistical surveys. At the same time, the need for a modern statistical information system (SIS) consistent with the GSBPM[3] model emerged.

Satisfaction of an increasing user's demand for statistical information as well as decrease of the administrative burden were the main driving forces for a new architecture of the SIS. The first important step in this endeavour was to design a new global architecture of the SIS. The main goal of the architecture was to strengthen organization and management of statistical work. The whole global architecture had several aspects (parts) – the content (what data are collected and from whom), processes (how the data are collected, processed, and disseminated), and modernization of IT infrastructure, which have been implemented in different time periods since 2005 to 2014. The project of the SIS redesign was an important step on the way from the traditional to the new generation statistical model.

The content part of the redesign consisted of maximum use of modelling, administrative data and use of data from one statistical task in another one. The new model was also based on the coordination of survey samples, rotation of an extended sample for individual NACE activities or rotation of variables for which a detailed structure is required. The principle of statistical coherence was one of the aspects of the reform. So-called principal statistical tasks were defined with the aim to determine an absolute value of surveyed (estimated, modelled) variables (by calibration or confrontation) in all relevant tasks or to be binding for determination of more detailed structure of these variables.

Different statistical domains had to respect consistency of published data (single figure principle). The principle of completeness was also very important, which meant that published outputs of core and standard variables cover the whole population (not only a fraction, e.g. only businesses with 10+ employees). If a statistical survey covered only a fraction of the population it was supposed that the below-threshold part estimate would be determined by modelling (e.g. based on administrative data or other surveys).

The Czech Statistical Office is nowadays at the beginning of planning the upgrade of the SIS (we can call it "Redesign 2.0"). It is necessary to take into account that it is already 15 years since the current SIS was designed. Unlike the then situation, we do not expect to reconstruct the system completely, but rather to modernize and complete certain parts of it. This proves that the concept of the current SIS is timeless and still valid. Unlike the original concept based on closely interlinked subsystems, we will prefer a modular concept of independent subsystems linked via interfaces that enables higher flexibility.

We will focus mainly on two statistical processes: collection of data and dissemination. Concerning the first one, we expect a higher share of input data from administrative data, registers, and private databases (e.g. scanner data and ERP systems) in the future. It will be necessary to prepare interfaces for an automatic transmission of a data batch based on common standards for the government registers. One concrete example in the domain of demographic statistics is so-called "Census Information System" built as a part of the 2021 Population Census.

The upcoming Population Census is also an opportunity to upgrade our Public database, which was developed as a part of the Redesign of the SIS in 2014. One of the most important tools within the Public database will be construction of hypercubes that enable to experienced users to create tailored made tables, charts or cartograms. However, the modern dissemination strategy is not only a question of technical tools; it is also about how statisticians are able to communicate with users. In the last years, the CZSO made significant progress in this way and stands out with many other NSIs in the EU. We have introduced info-graphics and published them in the social media, our magazine "Statistics and Us" is widely used by the media and various stakeholders including politicians, we are increasing our presence in the media by providing citations as a part of press releases and many other activities.

[3] GSBPM = General Statistical Business Process Model.

CONCLUSION

The official statistics in the Czech Republic celebrated its centenary. The environment in which it exists is changing rapidly and it is not easy to keep up with the times. In the same way as companies and institutions in other business areas, it needs to innovate its products and production processes. In the world of official statistics, we can observe movement from the traditional "stovepipe" model to the new generation model, which brings a new paradigm. It is connected with three fundamental questions for official statistics: what to measure, what data to use, and how to communicate with users. Modernization of statistical processes is also a key target for the Czech Statistical Office. The activities are focused mainly on the way the input data are obtained, increasing share of administrative or other data alternative to statistical surveys, and modernization of dissemination and communication tools. One of the priorities is also increased consistency across statistical domains. A very good example of a successful harmonization in the last years are external trade statistics.

References

CZECH STATISTICAL OFFICE. *External trade in the national approach – methodology* [online]. Prague: Czech Statistical Office, 2011. <https://www.czso.cz/csu/czso/2-vzonu_m>.

EUROSTAT. *Communication from the Commission to the European Parliament and the Council on the production method of EU statistics: a vision for the next decade.* Luxembourg: Eurostat, 2009.

NYMAND-ANDERSEN, P. *Preparing a statistics communication strategy.* Geneva: Conference of European Statisticians, 2017.

OUTRATA, E. Some Future Challenges for Czech Official Statistics [online]. *Statistika: Statistics and Economy Journal*, 2019, Vol. 99, No. 2, pp. 218–224. <https://www.czso.cz/documents/10180/88506448/32019719q2_218_outrata_anniversary.pdf/380397a3-9d39-43d4-ae2f-8952581dc534?version=1.0>.

PIŠTORA, L. Hundred Years of the Czech Statistics [online]. *Statistika: Statistics and Economy Journal*, 2018, Vol. 98, No. 4, pp. 385–389. <https://www.czso.cz/documents/10180/61266315/32019718q4385.pdf/0b0a6858-1f14-4c67-b300-6dca49300618?version=1.2>.

ROJÍČEK, M. *Redesign of the Statistical Information System: a Czech experience.* Madrid: European Conference on Quality in Official Statistics, 2016.

ZÁVODSKÝ, P. AND ŠIMPACH, O. A Centenary of the State Statistical Office [online]. *Statistika: Statistics and Economy Journal*, 2019, Vol. 99, No. 1, pp. 77–92. <https://www.czso.cz/documents/10180/88506450/32019719q1_077.pdf/f0b02e9f-01df-46f4-b028-36afad714e17?version=1.0>.

Recent Publications

New publications of the Czech Statistical Office

Demographic Yearbook of the Czech Republic 2018. Prague: CZSO, 2019.
Indicators of Social and Economic Development of the Czech Republic from 2000 till the end of Q2 2019. Prague: CZSO, 2019.
Věda, výzkum a informační technologie v mezikrajském srovnání v období 2007 až 2017 (Science, research and information technology in interregional comparison in the period 2007–2017). Prague: CZSO, 2019.

Other selected publications

Eurostat regional yearbook. 2019 Ed. Eurostat, 2019.
Main Economic and Social Indicators of the Czech Republic 1990–2017. Prague: VUPSV, 2018.
Slovak Republic in figures 2019. Bratislava: SOSR, 2019.

Recent Events

Conferences

The ***12nd International Scientific Conference RELIK 2019 (Reproduction of Human Capital – mutual links and connections)***, organized by the Department of Demography, Faculty of Informatics and Statistics, was held **during 7–8 November 2019 at the University of Economics, Prague, Czech Republic**. More information available at: <*https://relik.vse.cz*>.

The international scientific konference ***Quantitative Methods in Economics***, organized by the Slovak Society for Operations Research and the Department of Operations Research and Econometrics, Faculty of Economic Informatics, University of Economics in Bratislava, will take place **from 27th to 29th May 2020 in Púchov, Slovakia.** More information available at: <*http://www.fhi.sk/ssov/conference*>.

The ***10th European Conference on Quality in Official Statistics (Q2020)*** will be held **during 9–12 June 2020 in Budapest, Hungary**. More information available at: <*http://www.q2020.hu*>.

Papers
We publish articles focused at theoretical and applied statistics, mathematical and statistical methods, conception of official (state) statistics, statistical education, applied economics and econometrics, economic, social and environmental analyses, economic indicators, social and environmental issues in terms of statistics or economics, and regional development issues.

The journal of *Statistika* has the following sections:

The *Analyses* section publishes high quality, complex, and advanced analyses based on the official statistics data focused on economic, environmental, and social spheres. Papers shall have up to 12 000 words or up to twenty (20) 1.5-spaced pages.

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.

The *Methodology* section gives space for the discussion on potential approaches to the statistical description of social, economic, and environmental phenomena, development of indicators, estimation issues, etc. Papers shall have up to 12 000 words or up to twenty (20) 1.5-spaced pages.

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.

The *Book Review* section brings reviews of recent books in the fieled of the official statistics. Reviews shall have up to 600 words or one (1) 1.5-spaced page.

The *Information* section includes informative (descriptive) texts, information on latest publications (issued not only by the CZSO), 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 (e.g. On Laconic and Informative Titles) — Authors and Contacts — Abstract (max. 160 words) — Keywords (max. 6 words / phrases) — JEL classification code — Introduction — (chapters: 1, 2, …) — Conclusion — Acknowledgments — References — Annex — Tables and Figures (for print, for review process in the text)

Authors and Contacts
Rudolf Novak*, Institution Name, Street, City, Country
Jonathan Davis, Institution Name, Street, City, Country
* Corresponding author: e-mail: rudolf.novak@domainname.cz, phone: (+420) 111 222 333

Main Text Format
Times 12 (main text), 1.5 spacing between lines. Page numbers in the lower right-hand corner. *Italics* can be used in the text if necessary. *Do not* use **bold** or underline in the text. Paper parts numbering: 1, 1.1, 1.2, etc.

Headings
1 FIRST-LEVEL HEADING (Times New Roman 12, bold)
1.1 Second-level heading (Times New Roman 12, bold)
1.1.1 Third-level heading (Times New Roman 12, bold italic)

Footnotes
Footnotes should be used sparingly. Do not use endnotes. Do not use footnotes for citing references.

References in the Text
Place reference in the text enclosing authors' names and the year of the reference, e.g. "White (2009) points out that…", "… recent literature (Atkinson et Black, 2010a, 2010b, 2011; Chase et al., 2011, pp. 12–14) conclude…". Note the use of alphabetical order. Include page numbers if appropriate.

List of References
Arrange list of references alphabetically. Use the following reference styles: [for a book] HICKS, J. *Value and Capital: An inquiry into some fundamental principles of economic theory.* 1st Ed. Oxford: Clarendon Press, 1939. [for chapter in an edited book] DASGUPTA, P. et al. Intergenerational Equity, Social Discount Rates and Global Warming. In: PORTNEY, P. AND WEYANT, J., eds. *Discounting and Intergenerational Equity*. Washington, D.C.: Resources for the Future, 1999. [for a journal] HRONOVÁ, S., HINDLS, R., ČABLA, A. Conjunctural Evolution of the Czech Economy. *Statistika: Statistics and Economy Journal,* 2011, 3 (September), pp. 4–17. [for an online source] CZECH COAL. *Annual Report and Financial Statement 2007* [online]. Prague: Czech Coal, 2008. [cit. 20.9.2008]. <http://www.czechcoal.cz/cs/ur/zprava/ur2007cz.pdf>.

Tables
Provide each table on a separate page. Indicate position of the table by placing in the text "insert Table 1 about here". Number tables in the order of appearance Table 1, Table 2, etc. Each table should be titled (e.g. Table 1 Self-explanatory title). Refer to tables using their numbers (e.g. see Table 1, Table A1 in the Annex). Try to break one large table into several smaller tables, whenever possible. Separate thousands with a space (e.g. 1 528 000) and decimal points with a dot (e.g. 1.0). Specify the data source below the tables.

Figures
Figure is any graphical object other than table. Attach each figure as a separate file. Indicate position of the figure by placing in the text "insert Figure 1 about here". Number figures in the order of appearance Figure 1, Figure 2, etc. Each figure should be titled (e.g. Figure 1 Self-explanatory title). Refer to figures using their numbers (e.g. see Figure 1, Figure A1 in the Annex).

Figures should be accompanied by the *.xls, *.xlsx table with the source data. Please provide cartograms in the vector format. Other graphic objects should be provided in *.tif, *.jpg, *.eps formats. Do not supply low-resolution files optimized for the screen use. Specify the source below the figures.

Formulas
Formulas should be prepared in formula editor in the same text format (Times 12) as the main text 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 372 (incl. postage) for the Czech Republic,
EUR 117 or USD 174 (incl. postage) for other countries.
Printed copies can be bought at the Publications Shop of the Czech Statistical Office (CZK 66 per copy).
address: Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscriptions and orders
MYRIS TRADE, s. r. o.
P. O. BOX 2 | 142 01 Prague 4 | Czech Republic
phone: (+420) 234 035 200,
fax: (+420) 234 035 207
e-mail: myris@myris.cz

Design: Toman Design
Layout: Ondřej Pazdera
Typesetting: Družstvo TISKOGRAF, David Hošek
Print: Czech Statistical Office

99th 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