STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **99** (2) 2019

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

EXECUTIVE BOARD

MANAGING EDITOR

CONTENTS

ANALYSES

DISCUSSION

METHODOLOGY

100th ANNIVERSARY

INFORMATION

About Statistika

The journal of Statistika has been published by the Czech Statistical Office since 1964. Its aim is to create a platform enabling national statistical and research institutions to present the progress and results of complex analyses in the economic, environmental, and social spheres. Its mission is to promote the official statistics as a tool supporting the decision making at the level of international organizations, central and local authorities, as well as businesses. We contribute to the world debate and efforts in strengthening the bridge between theory and practice of the official statistics. Statistika is professional double-blind peer reviewed open access journal included in the citation database of peer-reviewed literature **Scopus** (since 2015), in the **Web of Science** *Emerging Sources Citation Index* (since 2016), and also in other international databases of scientific journals. Since 2011, Statistika has been published quarterly in English only.

Publisher

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

Contact us

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

Modelling Deprivation of the 50-plus Population of the Czech Republic Based on the Share Survey

Ivana Malá[1] | *University of Economics, Prague, Czech Republic*

Abstract

In the paper, two composite indicators – the indices of material and social deprivation – are analyzed based on the Survey of Health, Ageing and Retirement in Europe exploring the population aged 50 and over. A finite normal mixture model for the social deprivation index and a finite mixture of Bernoulli and normal distributions are used to model the distribution of the indices of deprivation for the Czech Republic in 2013. Applying a logistic regression model, the parameter of Bernoulli distribution is supposed to be dependent on explanatory variables. In terms of material deprivation, the situation in the Czech Republic is comparable to other European countries, the social deprivation index, showing, however, higher values.

Keywords	***JEL code***
Material deprivation, social deprivation, composite indicators, normal mixed model, logistic regression	*C21, I31*

INTRODUCTION

Improving the quality of life of the population and reducing deprivation and social exclusion are the ambitious goals of the European Union (EU) and other developed countries. Unfortunately, we are not able to accurately measure these serious issues so that they can be subjected to quantitative analysis. In order to address this complex problem, it is important to quantify subjectively and/or emotionally perceived experience, such as life satisfaction and quality, or material and social deprivation (Bellani and D´Ambrosio, 2011). For this purpose, we use either questionnaires that ask directly about respondents' subjective feelings, or composite indicators describing a given phenomenon by means of objective variables rather than subjective judgements. Usually, the result of such an effort is a measurement scale, composite indicator or index. More items from different areas of interest are included and the composite indicator is the result of weighting procedure. The result depends on both inputs – the choice of questions and the weights. There exists a relatively large spectrum of indicators focusing on different populations, areas of interest and data sources. Moreover, there is no widely accepted measure of the quality of life or deprivation of different

[1] University of Economics, Prague, Faculty of Informatics and Statistics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: malai@vse.cz, phone: (+420)224095486.

type or origin. A potential danger when applying and interpreting the above mentioned indicators consists in confusing them with the underlying phenomenon itself. They are just useful constructs, descriptions or quantifiers, reflecting reality, not copying it nor being identical.

As for material deprivation, there are more composite indicators that include other characteristics than just income, allowing to reliably reflect the situation in households. The study covers a sample of the population aged 50 and above. For the targeted group (the elderly population above 50 in particular), social exclusion and isolation along with material deprivation pose a real problem worth critical attention. The indices analyzed in this paper are adapted to the target population.

The aim of the present research is to describe and model deprivation indices for the Czech Republic included in the fifth wave of the Survey of Health, Ageing and Retirement in Europe (SHARE) conducted in 2013, comparing outputs to other European countries.

1 LITERATURE SURVEY

The investigated indices refer to older populations. Due to aging of the European population, general criteria and benchmarks need to be modified, putting more emphasis, for example, on health and retirement, highlighting both negative (deprivation or other disadvantage) and positive (life activity and satisfaction) socio-economic, cultural and psychological factors.

The United Nations Economic Commission for Europe uses the Active Ageing Index (AAI, 2018) as a tool to measure the untapped potential of older people for active and healthy aging worldwide. It monitors and compares the levels of independent living of the elderly, their participation in paid employment and social activities as well as the ability to actively age.

In the SHARE survey, many questions concerning the social and material situation or quality of life are usually asked. The life satisfaction variable CASP (CASP19, 2018), for example, was designed to capture the impact of factors that affect the quality of life of old people. The general scale consists of four sub-scales whose initial letters make up the abbreviation CASP, namely Control, Autonomy, Self-Realization and Pleasure. The CASP values are positive integers in the range from 12 to 48; see Hyde (2003) and CASP19 (2018).

Loneliness and isolation seem to accompany aging of many people. A 20-item scale – the revised UCLA Loneliness Scale – was developed to measure such feelings on an integer scale of 3 (not lonely) to 9 (very lonely) (Russell et al., 1978, 1980). As part of the SHARE survey, this scale was used in the present paper to be compared to the analyzed indices, especially that of social deprivation.

The Indices of Deprivation (ID) provide a set of relative deprivation measures grouped into seven domains for small areas (Lower-layer Super Output Areas) across England for each year (last one 2015). The index is based on the principle of distinct dimensions of deprivation which can be recognized and measured separately, and then are combined into a single complex measure – the overall Index of Multiple Deprivation – using the following weights Income Deprivation (22.5%), Employment Deprivation (22.5%), Education, Skills and Training Deprivation (13.5%), Health Deprivation and Disability (13.5%), Crime (9.3%), Barriers to Housing and Services (9.3%) and Living Environment Deprivation (9.3%) (Ralston, 2014).

The material deprivation index is regularly published by the Czech Statistical Office for the Czech Republic (CZSO, 2018) and by Eurostat for the whole European Union (EUROSTAT, 2018). According to the standard procedure, households that meet three or four out of nine selected material indicator criteria are marked as deprived. In the Czech Republic in 2013, age groups of 50–64 and 65+ had 15.1 and 16.6 percent of deprived households, respectively, satisfying three such criteria items, and 6.6 and 5.3 percent, respectively, meeting four of them. In terms of composite indicators (Saisana et al., 2005), constant weights (equal to 1/9) are used, the indicator equaling the relative frequency of positive items with the deprivation threshold set at 3/9 = 0.33 or 4/9 = 0.44.

The Scottish index of multiple deprivation (in its present 2016 form – SIMD 16) is applied by Scottish local authorities and central government in the most needy areas. The index includes the following domains: current income, employment, health, education, skills and training, housing, geographic access and crime (Ralston, 2014).

If we compare criterion domains used in the above-mentioned scales and analyzed indices (Tables 1 and 2), many common features can be identified.

2 METHODOLOGY

When employing statistical procedures in insurance, zero-inflated models and a mixture of Bernoulli and Poisson distributions are commonly used (Dalrymple et al., 2003) to model the number of insurance claims. In the present study, we apply similar approach to material deprivation.

In the analyzed data set, there are too many zeros for modelling continuous distributions. Therefore, in our model of material deprivation index Y_{mat}, we combine the Dirac measure (discrete part of $Y_{mat} = 0$) with a mixture of normal densities (for positive values of $Y_{mat} > 0$). Let $\pi_0(\mathbf{x}) = P(Y_{mat} = 0|\mathbf{x})$ denote the probability of zero deprivation, given the vector of $m \geq 1$ explanatory variables $\mathbf{x} = (x_1, x_2, ..., x_m)'$. The logistic regression model is applied in the form:

$$\operatorname{logit}\left(P\left(Y_{mat} = 0|\mathbf{x}\right)\right) = \operatorname{logit}\left(\pi_0\left(\mathbf{x}\right)\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)'$ is an $(m + 1)$-dimensional vector of unknown regression parameters. Using this notation, the normal mixture model with K_{mat} components is given by:

$$f_{mat}\,(y_{mat}|\mathbf{x}) = \pi_0(\mathbf{x}) + (1 - \pi_0(\mathbf{x}))\sum_{j=1}^{K_{mat}} \pi_j f_{j,\,mat}(y_{mat}), \tag{1}$$

where $0 \leq \pi_j \leq 1, j = 1, 2, ..., K_{mat}, \sum_{j=1}^{K_{mat}} \pi_j = 1$ and $f_{j,\,mat}, j = 1, 2, ..., K_{mat}$ are normal component densities specified by the component parameters (μ_j, σ_j^2). In the model, there are $(m + 1)$ logistic regression parameters, $2K_{mat}$ component parameters and $K_{mat} -1$ component probabilities, which makes a total of $m + 3K_{mat}$ unknown parameters.

For the social deprivation index Y_{soc} we use the normal mixture model with density f_{soc} given by:

$$f_{soc}\,(y_{soc}) = \sum_{j=1}^{K_{soc}} \pi_j\, f_{j,\,soc}\,(y_{soc}), \tag{2}$$

where K_{soc} is the number of components, $0 \leq \pi_j \leq 1, j = 1, 2, ..., K_{soc}, \sum_{j=1}^{K_{soc}} \pi_j = 1$ and $f_{j,\,soc}, j = 1, 2, ..., K_{soc}$ denoting normal component densities. In the model, we have $2K_{soc}$ component parameters and $K_{soc} -1$ component probabilities. It means $3K_{soc} -1$ unknown parameters altogether.

There are no exact rules for determining the number of components. Based on the histogram, we use two components to be included in the social deprivation index. For the material deprivation index, we took the AIC criterion along with the numerical stability of the solution into account. From this point of view, the choice of three components is an acceptable compromise.

For all calculations, statistical computing R software (R Core Team, 2013) was used. In order to obtain parameter estimates, the mixtool package (Benaglia et al., 2009) was applied. Maximum likelihood estimates were evaluated, bootstrap being used (1 000 replications drawn from the data) to estimate the standard deviations. GLM function with a binomial link function was employed to estimate a logistic regression model. ANOVA was applied to test the significance of explanatory variables.

In the SHARE survey, the weights of the individual data are provided for both the cross-section and longitudinal origin. In the analysis of deprivation, cross-section weights for the fifth wave of the survey can be included.

3 RESULTS

3.1 SHARE database and indices of deprivation

The Survey of Health, Ageing, and Retirement in Europe (SHARE, 2018; Börsch-Supan, 2016) is a multidisciplinary, cross-national panel database of microdata of the European population aged over 50. Currently, data from six waves (between 2004 and 2015) are available, the module of deprivation (based on Adena et al., 2015) was included only in the fifth wave in 2013. The survey in the 5th wave took place in 14 European countries and Israel (IL) – apart from the Czech Republic (CZ), also in Austria (AT), Belgium (BE), Denmark (DK), Estonia (EE), France (FR), Germany (DE), Italy (IT), Luxembourg (LU), the Netherlands (NL), Slovenia (SI), Spain (ES), Sweden (SE) and Switzerland (CH).

Values of the two composite indicators for social and material deprivation from the above-mentioned module are used to model the distribution of both indices for the Czech Republic in 2013. The index of material deprivation is an aggregate measure of material conditions of Europeans aged over 50 years, comprising a set of 11 criterion items that refer to two broad domains: the inability to meet basic needs and financial difficulties (Table 1). The index includes only the information based on the situation of households, its value being equal for all members of the household. The index of social deprivation utilizes 15 survey questions representing a social dimension (Table 2), combining information on items related to participation in everyday life, social activities and the quality of the neighborhood; 8 criteria apply to households, 7 to individuals (Table 2, items 1, 5–11 and 2–4, 12–15, respectively).

Alternative answers yes (in case of problems) or no (if there are no problems) are weighted to the composite indices. Both indicators are transformed into a <0, 1> scale from none (0) to the highest (1) degree of deprivation.

Both indicators can be used together to identify Europeans suffering from both material and social deprivation, this joint index diagnosing the so-called severe deprivation. In the SHARE project, the threshold for material or social deprivation is set for the lower quartile of all data (across the European

Table 1 Material deprivation index criteria

1	Your household does not eat meat, fish or chicken more often than three times per week because you cannot afford it.
2	Your household does not eat fruits or vegetables more often than three times per week because you cannot afford to eat it.
3	Can your household afford to regularly buy necessary groceries and household supplies?
4	Could your household afford to go for a week-long holiday away from home at least once a year?
5	Could your household afford to pay an unexpected expense without borrowing any money?
In the last twelve months, to help you keep your living costs down, have you	
6	continued wearing clothing that was worn out because you could not afford replacement?
7	continued wearing shoes that were worn out because you could not afford replacement?
8	put up with feeling cold to save heating costs?
9	gone without or not replaced glasses you needed because you could not afford new ones?
10	postponed visits to the dentist?
11	Was there a time in the past 12 months when you needed to see a doctor but could not because of the cost?

Source: SHARE Release Guide 6.0.0 (2018) population) to indicate the upper limit of 25 percent of the lowest values. These values for material and social deprivation are, 0.220 and 0.224, respectively. They both are lower than country-specific values for the Czech Republic, namely 0.261 and 0.375. The data for the Czech Republic show that material and social thresholds were exceeded by 26 (0.7) and more than 50 (0.7) percent of respondents, respectively. Severe deprivation occurs if people are deprived both materially and socially. In the whole data set, 11 percent of severely deprived respondents were identified, the relative frequency in the Czech Republic being only 4.6 percent.

Table 2 Social deprivation index criteria

1	Less than one room per person in the household
2	Poor reading or writing skills
3	Poor computer skills or never used a computer
4	Not feeling part of the local area
5	Vandalism in the local area
6	Local area not clean
7	No helpful people in the local area
8	Difficult access to the bank
9	Difficult access to grocery shop
10	Difficult access to a pharmacy
11	Waiting too long to see a doctor
12	Not attending any course in the past 12 months
13	Not taking part in any organization in the past 12 months
14	People cannot be trusted
15	Feeling left out of things

Source: SHARE Release Guide 6.0.0 (2018)

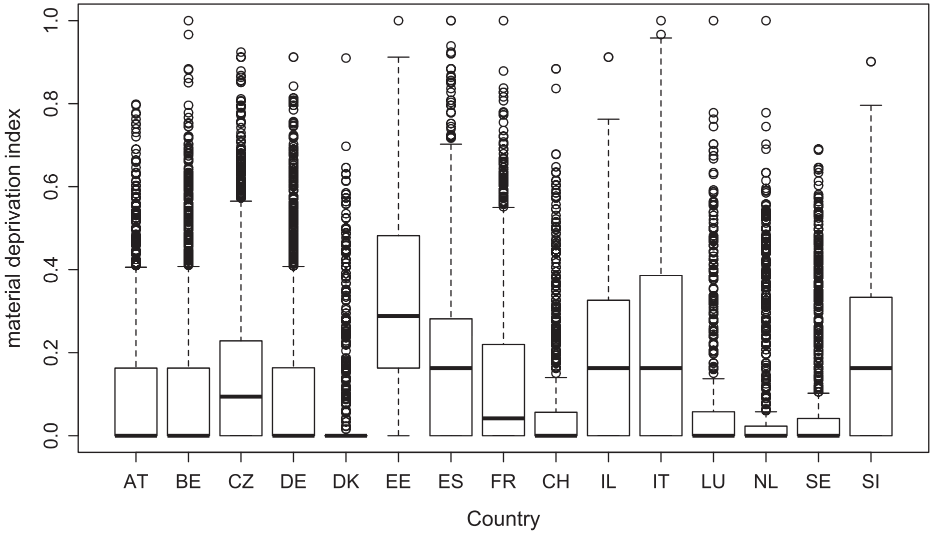
3.2 Deprivation statistics – the European Union

The empirical distribution of sample values is highly country-specific. Box plots for material and deprivation indices are shown in Figures 1 and 2, respectively. For material deprivation, median values are equal to zero, more than half of Austrian, Belgian, Danish, Swiss, Luxembourg, Dutch and Swedish respondents reporting no deprivation. The highest median is reported by Estonia followed by a group of countries with similar median values (Spain, Italy, Slovenia and Israel).

Social deprivation index outcomes are different, all lower quartiles being positive. The lowest and highest median values were recorded in Denmark and the Czech Republic, respectively.

It is possible to compare the analyzed indices to CASP and UCLA Loneliness scales. There is a relatively high negative correlation between CASP and the indices, respectively, –0.404 and –0.559 for social and material deprivation. Dependence between the loneliness variable and the analyzed indices, on the other hand, is expected to be relatively high, Spearman's correlation coefficient (due

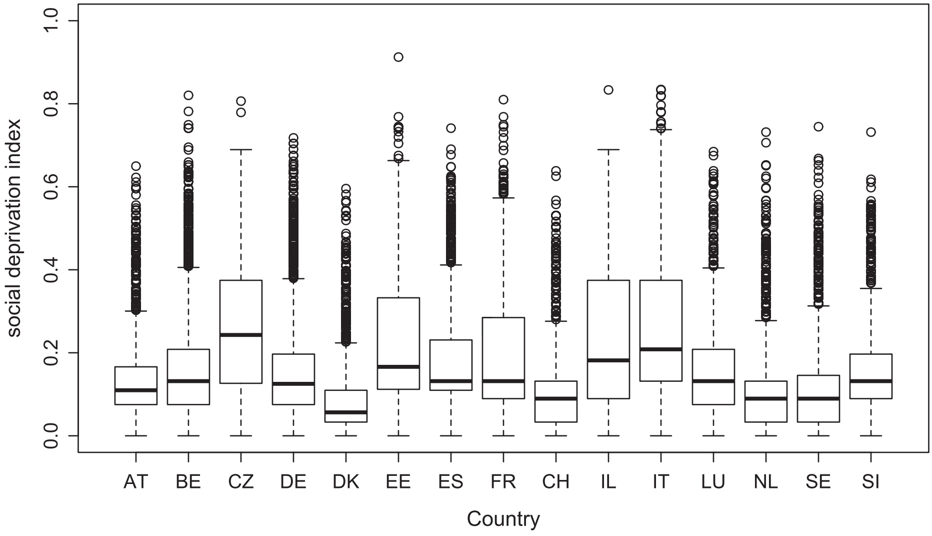
Figure 1 Comparison of distributions of the material index in analyzed countries



Source: Own computations

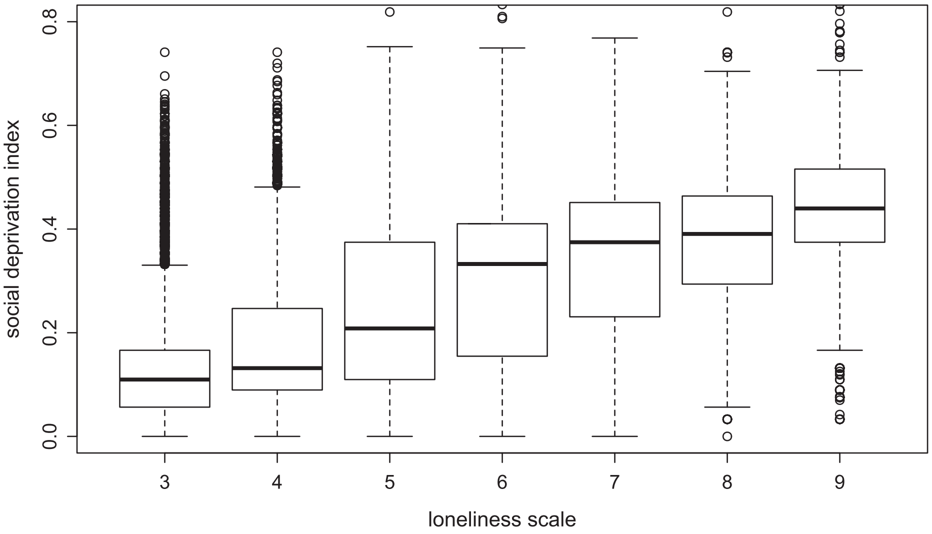
to the discrete distributions) equaling 0.378 and 0.175 for social and material deprivation, respectively. In Figure 3, the nonlinear relationship between the loneliness scale and the social index is obvious. The focus of both the above indicators is similar, while the areas of the index analyzed in this study are much broader.

Figure 2 Comparison of distributions of the social index in analyzed countries



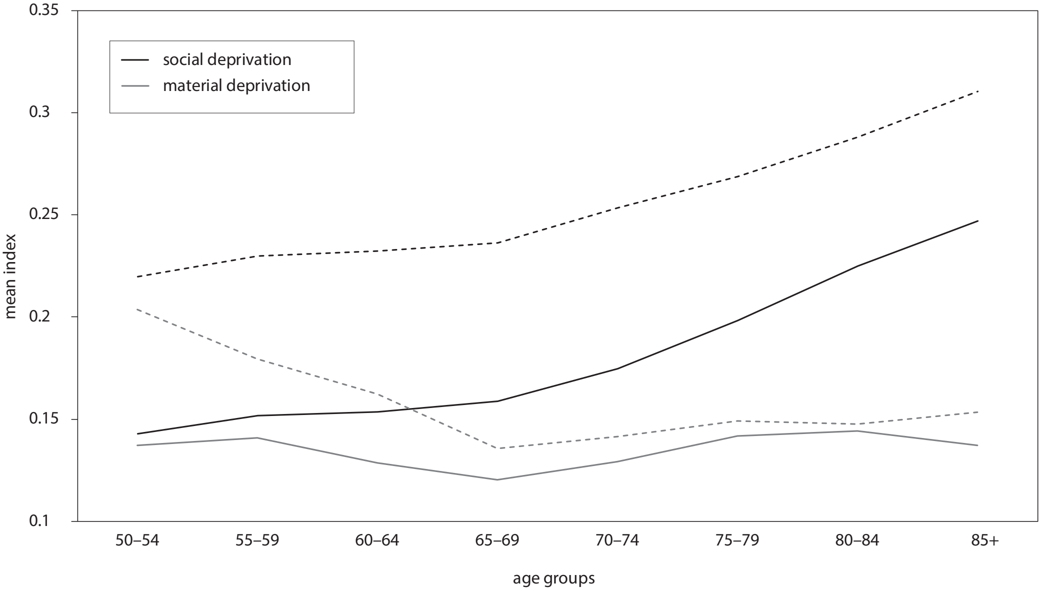
Source: Own computations

Figure 3 Comparison of loneliness scale and social deprivation index



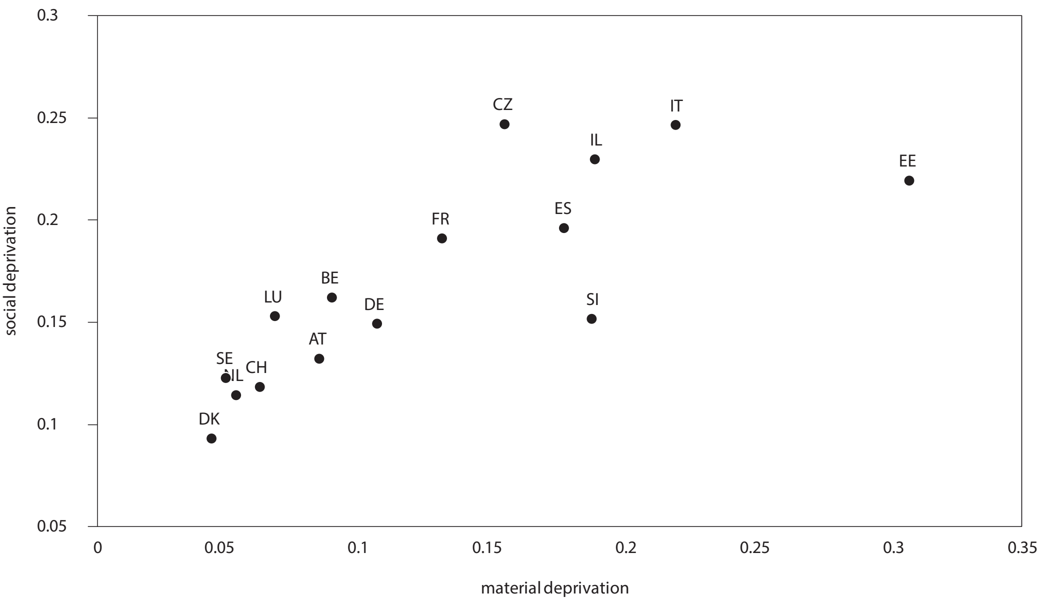
Source: Own computations

Figure 4 Mean indices for the EU (solid lines) and the Czech Republic (dashed lines)



Source: Own computations

Figure 5 Mean indices for European countries



Source: Own computations

Thanks to European welfare states' care for aging populations, the dependence of both indices on age is rather weak, the mean material deprivation index not rising with increasing age. Figure 4 displays data for all countries and for the Czech Republic, grouping respondents into five-year age groups. For the Czech Republic, almost the highest value of mean material deprivation was recorded in the group of active people between 50 and 60 years of age – obviously because of the problems these people face in the labor market – the mean apparently being independent of time. The social deprivation index, on the other hand, signals deterioration in living conditions, its values clearly increasing with age.

Spearman's coefficient of correlation between both indices is equal to 0.39, in the Czech Republic, however, we obtain only 0.22. Both values of coefficient are highly statistically significant because of the research sample size. Mean indices for all countries are given in Figure 5. In the bottom left corner of the chart are the "old" EU members (Denmark, the Netherlands, Sweden, Luxemburg, Austria, Belgium, Germany) and Switzerland. Another, less homogenous group of countries (the worse-off ones) consists of the Czech Republic, Italy, Spain, Slovenia and Israel. France ranks between both groups, Estonia, due to its high material deprivation (Figure 1), being an outlier in the set of the countries analyzed.

3.3 Deprivation statistics – the Czech Republic

In this section, we will assess the situation in the Czech Republic. From 5 646 fifth-wave survey respondents, we took those aged 50 and above with information of both types of deprivation. We obtained $n = 3\ 954$ respondents with the mean age of 67.8 years (standard deviation = 8.7). The sample includes 2 289 women (57.6 %) and 1 685 men (42.4 %) with age means of 67.5 (8.8) and 68.2 (8.6), respectively.

Table 3 compares the age structure of the 2013 population in the Czech Republic (CZSO, 2018) with that of the research sample data on the 50-plus population divided into labor-active and inactive groups, 50–64 and 56+ years of age, respectively.

Table 3 Comparison of CR and research sample population (2013)

group	**50+**		**50–64**		**65+**	
	population	share (%)	population	share	1 980	2003
population	3 885 926	37.0	2 089 667	53.4	1 796 259	46.3
males	1 762 591	34.1	1 024 944	58.1	737 647	41.9
females	2 123 335	39.7	1 064 723	50.0	1 058 612	50.0
males / females (%)	45/55		49/51		41/59	
sample	3 974		1 542		2 432	
males	1 685		617		1 068	
females	2 289		925		1 364	
males / females (%)	42/58		40/60		44/56	

Source: CZSO, own computations

Table 4 Descriptive statistics of analyzed indices (SD = standard deviation, lq = lower quartile, uq = upper quartile)

Index	**Mean**	**Median**	**SD**	**lq**	**uq**
y_{mat}	0.159	0.114	0.186	0.000	0.261
$y_{mat} > 0$	0.267	0.220	0.171	0.163	0.383
y_{soc}	0.248	0.243	0.144	0.126	0.375

Source: Own computations

Descriptive statistics of the sample are presented in Table 4. 1 640 respondents show a zero value of the material deprivation index, indicating no problem in any of the areas studied. The table therefore shows the frequency of zeros, along with all numerical characteristics and separate positive values. In terms of social deprivation, however, only 25 observations are equal to zero.

In the logistic regression model for the material deprivation we use three explanatory variables: gender, household size and NUTS3 regions of domicile (the latter being of particular importance in the Czech Republic). The reference combination of explanatory variables consists of a man living in a two-person household in the Central Bohemian Region. The age variable, both continuous and discrete, was excluded from the research due to its low explanatory power. In Table 5, the estimated parameters are given along with standard deviations, exponential transformations to odds and p-values for the test of zero parameters. All three explanatory variables (gender, household size, domicile) are statistically significant in the model (the ANOVA table not presented herein).

The components in the mixture model (1) with $K_{mat} = 3$ are artificial as the component membership is not observable. Three centers 0.153, 0.372 and 0.495 were identified. They might be interpreted as the component of low deprivation and medium deprivation. The third component (21% of the continuous part of the distribution) with the highest standard error includes those with relatively high deprivation to describe a tail of the material index distribution. The empirical distribution of the continuous part of the material deprivation index has two local modes, the selected three normal component model allowing for modelling a relatively high positive skewness.

All weights (given in Table 6) should be multiplied by the probability complementary to the $\hat{\pi}_0(\mathbf{x})$ where the vector of explanatory variables $\mathbf{x}$ depends on a particular respondent. We performed 1 000 bootstrap iterations (using *mixtools* package (Benaglia et al., 2009)) in order to estimate standard errors of the estimated parameters.

In the case of the social deprivation index, two subgroups (of low and high deprivation levels) were identified. The levels of deprivation in the mixture components are estimated to reach 0.112 and 0.367, respectively. 98 percent for the low deprivation component and 96 percent of the high deprivation component are below and above the European social deprivation limit 0.224, respectively. The European bandwidth 0.224 well distinguishes both components, the fit corresponds to the empirical frequencies.

Table 5 Logistic regression model results

Coefficients	$\hat{\beta}$	$S_{\hat{\beta}}$	$\exp(\hat{\beta})$	*p*-value
intercept	0.156	0.115		0.176
gender female	–0.202	0.069	0.816	0.003
hh size 1	–0.403	0.087	0.668	$<10^{-6}$
hh size 3	–0.236	0.105	0.789	0.025
hh size 3+	–0.177	0.132	0.837	0.181
nuts3 Hradec Kralove	–0.560	0.148	0.570	$<10^{-3}$
nuts3 South Bohemian	–0.099	0.161	0.905	0.539
nuts3 Zlin	–0.863	0.163	0.421	$<10^{-5}$
nuts3 Karlovy Vary	–0.152	0.264	0.858	0.564
nuts3 Liberec	0.342	0.228	1.408	0.134
nuts3 Moravian-Silesian	–0.386	0.135	0.679	0.004
nuts3 Olomouc	0.157	0.187	1.170	0.402
nuts3 Pardubice	–0.573	0.210	0.563	0.006
nuts3 Plzen	–0.209	0.183	0.810	0.253
nuts3 Prague	0.075	0.148	1.078	0.609
nuts3 South Moravian	0.055	0.152	1.056	0.718
nuts3 Usti nad Labem	–0.568	0.159	0.566	$<10^{-3}$
nuts3 Vysocina	–0.492	0.193	0.610	0.011

Source: Own computations

Table 6 Estimates of component distribution parameters for both mixture models; point estimate (1st row) and standard deviation (precision, 2nd row)

Material deprivation

Component 1			**Component 2**			**Component 3**		
π_1	μ_1	σ_1	π_2	μ_2	σ_2	π_3	μ_3	σ_3
0.604	0.153	0.068	0.213	0.372	0.070	0.183	0.495	0.164
0.024	0.003	0.002	0.057	0.008	0.012	0.056	0.060	0.022

Social deprivation

π_1	μ_1	σ_1	π_2	μ_2	σ_2
0.472	0.112	0.050	0.527	0.367	0.079
0.008	0.001	0.001	0.008	0.002	0.001

Source: Own computations

CONCLUSION

In the text, two composite indicators of material and social deprivation for the Czech population aged above 50 based on the SHARE survey are analyzed. The indices of material deprivation published periodically by the Czech Statistical Office contain also results for 50–64 and 65 and above age groups. They are consistent with those based on the SHARE panel data, given the possible comparisons. In the SHARE survey, however, not all questions asked to form the CZSO material deprivation index are included, thus it is not possible to compare both approaches on the level of individual values.

All conclusions are based on the particular composite indicators used in the analysis. The interpretation should be limited to them and their ability to describe deprivation in the population of elderly inhabitants of the EU.

For the Czech Republic, empirical distributions of the two indices are very different. In terms of social deprivation, the distribution is bimodal and the mixture of the two normal components is well applicable. The model identifies two clearly distinct (artificial) subgroups. The material index acquires zero value for over 40 percent of respondents. Therefore, the mixture of one Dirac measure (its parameter being estimated by logistic regression) and three normal distributions was applied to model a discrete part and a continuous, highly positively skewed model.

The Czech Republic seems to be really non-homogenous with respect to the material deprivation index. This fact and the analysis of differences are in agreement with the well-known diversity of region with respect to the quality of life and economic problems.

In terms of material deprivation, the Czech Republic and other European countries are comparable. Values of the social deprivation index, on the other hand, are higher in the Czech Republic, the mean approximately equaling the upper quartile in other European countries. The growth of social deprivation is common to all participating countries, but in the Czech Republic is even steeper (Figure 4). In Figure 5, the position of the Czech Republic in the participating countries is shown. The mean value of a material index is higher than those in a cluster of "old" European countries, but it is lower than for Spain, Italy, Israel or Slovenia and Estonia (see also Figure 1 for medians or quartiles). The median of the deprivation index is the highest from all counties in the survey (Figure 2). The mean for the deprivation index is comparable to the worst values of Estonia, Italy and Israel. It follows, that the problem of social deprivation of elderly population seems to be more serious in the Czech Republic than material deprivation.

The deprivation module being part of the fifth wave of the SHARE research (the only phase that used the questionnaire defining the deprivation indices), the results of the present study, unfortunately, cannot be compared to other stages of the panel survey, thus lacking the time dimension.

ACKNOWLEDGMENTS

Institutional support from the funds for the long-term conceptual advancement of science and research, number IP 400 040, at the Faculty of Informatics and Statistics, University of Economics, Prague, Czech Republic is gratefully acknowledged.

References

AAI. *Active ageing index* [online]. [cit. 12.7.2018] <https://statswiki.unece.org/display/AAI/Active+Ageing+Index+Home>.

ADENA, M., MYCK, M., OCZKOWSKA, M. Innovation for better understanding deprivation index. In: *Ageing in Europe – Supporting Policies for an Inclusive Society*, De Gruyter, 2015.

BELLANI, L. AND D´AMBROSIO, C. Deprivation, social exclusion and subjective well-being. *Social Indicators Research*, 2011, pp. 67–86.

BELLANI, L. Multidimensional indices of deprivation: the introduction of reference groups weights. *Econ Inequal*, 2013, pp. 495–515.

BENAGLIA, T., CHAUVEAU, D., HUNTER, D., YOUNG, D. Mixtools: An R Package for Analyzing Finite Mixture Models [online]. *Journal of Statistical Software*, 2009, 32, pp. 1–29. <http://www.jstatsoft.org/v32/i06>.

BÖRSCH-SUPAN, A. *Survey of Health, Ageing and Retirement in Europe (SHARE) Wave 5.* Release version: 5.0.0, SHARE-ERIC, Data set, 2016.

CASP19. *Control, Autonomy, Self-Realization and Pleasure. Measuring quality of life in later ages* [online]. [cit. 25.6.2018] <https://casp19.com/background>.

CZSO. *Household Income and Living Conditions – 2013* [online]. Prague: CZSO. [cit. 20.4.2018] <https://www.czso.cz/csu/czso/household-income-and-living-conditions-2013-ia0fwqxyxa>.

DALRYMPLE, M. L., HUDSON, I. L., FORD, R. P. K. Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Computational Statistics & Data Analysis*, 2003, 41(3–4), pp. 491–504.

Europe 2020 Strategy [online]. 2017. [cit. 12.4.2018] <https://ec.europa.eu/info/strategy/european-semester/framework/europe-2020-strategy_en>.

EUROSTAT. *Depth of material deprivation – EU-SILC survey* [online]. [cit. 12.4.2018] <http://ec.europa.eu/eurostat/web/productsdatasets/tessi150http://ec.europa.eu/eurostat/tgm/refreshTableAction.do?tab=table&plugin=1&pcode=tessi082&language=en>.

HYDE, M. A measure of quality of life in early old age: The theory, development and properties of a needs satisfaction model (CASP19). *Aging and mental health*, 2003, 7(3), pp. 186–194.

MYCK, M., NAJSZTUB, M., OCZKOWSKA, M. Measuring social deprivation and social exclusion. In: *Ageing in Europe – Suporting Policies for an Inclusive Society*, De Gruyter, 2015.

R CORE TEAM. R: *A language and environment for statistical computing* [online]. Vienna, Austria: R Foundation for Statistical Computing, 2013. <https://www.R-project.org>.

RALSTON, K., DUNDAS, R., LEYLAND, A. H. A comparison of the Scottish Index of Multiple Deprivation (SIMD) 2004 with the 2009 + 1 SIMD: does choice of measure affect the interpretation of inequality in mortality? *International Journal of Health Geographics*, 2014, pp. 13–27.

RUSSELL, D., PEPLAU, L. A., FERGUSON, M. L. Developing a measure of loneliness. *Journal of Personality Assessment*, 1978, 42, pp. 290–294.

RUSSELL D., PEPLAU L. A, CUTRONA CAROLYN, E. The Revised UCLA Loneliness Scale: Concurrent and Discriminant Validity Evidence. *Journal of Personality and Social Psychology*, 1980, pp. 472–80.

SAISANA, M., SALTELLI, A., TARANTOLA, S. Uncertainty and Sensitivity Analysis Techniques as Tools for the Quality Assessment of Composite Indicators. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 2005, 168, pp. 307–323.

SHARE Release Guide 6.0.0 [online]. [cit. 18.4.2018] <http://www.share-project.org/data-documentation/waves-overview/wave-6.html>.

Use of Logistic Regression for Understanding and Prediction of Customer Churn in Telecommunications

Jan Manďák[1] | *VŠB-Technical University of Ostrava, Ostrava, Czech Republic*
Jana Hančlová[2] | *VŠB-Technical University of Ostrava, Ostrava, Czech Republic*

Abstract

Customer churn, loss of customers due to switch to another service provider or non-renewal of commitment, is very common in highly competitive and saturated markets such as telecommunications. Predictive models need to be implemented to identify customers who are at risk of churning and also to discover the key drivers of churn. The aim of this paper is to use demographic and service usage variables to estimate logistic regression model to predict customer churn in European Telecommunications provider and to find the factors influencing customer churn. An interesting findings came out of the estimated model – younger customers who are shorter time with company, who use mobile data and sms more than traditional calls, having occasional problem with paying bills, with students account and ending contract in the near future are typical representatives of customers who tend to leave the company.

An interaction terms added as explanatory variables showed that effect of usage of data and voice vary depending on the year of birth. The quality of the logistic regression model was assessed by Hosmer-Lemeshow test and pseudo R squared measures. An independent testing data set was further used to evaluate the predictive ability of the model by computation of performance metrics such as the area under the ROC curve (AUC), sensitivity and precision. The resulting model was able to catch 94.8% of customers who in fact left the company. Quality of the model was confirmed also by high value of AUC metric equal to 0.9759. Logistic regression represents a very useful tool in prediction of customer churn not only thanks to its interpretability, but also for its predictive power.

Keywords

Customer churn, telecommunications, predictive analytics, logistic regression, sensitivity

JEL code

C35, C38, C53, L96

[1] VŠB-TU Ostrava, Faculty of Economics, Department of Systems Engineering, Sokolská třída 33, 702 00 Ostrava, Czech Republic. E-mail: jan.mandak@gmail.com, phone: (+420)737674192.

[2] VŠB-TU Ostrava, Faculty of Economics, Department of Systems Engineering, Sokolská třída 33, 702 00 Ostrava, Czech Republic. E-mail: jana.hanclova@vsb.cz, phone: (+420)597322285.

INTRODUCTION

With continuously decreasing costs of data storage, telecommunications companies have access to various data sources, which can be beneficial to various types of customer analyses. Traditional transactional data stored in databases can be combined with unstructured data such as complaints or feedback scraped from social networks or call recordings. These data are used to create predictive models with the use of algorithms such as logistic regression, decision trees, random forests or neural networks. Predictive models can help decision makers to identify customers who are likely to churn (Balasubramanian and Selvarani, 2014). Telecommunications companies can then offer customers new incentives to stay. But it is not sufficient to predict who is likely to churn, but also what are the key factors causing the churn. With this knowledge in hands marketing departments can target retention campaigns to the right customer groups and also change the whole range of services.

The goal of this paper is to predict the customer churn in European Telecommunications Company with the use of logistic regression. The estimated model should reveal especially the key factors leading customers to leave the company. Another aim of the paper is also to assess quality of both logistic regression model and its predictions. This paper should also confirm the suitability of the usage of logistic regression for customer churn prediction.

The theoretical background of customer churn, its types, dimensions, consequences and benefits of proper churn management process as well as review of current literature are described in the introductory part of the paper. The methodological part contains definition of logistic regression, maximum likelihood method used for estimation of beta coefficients, statistical tests such as Wald test or Hosmer Lemeshow test, pseudo R-squared measures, ROC curve and performance statistics sensitivity and precision. Then, the variables used for creation of the model are at first introduced, interesting patterns explored during graphical analysis are shown and finally estimation results and results of statistical tests together with performance metrics are listed. The most interesting results as well as comparison with similar studied are content of the last part of the paper.

1 CUSTOMER CHURN

The term *"customer churn"* is used to indicate those customers who are about to leave for a competitor or end their subscription. Customer churn or customer attrition has become an important issue for organizations particularly in subscription based businesses, where customers have a contractual relationship which must be ended. Customer churn in telecommunications industry is really hot topic, because it is saturated market and where it is difficult to attract new customers and because it is relatively easy to switch to another company (Canale and Lunardon, 2014). It is generally accepted that acquiring a new customer costs five to six times more than retaining the existing one. For telecom operators it's preferable to invest into existing customers and renew their trust, rather than attract new ones characterized by a higher churn rate. *Churn management* is the process of identifying customers, who are valuable for company and are likely to churn, and taking proactive steps to retain them. The measurement for the number of customers moving out during a specific period of time is called *Churn rate*. People responsible for the churn management process should take care of these three dimensions: *WHO* (which customers are likely to churn), *WHEN* (will the customers churn in a week, month or year?) and *WHY* (what are the reasons of customer churn).

It is necessary to distinguish between two types of churn (Lazarov and Capota, 2007). In case of *voluntary churn* – the customer decides to cancel his contract and to switch to another provider. For companies it is necessary to know the reason of churn before applying the right retention strategy. Reasons for churn may include dissatisfaction with the quality of the service, too high costs, not competitive price plans, no rewards for customer loyalty, bad support, long time of problem solutions, privacy concerns, etc. *Involuntary churn* is a situation when the company discontinues the contract itself, e.g. because of fraud or non-payment. This type of churn can be healthy for a company, because company loses non-profitable or problem-causing customers.

Dahiya and Talwar (2015) emphasize that machine learning models work well if there is enough time spent to prepare meaningful features. Thus, having the right features is usually the most important thing. With the still decreasing costs of data storage, telecommunications companies have access to various data sources, which can be beneficial for analysis of customer churn. It is therefore necessary to invest time into feature engineering, because well prepared features can also help us identify the reasons of churn.

Proper churn management can undoubtedly save company a huge amount of money. Van den Poel and Lariviere (2004) summarize the economic value of customer retention, which may have several benefits: satisfied customers can bring new ones and share positive references, long-term customers tend to buy more and are less sensitive to competitors marketing offers and company can also focus on satisfying the existing customer's needs.

There are many approaches applicable to distinguish churners and non-churners, such as association rules, classification models, clustering, or various types of visualization. Logistic regression, together with decision trees and neural networks belongs to the most frequently used classification methods for churn prediction. Several authors have used logistic regression to detect potential churners in Telecommunications.

Olle and Cai (2014) gathered dataset of 2000 subscribers from an Asian Telecommunications operator. Location, Age, Tenure or Tariff were some of their explanatory variables. Logistic regression model estimated in WEKA achieved precision 0.72 and recall 0.75. Gürsoy (2010) analyzed churn in a major Telecommunication firm in Turkey. Logistic regression model was estimated in SPSS Clementine software. The following explanatory variables were confirmed as statistically significant using Wald test: discount package, customer age or average length of call. Ahn et al. (2006) investigated determinants of customer churn in the Korean mobile telecommunications service market using logistic regressions. Their results indicate that call quality-related factors influence customer churn. Also heavy users are more likely to leave the service provider. Sebastian and Wagh (2017) gathered dataset with over 2000 customers described by 22 variables. They achieved accuracy 0.8 by the use of backward stepwise logistic regression. They also made the results more understandable by visualization in Power BI. Modelling telecom customer attrition is important also in African countries, Oghojafor et al. (2012) state that e.g. in Nigeria the annual churn rate come up to 41%. They created a well-structured and compliant questionnaires and obtained 6000 subscribers of Telecom service providers. They distinguish churners from non-churners by questioning respondents whether they would you like to change their current service provider. Stepwise logistic regression model revealed that call expenses, providers' advertisement medium, type of service plan, number of mobile connections and providers' service facilities are the most important factors driving churn.

2 LOGISTIC REGRESSION

Logistic regression is member of a class of models called *generalized linear models* (Zumel and Mount, 2014). The aim of generalized linear models for a binary dependent variable is to estimate a regression equation that relates the expected value of the dependent variable y to one or more predictor variables, denoted by $\boldsymbol{x}$ (Heeringa et al., 2010). A naïve approach is to model y as a linear function of $\boldsymbol{x}$, but linear regression doesn't capture the relationship between y and $\boldsymbol{x}$ and moreover it may produce predictions that are outside the permissible range 0-1. A better alternative is a nonlinear function that yields a regression model that is linear in coefficients and it is possible to transform the resulting predicted values to the range 0-1. These functions are called in the terminology of generalized linear models link functions (Heeringa et al., 2010). The two most common link functions used to model binary survey variables are the *logit* and the *probit*. The logit, natural logarithm of the odds, can be modelled by a linear regression model:

$$ln\left(\frac{\pi(x)}{1-\pi(x)}\right)= \beta_0+\beta_1 x_1+\cdots+\beta_{kp}x_k\,, \tag{1}$$

where $\pi(\boldsymbol{x})$ is the conditional probability that $y = 1$ given the covariate vector $\boldsymbol{x}$, β_0, β_1, … , β_k are estimated regression coefficients of the logit model and x_1, x_2, … , x_k are the explanatory variables. The left-hand side of the Formula (1) is called the *log-odds* or *logit* and can take values from the interval (-∞; ∞). The expression $\frac{\pi(x)}{1-\pi(x)}$ is called the *odds* and can take on any value between 0 and ∞. Values close to 0 indicate very low and values close to ∞ indicate very high probability.

The usual practice after estimation of the model coefficients is to assess the significance of the explanatory variables (Hosmer and Lemeshow, 2000). *Wald test* can be used to test the statistical significance of the coefficients $\boldsymbol{\beta}$ in the model. Wald test calculates a Z statistic (2), which is for i-th variable computed as:

$$Z=\frac{\widehat{\beta}_i}{\widehat{SE}(\widehat{\beta}_i)}\,, \tag{2}$$

where $\widehat{SE}(\widehat{\beta}_i)$ is an estimated standard error of the estimated regression coefficient $\widehat{\beta}_i$. This Z value is then squared, yielding a Wald statistic with a chi-square distribution.

An important step after the model building is assessing the fit of the model. It can be done using e.g. the *Hosmer-Lemeshow test*. The Hosmer-Lemeshow test is a goodness of fit test, which tells how well the model fits the data. Specifically, it calculates if the observed event rates match the expected event rates in population subgroups. Data is first regrouped by ordering the predicted probabilities and forming the number of groups, g. The Hosmer-Lemeshow test statistic (3) is calculated with the following formula:

$$G^2_{HL}=\Sigma^{g}_{j=1}\,\frac{(O_j-E_j)^2}{E_j(1-E_j/n_j)}\,, \tag{3}$$

where n_j = number of observations in the j^{th} group, O_j = number of observed cases in the j^{th} group and E_j = number of expected cases in the j^{th} group. This statistics follows a *chi-squared* distribution with (g – 2) degrees of freedom.

There have also been proposed several *pseudo-R^2 measures* for measuring predictive power of logistic regression. The most often used R squared measures in statistical software's appear to be one proposed by McFadden (1974) and Cox and Snell (1989) along with its corrected version known as Nagelkerke's R^2 (Nagelkerke, 1991). All three measures are based on comparison of the value of likelihood function for model with no predictors and model being estimated.

For the classification tasks where the class which we want to predict is much less frequent, the most known performance metric – *accuracy* – is not sufficient. Two other metrics, *precision* and *sensitivity (recall)*, are much more important. *Accuracy*, *precision or sensitivity* can be computed from the information available in *confusion matrix* (Lantz, 2013), which categorizes predictions according to whether they match the actual value in the data. One dimension indicates the possible categories of predicted values while the other dimension indicates the same for actual values. There are four categories in 2 x 2 confusion matrix (Lantz, 2013):

- True Positive (TP): Correctly classified as the class of interest;
- True Negative (TN): Correctly classified as not the class of interest;
- False Positive (FP): Incorrectly classified as the class of interest;
- False Negative (FN): Incorrectly classified as not the class of interest.

From the business point of view the two most important predictive measures are *sensitivity* and *precision*. By *precision* we mean the ratio of correct predictions, where the model predicts that the customers should churn. A company should avoid low values of *precision*, because it means that the money is spent in retention campaigns to those customers, who would stay regardless beneficial retention offer. It is computed as true positives divided by sum of true positives and false positives:

$$Precision = \frac{TP}{TP + FP}. \tag{4}$$

Sensitivity in our case measures the ratio of customers, who churned, and the model was able to predict them. Formally it is calculated as the ratio of true positives and sum of true positives and false negatives:

$$Sensitivity\ (recall) = \frac{TP}{TP + FN}. \tag{5}$$

Another metric, *Specificity*, is the proportion of correctly classified non-churners. Formally it is calculated as the ratio of true negatives and sum of true negatives and false positives:

$$Specificity = \frac{TN}{TN + FP}. \tag{6}$$

The overall performance of a classifier summarized over all possible thresholds is given by the *area under the (ROC) curve (AUC)*. *AUC* ranges from 0.5 (classifier with no predictive value) to 1 (a perfect classifier).

To sum up, the first and probably the most important and difficult thing is to gather appropriate input variables. Then, an exploratory data analysis is necessary to get an idea of which variables could be useful for classification. Further the model is estimated and its quality is tested e.g. using Wald test, Hosmer-Lemeshow test or pseudo R-squared measures. The final step is the computation of predictive performance measures like sensitivity or AUC with the use of independent test data set.

3 RESULTS AND DISCUSSION

This section of the paper is devoted to application of logistic regression on two real data sets of approximately 50 000 customers from a European Telecommunications. Two independent data sets were available, training data set, whose purpose is to train classification models, and testing data set to evaluate the predictive performance. The input variables divided into demographic group and service usage group are described at the beginning of this part. Then, some interesting patterns uncovered during exploratory data analysis are shown. The following step is to estimate a logistic regression models using training data set. The estimated model is further described, explained and tested. The testing data set is then used to calculate performance statistics to check the behavior of the model on unseen data.

3.1 Description of the data

The input data for modelling were downloaded from relational tables stored in company data warehouse. Two data sets, both with roughly 50 000 different customers, were randomly selected from the population of approx. 1 million customers. As mentioned earlier, the first one was used to estimate logistic regression model while the second one was left for testing of model's performance. The variables which could help reveal leaving customers were thoroughly selected on the basis of cooperation with customer service managers and IT database experts. Customer managers wanted to see mainly two categories of data – demographic data such as age, customer lifetime or type of account (see Table 1) and service usage data such as consumption of mobile data, voice or monthly invoice paid (see Table 2). They had another interesting suggestions for input variables such as number of calls to other networks, but this information was not easily extractable from the data warehouse.

Three *R* packages were used to further process the data in *R* – *plyr* (Wickham, 2011), *dplyr* (Wickham and Francois, 2016) and *reshape2* (Wickham, 2007). Demographic input variables are shown in Table 1.

Majority of the variables are self-explanatory, but account type and city should be more explained. There are 3 various types of accounts – type A corresponds to personal account, type B to account for students and type D to family account. The variable city is used to distinguish behavior of customers in big and small cities. The five biggest cities were selected based on discussion with company representatives. The living standard of these five cities is higher than standard in the smaller ones.

Table 1 Demographic variables

Variable name	Description
birthYear	Customer's year of birth
delinquent	Did the customer have problem with paying bills at least once in last year?
duration	Customer lifetime
accType	Type of account
contractDuration	Days till the end of contract
city	Indication, whether a customer is from the 5 biggest cities in the country

Source: Company database

The input variables in the service usage group are visible in Table 2. It is important to note that all averages are monthly averages.

The dependent binary variable is called *portout* and tells whether the given customer left the company in 45 days after the date when the input variables were calculated.

Table 2 Service usage variables

Variable name	Description
avgInvoice	Average amount of invoice for last year
avgExtra3	Average overpayment for last 3 months
avgExtra6	Average overpayment for last 6 months
avgExtra12	Average overpayment for last 12 months
avgData	Average monthly data consumed (GB's)
avgVoice	Average monthly voice consumed (min's)
avgSMS	Average monthly number of sent SMS
avgMMS	Average monthly number of sent MMS
VAS	Whether the customer has currently value added services
portout	Whether the customer left the company

Source: Company database

You can see descriptive statistics (minimum, first quartile, median, mean, third quartile and maximum) of numeric variables in Table 3. There are clearly visible some outliers, e.g. maximal values for avgSms and avgMms. Outliers were removed and replaced by medians. The negative values of variable contractDuration means that a given customer had specific time limited (e.g. 1 year) contract which ended in past. Now, the customer does not have time limited contract and pays invoice based on actual usage of services.

Table 3 Descriptive statistics of numeric variables

Variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
birthYear	1 910	1 959	1 968	1 968	1 980	2003
contractDuration	–6 890	–2 583	–1 274	–1 695	–366	4 866
avgInvoice	0.16	331.69	431.37	567.95	674.85	746.64
avgExtra12	0.00	9.35	46.31	92.19	25.16	6 133.47
avgExtra6	0.00	7.00	48.44	100.91	136.43	6 533.47
avgExtra3	0.00	2.00	42.18	101.82	136.16	9 666.17
avgData	0.00	366.70	2 505.60	10 746.80	8 968	60 602.80
avgVoice	0.00	0.00	0.00	49.15	64.00	2 723.72
avgSms	0.00	0.00	0.00	2.85	0.00	594.55
avgMms	0.00	0.00	0.00	0.11	0.00	680
duration	15	1 335	2 196	2 639	3 833	7 070

Source: Company database

Counts of individual levels for categorical variables are visible in Table 4. The majority of customers have account type A and have no problems with paying bills. Roughly three fifths of customers have value added services and only one fifth of customers is from the 5 biggest cities.

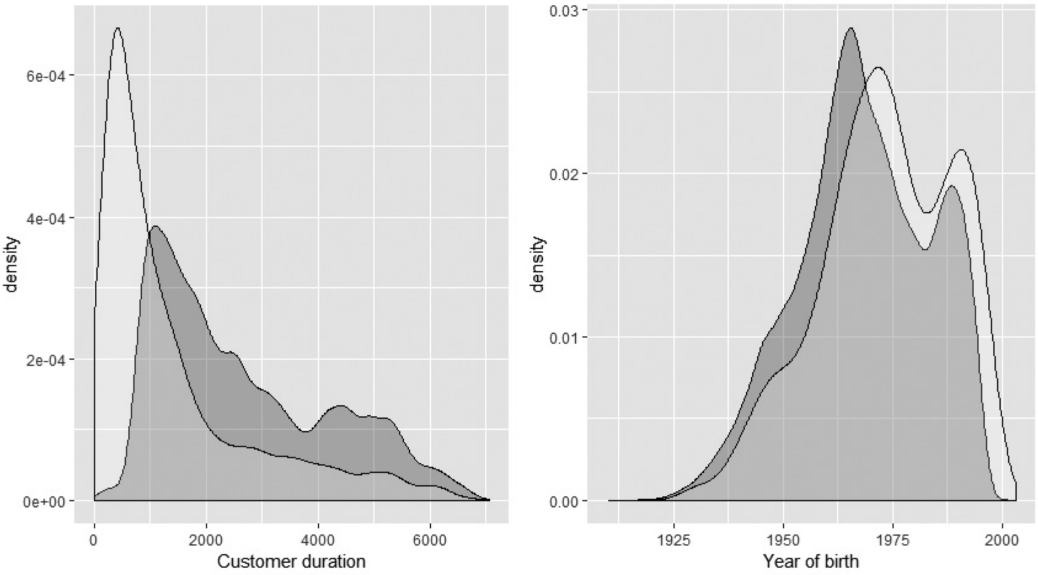
Table 4 Frequencies of levels of categorical variables

accType		delinquent		vas		city		portout	
A	49 571	*yes*	49 948	*yes*	31 623	*yes*	10 241	*yes*	49 225
B	373	*no*	112	*no*	18 437	*no*	39 819	*no*	835
D	116								

Source: Company database

Two interesting patterns appeared during exploratory data analysis. Overlapping density plots were used for comparison of differences of numeric variables *duration* and *birthYear* grouped by binary dependent variable *portout* (see Figure 1). The density plots were created using the kernel smoothing method with Gaussian kernel. The lighter gray color of area under density curve represents customers, who left company, the darker gray area under density curve represents customers, who stayed. It is evident that churners are younger and are with company shorter time.

Figure 1 Overlapping density plots for duration and birth year



Source: Own construction

3.2 Description of the estimated model

Estimation results of logistic regression model, which consists of estimated regression coefficient, odds ratio, standard error, value of Wald *z* statistic and corresponding *p-value* are shown in Table 5. According to the regression coefficients, higher values of these numeric variables *increases probability to churn*: birth year, average data and average sms. Also delinquent customers have 2.61 times higher odds to churn than non-delinquent ones and customers with students account type (B) have 2.53 times higher odds to churn than customers with regular personal account. We can infer that younger customers, who use mobile data and sms, with occasional problem with paying their bills and with students account have generally higher tendency to leave the company.

On the other hand, higher values of these numeric variables *decreases probability to churn*: average invoice, days till the end of contract, duration, average overpayment, average voice and average mms. Company should focus retention activities mainly on customers with ending contract in the near future and are shorter time with the company.

Customers with family account in comparison to customers with personal account and customers which have value added services in comparison to those without value added services have minimal odds to churn (in fact, in the training data set there were not a single customer with value added services or with account type D, which left company). Customers from the 5 biggest cities have 0.96 times lower odds to churn than customers from other cities and are therefore less likely to churn.

The behavior of customers is most probably different for various levels of their age and this fact could influence estimated regression coefficients. Interaction terms between birth year of customers and average voice, data and duration were, therefore, embedded into logistic regression model to monitor the effect of year of birth to regression coefficients of these variables. These variables were selected because it is expected that younger customers use mobile data more and are shorter time with company whereas older customers use voice calls more and are longer time with company.

Table 5 Estimation results

Variable	Estimate	Odds	Std. Error	Z value	p-value
(Intercept)	–33.760	0.000	8.567	–3.941	0.000
birthYear	0.017	1.017	0.004	3.839	0.000
contractDuration	–0.004	0.996	0.000	–29.371	< 2e-16
delinquent true	0.961	2.614	0.262	3.665	0.000
accType B	0.926	2.525	0.488	1.899	0.058
accType D	–15.308	0.000	1 628	–0.009	0.992
avgInvoice	–2.64e-05	1.000	1.32e-04	–0.200	0.841
avgExtra12	–1.18e-04	1.000	1.96e-04	0.601	0.548
avgData	3.05e-04	1.000	1.46e-04	2.093	0.036
avgVoice	–0.084	0.919	0.053	–1.597	0.110
avgSms	0.004	1.004	0.002	1.775	0.076
avgMms	–0.017	0.983	0.005	–0.335	0.737
vas true	–18.983	0.000	185.3	–0.102	0.918
city yes	–0.045	0.956	0.103	–0.439	0.661
duration	-0.013	0.987	0.004	3.008	0.003
avgVoice·birthYear	4.22e-05	1.000	2.68e-05	1.576	0.115
avgData·birthYear	–1.55e-07	1.000	7.42e-08	–2.090	0.037
duration·birthYear	–6.86e-06	1.000	2.22e-06	–3.092	0.002

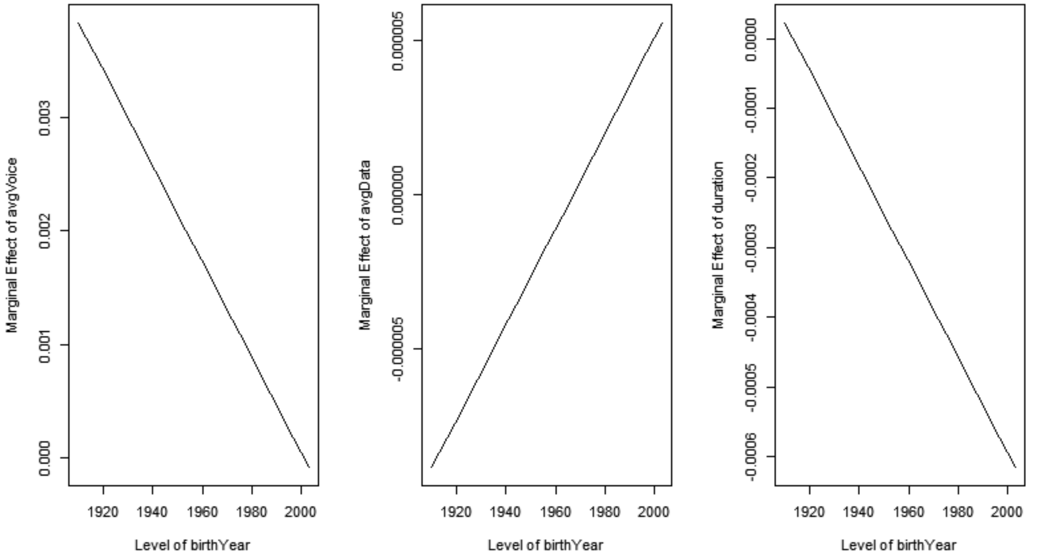
Source: Own construction

The effect of avgVoice, avgData and duration on the probability to churn is not the same depending on the values of birthYear. We can see values of birth year on the x axis and marginal effect of corresponding variable on y axis in Figure 2. The lines in three charts show how the effect of a given variable changes depending on the value of birth year. For example the marginal effects for average voice are calculated as

$$ME_{avgVoice} = \widehat{\beta}_{avgVoice} + \widehat{\beta}_{avgVoice \cdot birthYear} \cdot birthYear. \tag{7}$$

From Figure 2 it is clear that the effect of avgVoice and also duration is higher for older customers while the effect of avgData is higher for younger customers.

Figure 2 Effect of interaction terms



Source: Own construction

The *Hosmer-Lemeshow* test was used as the first one to assess the fit of the model. Hosmer and Lemeshow (2000) recommends to set the parameter g (number of subgroups) as $k + 1$ (number of variables + 1), this parameter was therefore set to 18. The test statistics follows χ^2 distribution with $g - 2 = 16$ degrees of freedom. The *p-value* of test statistic is equal to 2.2e – 16, which is below alpha = 0.05, so the null hypothesis that the observed and expected proportions are the same across all subgroups is rejected. This negative result of the test can be caused by the large data set or by presence of nonlinearities in the model.

Another possibility to assess the fit of the model is by using R^2 measures. R squared measures should be used only as an additional tool for assessing model fit (Hosmer and Lemeshow, 2000). These measures are suitable for comparison of competing models fit to the same set of data. The relatively low values of pseudo R^2 measures (see Table 6.) comparing to the R^2 values of good linear models are the norm and should not be understood as a signal of bad model.

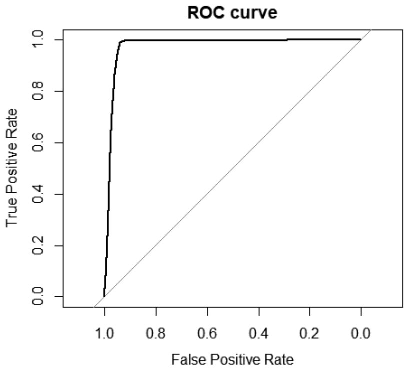
Table 6 Pseudo R squared metrics

Pseudo R squared	Value
McFadden	0.478
Nagelkerke (Cragg and Uhler)	0.499

Source: Own construction

In the next part the quality of the model is assessed using test data set of 50 060 customers which wasn't used in the process of model training. Complex assessment of the ability of classification model to distinguish between classes is possible with the use of *ROC curve* and corresponding predictive measure AUC. From Figure 3 it is evident that the *ROC curve* is close to the perfect classifier, which has a curve that passes through the point at 100 percent true positive rate and 0 percent false positive rate. With the AUC value equal to 0.976 the logistic regression classifier is according to the Lantz (2013) outstanding.

Figure 3 ROC curve



Source: Own construction

To compute the values of *sensitivity* and *precision*, it is necessary to determine the probability cutoff (threshold). Hosmer and Lemeshow (2000) recommends choose the threshold in the intersection of sensitivity and specificity. In this case the threshold was set to value 0.065 – customers with predicted probability to churn lower than or equal to 0.065 are predicted to churn and customer with probability higher than 0.065 is predicted to stay.

Table 7 shows the resulting confusion matrix computed for the probability threshold 0.065.

Table 7 Confusion matrix for the threshold 0.065

Prediction / Reference	stay	leave
stay	46 669	43
leave	2 556	792

Source: Own construction

Information in confusion matrix is used to calculate sensitivity and precision. Sensitivity is computed using this formula:

$$Sensitivity = \frac{792}{792 + 43} = 0.948. \qquad (8)$$

The value 0.948 tells us that the logistic regression model is able to catch 94.8% of customers, who in reality left company. Another important performance measure is *precision*. *Precision* is calculated as follows:

$$Precision = \frac{792}{2\,556 + 792} = 0.237. \qquad (9)$$

The value of *precision* 0.237 means that 23.7% of customers predicted to leave in fact left the company. This ratio is not so big in comparison with the sensitivity, but is acceptable because of the higher importance of *sensitivity* for the company.

It was proved that logistic regression is a useful tool applicable to the area of prediction of customer churn. There are mainly two advantages of using logistic regression – the first one is its interpretability, which is understood as an ability to reveal important factors, their strength and direction. The second one is the predictive power of the algorithm tested and confirmed on the independent testing data set.

CONCLUSIONS

The aim of this paper was to use logistic regression for understanding and prediction of customers who are about to leave Telecommunications provider, which resides in the European Union. Demographic variables such as year of birth, customer lifetime or account type as well as service usage variables such as average monthly voice, data, overpayment or invoice were extracted from the company's data warehouse. Exploratory data analysis revealed that younger customers and customers with shorter lifetime tend to churn more. The estimated logistic regression model confirmed this finding – younger customers who are shorter time with a company, who use mobile data and sms more than traditional calls, with occasional problem with paying bills, with students account and ending contract in near future are typical representatives of customers who tend to leave the company. The interaction terms between year of birth and variables avgVoice, avgData and duration show that the marginal effects of avgVoice and duration are higher for older customers, while the marginal effect of avgData is higher for younger customers. It seems natural that younger customers use their smartphones mainly with connection to the internet whereas older customers use mobile mainly for calling. The quality of the model was confirmed by for logistic regression relatively high value of McFadden and Nagelkerke R-squared measures equal to 0.478 and 0.499, respectively. Three other metrics for assessing quality of the model on independent test data set were calculated – AUC, sensitivity and precision. High values of AUC (0.9759) and sensitivity (0.948) validated that the estimated model is able to predict customers intending to churn. According to the value of sensitivity, the logistic regression model was able to successfully predict 94.8% of customers who in fact actually left the company. To sum up, logistic regression was successfully applied in Telecommunication Company for detection of customers tending to leave the company and also for discovery of the most important drivers of churn.

ACKNOWLEDGEMENTS

This research was supported by the European Social Fund within the project CZ.1.07/2.3.00/20.0296, the Student Grant Competition of Faculty of Economics, VŠB-Technical University of Ostrava within project SP2016/116.

References

AHN, J. H., HAN, S. P., LEE, Y. S. Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications Policy*, 2006, 30(10–11), pp. 552–568.

BALASUBRAMANIAN, M. AND SELVARANI, M. Churn Prediction in Mobile Telecom System Using Data Mining Techniques. *International Journal of Scientific and Research Publications*, 2014, 4(4), pp. 1–5.

CANALE, A. AND LUNARDON, N. *Churn Prediction in Telecommunications Industry.* A Study Based on Bagging Classifiers, Carlo Alberto Notebooks, 350 p.

COX, D. R. AND SNELL, E. J. *Analysis of binary data.* New York: Chapman and Hall, 1989.

DAHIYA, K. AND TALWAR, K. Customer churn prediction in telecommunication industries using data mining techniques – a review. *International journal of advanced research in computer science and software engineering*, 2015, 5(4), pp. 417–433.

GÜRSOY, S. Customer churn analysis in telecommunication sector. *İstanbul Üniversitesi İşletme Fakültesi Dergisi*, 2010, 39(1), pp. 35–49.

HOSMER, D. W. AND LEMESHOW, S. *Applied logistic regression.* New York: Wiley, 2000.

HEERINGA, S., WEST, B., BERGLUND, P. *Applied survey data analysis.* Boca Raton, FL: Chapman & Hall/CRC, 2010.

JAMES, G. R. *An introduction to statistical learning: with applications in R.* New York: Springer, 2013.

LANTZ, B. *Machine learning with R.* Birmingham: Packt Publishing, 2013.

LAZAROV, V. AND CAPOTA, M. *Churn Prediction, Business Analytics Course.* TUM Computer Science, 2007.

MCFADDEN, D. Conditional logit analysis of qualitative choice behavior. In: ZAREMBKA, P. eds. *Frontiers in Econometrics.* New York: Academic Press, 1974.

NAGELKERKE, N. J. D. A note on a general definition of the coefficient of determination. *Biometrika*, 1991, 78, pp. 691–692.

OGHOJAFOR, B. et al. Discriminant Analysis of Factors Affecting Telecoms Customer Churn. *International Journal of Business Administration*, 2012, 3(2), pp. 59–67.

OLLE OLLE, G. AND CAI, S. A Hybrid Churn Prediction Model in Mobile Telecommunication Industry. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 2014, 4(1), p. 55–62.

SEBASTIAN, H. AND WAGH, W. Churn Analysis in Telecommunication using Logistic Regression. *Oriental Journal of Computer Science & Technology*, 2017, 10(1), pp. 207–212.

VAN DEN POEL, D. AND LARIVIERE, B. Customer Attrition Analysis for Financial Services Using Proportional Hazard Models. *European Journal of Operational Research*, 2004, 157(1), pp. 196–217.

WICKHAM, H. AND FRANCOIS, R. *dplyr: A Grammar of Data Manipulation. R package version 0.5.0* [online]. 2016. <https://CRAN.R-project.org/package=dplyr>.

WICKHAM, H. Reshaping Data with the reshape package. *Journal of Statistical Software*, 2007, 21(12), pp. 1–20.

ZUMEL, N. AND MOUNT, J. *Practical data science with R.* Shelter Island, NY: Manning Publications Co., 2014.

Application of Hierarchical Cluster Analysis in Educational Research: Distinguishing between Transmissive and Constructivist Oriented Mathematics Teachers

Janka Medová[1] | *Constantine the Philosopher University in Nitra, Nitra, Slovakia*
Jana Bakusová | *Constantine the Philosopher University in Nitra, Nitra, Slovakia*

Abstract

Special questionnaire containing two sets of items related to teachers' beliefs and current pedagogies was completed by 30 mathematics in-service teachers from 26 different lower secondary schools in Slovakia. Their responses were analysed by means of hierarchical cluster analysis. The clustering method was chosen by the correlation coefficient between Euclidean and dendrogram-predicted distances. Cluster analysis grouped the items of this questionnaire into four clusters, describing the following aspects: (1) discipline and classroom culture; (2) pedagogies and problem solving; (3) applications of mathematics and students' activity during the lesson; and (4) teachers' attitudes towards students' individuality and mathematics. Teachers were grouped into two clusters. Based on the differences between responses we consider one cluster of teachers as transmissive-oriented and the second one as constructivist-oriented.

Keywords	***JEL code***
Hierarchical cluster analysis, mathematics education, teachers' beliefs, instructional approaches	*C38, I29*

INTRODUCTION

One of the main aims of mathematics education is to motivate students through the properly posed questions and problems. Students gain understanding, their imageries are developed and concepts become

[1] Department of Mathematics, Faculty of Natural Sciences, Constantine the Philosopher University in Nitra, Tr. A. Hlinku 1, 949 74 Nitra, Slovakia. Corresponding author: e-mail: jmedova@ukf.sk, phone: (+421)376401691.

crystallised (Hejný et Kuřina, 2001). However, the everyday reality of schools is often very different. Very often, in advance prepared facts are routinely presented to students, and which may lead to rote learning and acquisition of merely formal knowledge of mathematical concepts (Lloyd, 2018). Traditionalist or transmissive teachers believe that they should use "didactic instructional practices, directly offering students standardized, prescriptive methods for completing mathematical task. These methods are practiced on homework tasks, memorized and applied on similar … test tasks for which correct responses indicate mathematical understanding" (Lloyd, 2018). By contrast, Beerenwinkel et von Arx (2017) have described the constructivist-oriented teacher as a person who activates students' pre-knowledge, provides them with suitable problems often related to everyday context. During the problem-solving activity of students, the teacher creates the required and necessary space for independent learning, encourages rethinking and seeks to demonstrate the scientific approach to knowledge generation which fosters critical thinking (Grofčíková et al., 2018).

Teachers have to face a lot of issues and challenges when implementing constructivist approaches. Especially, those teachers who lack any experience in the constructivist classroom work often struggle with setting such a learning environment that is required for the constructivist perspective (Windschitl, 2002).

Even though earlier research has shown that teachers' perception of mathematics does not have significant impact on teachers' work, their perceptions of mathematics teaching form teachers' practice (Andrews et Hatch, 1998).

The objective of this study is to propose a mean that could help classify teachers as preferring either transmissive or constructivist pedagogies. Previously, teachers' beliefs had been investigated via questionnaires (e.g., Andrews et Hatch, 1998) and through the observation of their practice (e.g., Beerenwinkel et von Arx, 2017). Moreover, Lloyd (2018) performed a typological analysis of pre-service mathematics teachers based on analysis of both qualitative data, including video responses and lesson plans, and quantitative data from the Likert-scale questionnaires.

In order to divide teachers into several groups according to their beliefs about teaching and self-reporting their teaching practices the hierarchical cluster analysis (Tan et al., 2019) was performed. Hierarchical cluster analysis had previously been used for grouping teachers according their practice based on analysis of video-recordings of their lessons (Beerenwinkel et von Arx, 2017), for grouping students according their responses to the Likert-scale items (da Recha Seixas et al., 2016), or by their learning styles (Liu et al., 2017). It was also used for grouping items of questionnaires according to similar responses (Hörstermann et Krolak-Schwerdt, 2012).

1 METODOLOGY OF RESEARCH

In order to distinguish between the constructivist and transmissive teachers a questionnaire was designed, taking into account their beliefs and practice into account was designed. This questionnaire comprises of specific items focused on the professional practice of teachers, their professional development, the description of current classroom practice, and their beliefs. The questionnaire had been applied within a larger study, investigating the relation between the teachers' pedagogies and mathematical knowledge for teaching combinatorics.

1.1 Teachers' beliefs and current pedagogies

The items focused on the teachers' beliefs and description of current classroom practices were taken from the questionnaire described in more details in (Engeln et al., 2013; Engeln, 2013), designed within the project PRIMAS (PRIMAS, 2018) where the first author of the here-presented study was involved as well.

The first eight questions were aimed at identification of the teacher alone. The battery 9 contained such items that related to the students' role, including social interaction (9a, 9f and 9o) and use of problems (9e, 9g and 9s). The battery 10 comprised of the items related to teacher's role, with the special focus on applications of mathematics (10a, 10e and 10h) and usual pedagogies (10b. 10f, 10i, 10l and 10n). Item 11 was an open question, asking about the problem the teacher had to deal in mathematics instruction. Remaining three batteries were aimed at the use of inquiry-based learning in participants' teaching.

The aim of the PRIMAS questionnaire was to measure teachers' tendency to use inquiry-based pedagogies in their classroom. As the PRIMAS questionnaire was intended for both mathematics and science teachers, the items related to science education were omitted. When applicable, the word *subject* (meaning math, or science in the original version) was for our purposes substituted by the word *mathematics*. The first set of the adjusted questionnaire (battery 9 of the PRIMAS questionnaire) consisted of 18 statements about students' activity during mathematics lessons, and the second one (battery 10) comprised of 14 statements about the teacher. The teachers were asked to assign the frequency at which the described situations occurred in their classroom to the values on Likert scale from 1 (never or hardly ever) to 4 (almost in each lesson). Henceforth we will address the items by their numbers in the PRIMAS questionnaire.

1.2 Participants

The designed questionnaire was distributed to lower secondary mathematics teachers via online portal for teachers. Altogether 30 teachers from 27 schools in Slovakia completed the questionnaire. The sample contains teachers from all regions in Slovakia. As for gender, there were only two men in the sample. Twenty-one teachers were from mixed-ability schools. Among the rest of the teachers there were teachers from schools for gifted children (2), schools with special programme in mathematics (1), sports (2), music (1) or languages (2), and a special school for students with hearing impairment (1). The sample cannot be considered as representative, so we cannot generalize the conclusions from the concluded analysis.

1.3 Statistical methods

Hierarchical cluster analysis is usually employed in order to divide data into meaningful and/or useful groups (or clusters). It helps to provide understanding of the structure and further classification of involved elements (e.g., participants, samples, items) (Tan et al., 2019).

The inputs to the analysis are usually n objects with d descriptors for each further called x_i, $i \in \{1,2, \dots, d\}$. Altogether they form an $n \times d$ matrix with entries m_{ij}. The descriptors can be continuous, discrete or binary. The data should be standardized with the aim to diminish influences due to measure units of different descriptors (Kráľ et al., 2009). In our case, an integer value between one and four was assigned to each questionnaire item by respondents, thus the standardisation was not needed.

As the first step of the analysis, a distance matrix $\mathbf{D}_{ij} = D(x_i, x_j)$ is constructed. Euclidean distance defined as $D(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (m_{ik} - m_{jk})^2}$ is often used. The minimal value $D(C_i, C_j) = \min_{1 \le k,\, l \le n,\, k \ne l} D(C_k, C_l)$ is determined and the clusters C_i and C_j are merged into a new cluster C_{ij}. The row j is omitted and values of distance matrix $\mathbf{D}$ are recalculated by the formula:

$$D(C_k, C_{ij}) = \alpha_i\, D(C_k, C_i) + \alpha_j\, D(C_k, C_j) + \beta D(C_i, C_j) + \gamma \left| D(C_k, C_i) - D(C_k, C_i) \right|, \tag{1}$$

where the way to calculate coefficients α, β, γ determines the clustering method (Tan et al., 2019). The most frequently used methods are summarized in Table 1. The clusters with minimal distance are merged until there is only one cluster left. Then, the dendrogram T is constructed and analysed.

The relation between the Euclidean distance and dendrogram-predicted distance can be expressed by the cophenetic correlation coefficient:

$$c = \frac{\sum_{i<j} (D(x_i, x_j) - \bar{x})(t(x_i, x_j) - \bar{t})}{\sqrt{[\sum_{i<j} (D(x_i, x_j) - \bar{x})^2][\sum_{i<j} (t(x_i, x_j) - \bar{t})^2]}}, \tag{2}$$

Table 1 Table of coefficients for common hierarchical clustering approaches

Clustering method	α_i	α_j	β	γ
Nearest neighbour (single linkage)	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Unweighted pair group method using arithmetic mean	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	0	0
Weighted pair group method using arithmetic mean	$\frac{1}{2}$	$\frac{1}{2}$	0	0
Centroid	$\frac{n_i}{n_i + n_j}$	$\frac{n_j}{n_i + n_j}$	$-\frac{n_i n_j}{(n_i + n_j)^2}$	0
Ward's	$\frac{n_i + n_k}{n_i + n_j + n_k}$	$\frac{n_j + n_k}{n_i + n_j + n_k}$	$-\frac{n_k}{n_i + n_j + n_k}$	0

Note: n_i is the size of cluster C_i.
Source: Adapted from Tan et al. (2019)

where $t(x_i, x_j)$ is the distance between the objects x_i and x_j in dendrogram. The higher the correlation is, the better the model fits the data.

Cluster analysis of the responses to items in two sets from the PRIMAS questionnaire (Engeln, 2013) was performed, using the Euclidean distance and the five clustering methods in R environment (RCoreTeam, 2018). In order to evaluate the most appropriate clustering method for each direction of analysis, the correlation coefficients between Euclidean and dendrogram-predicted distances were calculated. The final clustering was visualized by a heatmap and dendrograms (Figure 1) using the gplots (Warnes et al., 2016) library. The medians of responses for the items were compared by the Moods' median test using the RVAideMemoire (Hervé, 2018).

2 RESULTS AND DISCUSSION

The five clustering methods listed in Table 1 lead to five dendrograms describing the similarity of responses to the questionnaire items. Next, the responding teachers were taken as objects and items as descriptors, and another five dendrograms were constructed. Correlation coefficients between the predicted and Euclidean distances were calculated for the dendrograms (Table 2). The most appropriate method for each direction of analysis is highlighted. UPGMA was found to be the most appropriate clustering method for items and Ward's method for clustering the responding teachers.

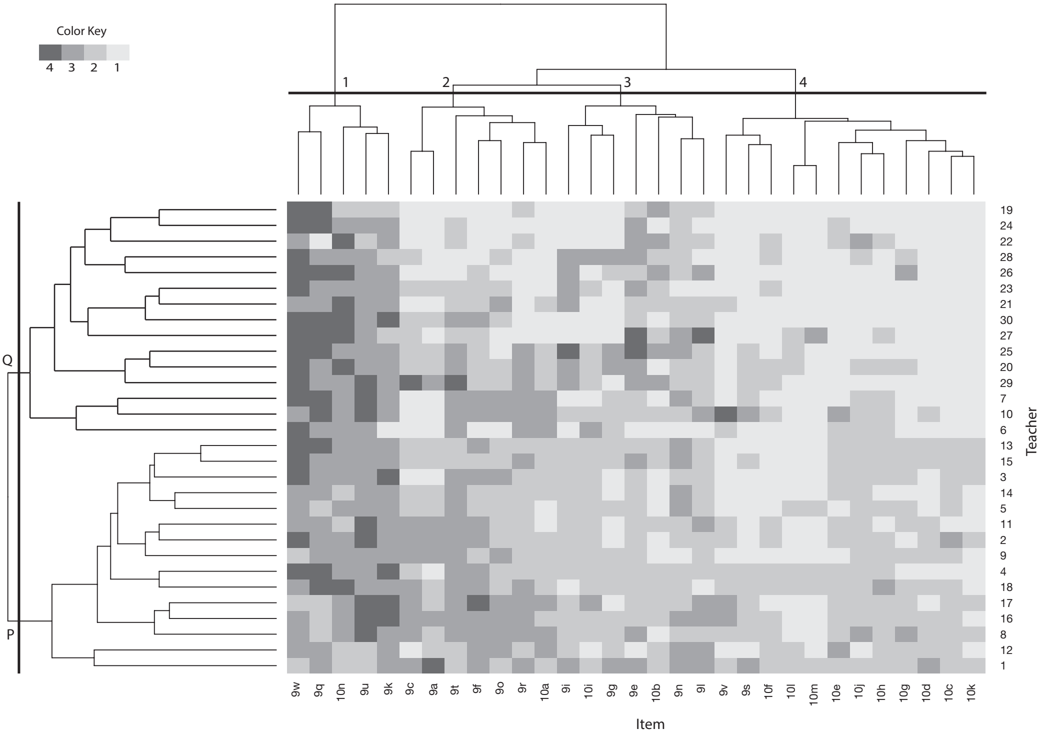
Table 2 Correlation coefficients for dendrograms constructed by different methods

Clustering method	Items	Teachers
Nearest neighbour (single linkage)	0.7085525	0.5842636
Unweighted Pair Group Method using Arithmetic mean (UPGMA)	**0.8613783***	0.6834309
Weighted Pair Group Method using Arithmetic mean (WPGMA)	0.8591654	0.5376688
Centroid	0.8258559	0.6662426
Ward's	0.8347634	**0.7624185***

Note: The highest correlation coefficient is asterisked.
Source: Authors' construction

Hierarchical cluster analysis (Figure 1) grouped all respondents into two clusters, henceforth labelled as cluster P and cluster Q. The questionnaire items were grouped into four clusters, henceforth labelled from cluster 1 to cluster 4. Further, we describe the clusters in more details. The medians and quartite spans for responses to item according the clustering are in the Annex (Table A1).

Figure 1 Heatmap of the teachers' responses to items concerning their beliefs and current classroom practices



Source: Authors' construction

Five items were grouped in cluster *1 Discipline and classroom culture*. All of them are related to discipline (9q, 9w) and classroom culture (9k, 9u, 10n). The cluster P teachers agreed significantly more with statement 9w (The students take long to settle down after the lesson begins, $p = 0.026$). Based on their responses we consider the cluster P teachers as more strict and requiring higher discipline in classroom. Čeretková et Janečková (2015) list the classroom culture supporting students' discussion as one of the five main characteristics of inquiry-based class-room, which is one of the approaches based on constructivism.

Cluster *2 Pedagogies and problem solving* consisted of the items related to teachers' activity during the lessons. Teachers in cluster P agreed more with items 9e (The students repeatedly practice the same method on many questions) and 9i (The students listen to what I say) what is typical for transmissive teachers (Wood, 1995). The teachers in cluster Q responded more affirmatively to item 10i (I summarise content and results, $p = 0.013$). This difference may stem from the different approaches to problem handling. Students taught by the cluster P teachers work out more on routine problems. The higher agreement with statement 10i might indicate more problem-solving activity (Schoenfeld, 1992) in practice of the cluster Q teachers.

There were nine items grouped in cluster *3 Applications of mathematics and students' activity during the lesson*. The items were related either to students' activity during the lessons (SA) in terms of teacher- or student-centred beliefs about teaching according to Murphy et al. (2004), or to applications of mathematics (Ap). Teachers in the cluster Q showed higher agreement with the items 9o (The students are involved in class debate or discussion, $p = 0.009$) and 9t (The students have an influence on what is done in the lesson, $p = 0.024$), what indicates that these teachers tend to offer more room for students' discussion. Windschitl (2002) put a special emphasis on providing students with opportunities for discussion in a constructivist classroom. Higher agreement with statements 9r (The students work on problems that are related to their real life experience, $p < 0.001$) and 10a (I use this subject to help the students understand the world outside school, $p = 0.004$) indicates the teachers' willingness to involve real-life problems in the classroom. Teachers in this cluster manifested their student-centred beliefs about mathematics teaching. Independent learning of students is considered as typical for constructivist teachers (Beerenwinkel et von Arx, 2017).

The last cluster, cluster *4 Teachers' attitudes towards students' individuality and mathematics*, put together two different issues. The items grouped here are reflecting the extent of teachers' individual approach to students (IS) and the teachers' attitude towards mathematics and its teaching (M). Teachers in the cluster Q demonstrated higher agreement with the item 10j (I help students with their learning, $p < 0.001$). Significant differences were confirmed in the items 10e (I show how mathematics is relevant to society, $p = 0.001$) and 10h (I explain the relevance of mathematics to students' daily lives, $p = 0.001$) with that the teachers in the cluster Q agreed more what confirmed their belief that mathematics should be related to everyday life of the students.

Based on the description of clusters and differences between teachers in the clusters P and Q we can see that the cluster P teachers lecture and explain, and, based on teachers' instruction, students are working on routine problems. We can conclude that the teachers in the cluster P tend to do transmissive instructions. On the other hand, according to teachers' responses, students in the cluster Q classrooms discuss and influence the lesson by their activity. Teachers reported themselves as persons summarising results rather than lecturing. The belief that students are able not only to use mathematical concepts, but also to discover and inquire into them, and to relate obtained information with previous knowledge, is usual for constructivist teachers (Krpec, 2015). They tend to involve their students by providing them with opportunities to discuss specific topics related to mathematics and, hence, influencing the lesson. These teachers also incline to pose problems from every-day life of their students. Based on our questionnaire we cannot declare whether they use real problems or pseudo-real tasks. The items related to every-day use of mathematics were present in the two clusters. Cluster 3 contained items related to the choice of problems. Some teachers have never encountered context-based mathematics tasks in their own education (Plothová et al., 2017), so they may differ in attitudes towards mathematics and do not see it useful for society or students' daily lives, as grouped in cluster 4. The teachers in cluster Q see themselves as being the facilitators of the educational process. Their teaching is in good fit with the inquiry-based pedagogies, as described by several scholars (Čeretková et Janečková, 2015; Engeln et al., 2013; Samková et al., 2015), therefore, we consider such teachers as constructivist-oriented.

CONCLUSIONS

This paper tried to shed more light on differences in pedagogies and beliefs of lower-secondary mathematics teachers. As early as 1984, Gonzales Thompson showed that teachers' beliefs play significant role in their instructional practices. Her research was qualitative, comprising several case studies. Subsequently, in the following decades the idea was further elaborated (e.g. Lloyd, 2018). We drew on work of Engeln et al. (2013) within the PRIMAS project who constructed a questionnaire for measuring the teachers' tendency to use inquiry-based learning and therefore the constructivist teaching in their practice.

Having performed the hierarchical cluster analysis while processing the results of questionnaire survey conducted among 30 lower-secondary mathematics teachers, two groups of teachers were identified. Four sets of statements were constructed when the same analysis was performed on transposed data. These four clusters allowed us to study the differences between the two clusters of teachers. We found that the main difference is in students' activity in the classroom. In classrooms taught by the cluster Q teachers there is a space for individual activity of students and strong focus on applications of mathematics.

The students in classrooms of the cluster P teachers more often repeatedly practice the same method on different questions, whereas students in the Q classrooms solve preferably real-life problems and discuss different topics related to mathematical concepts. Based on the literature in the field of mathematics education, we can refer to the cluster P teachers as transmissive-oriented, and the cluster Q teacher as constructivist-oriented.

We have shown that the hierarchical cluster analysis may be used as a reasonable tool for grouping teachers with certain characteristics obtained as an agreement with thoroughly chosen statements. The findings of the study cannot be generalized as the sample was not fully representative and comprised only 30 teachers. If more participants filled the questionnaire, the number of clusters could increase. The new cluster could obtain the more extreme (more transmissive or more constructivist) teachers, or some kind of balanced approach using both transmissive and constructivist pedagogies.

ACKNOWLEDGMENT

Authors would like to thank to two anonymous reviewers whose comments significantly increased the quality of the paper, and to Anna Tirpakova for her kind and fruitful advice.

This work was supported by the Slovak Research and Development Agency under the contract No. APVV-15-0368.

References

BEERENWINKEL, A. AND VON ARX, M. Constructivism in Practice: An Exploratory Study of Teaching Patterns and Student Motivation in Physics Classrooms in Finland, Germany and Switzerland. *Research in Science Education*, 2017, 47, pp. 237–255.

ČERETKOVÁ, S. AND JANEČKOVÁ, M. *Objavné vyučovanie matematiky (Inquiry-Based Learning in Mathematics).* 1st Ed. Nitra: Constantine the Philosopher University, 2015.

DA ROCHA SEIXAS, L., GOMES, A. S., DE MELO FILHO, I. J. Effectiveness of Gamification in the Engagement of Students. *Computers in Human Behavior*, 2016, 58, pp. 48–63.

ENGELN, K. *IBL Implementation Survey Report.* Freigurg: PH Freiburg, 2013.

ENGELN, K., EULER, M., MAASS, K. Inquiry-Based Learning in Mathematics and Science: A Comparative Baseline Study of Teachers' Beliefs and Practices across 12 European Countries. *ZDM*, 2013, 45, pp. 823–836.

GONZALEZ THOMPSON, A. The Relationship of Teachers' Conceptions of Mathematics and Mathematics Teaching to Instructional Practice. *Educational Studies in Mathematics*, 1984, 15, pp. 105–127.

GROFČÍKOVÁ, S., DUCHOVIČOVÁ, J., FENYVESIOVÁ, L. Development of Future Teachers' Critical Thinking through Pedagogical Disciplines. *Slavonic Pedagogical Studies Journal: The scientific educational journal*, 2018, 7, pp. 101–109.

HEJNÝ, M. AND KUŘINA, F. *Dítě, škola a matematika: Konstruktivistické přístupy k vyučování (Child, School and Mathematics: Constructivist Approaches to Education).* 1st Ed. Prague: Portál, 2001.

HERVÉ, M. *RVAideMemoire: Testing and Plotting Procedures for Biostatistics.* R package version 0.9-70, 2018.

HÖRSTERMANN, T. AND KROLAK-SCHWERDT, S. Teachers' Typology of Student Categories: A Cluster Analytic Study. In: GAUL, W. A., GEYER-SCHULZ, A., SCHMIDT-THIEME, L., KUNZE, J. eds. *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*, pp. 547–555, Berlin, Heidelberg: Springer, 2012.

KRÁĽ, P., et al. *Viacrozmerné štatistické metódy so zameraním na riešenie problémov ekonomickej praxe (Multivariate Statistical Methods Focusing on Solving Problems of Economic Practice).* Banská Bystrica: Univerzita Mateja Bela v Banskej Bystrici, 2009.

KRPEC, R. *Konstruktivistický přístup k výuce kombinatoriky (Constructivist Approach to Combinatorics Education).* Ostrava: Ostravská univerzita v Ostravě, Pedagogická fakulta, 2015.

LIU, Q. et al. A Study on Grouping Strategy of Collaborative Learning Based on Clustering Algorithm In: CHEUNG, S. K. L., MA, W., LEE, LK., YANG, H. eds. *International Conference on Blended Learning*, Cham: Springer International Publishing, 2017, pp. 284–294.

LLOYD, M. E. R. A Typological Analysis: Understanding Pre-Service Teacher Beliefs and How They Are Transformed. *International Journal of Mathematical Education in Science and Technology*, 2018, 49, pp. 355–383.

MURPHY, P. K., DELLI, L. A. M., EDWARDS, M. N. The Good Teacher and Good Teaching: Comparing Beliefs of Second-Grade Students, Preservice Teachers, and Inservice Teachers. *The Journal of Experimental Education*, 2004, 72, pp. 69–92.

OECD. *PISA 2006 Technical Report.* Paris: PISA, OECD Publishing, 2009.

PLOTHOVÁ, L. et al. An Analysis of Students' Use of Mathematical Models in Solving Tasks with Real-life Context. In: *APLIMAT 2017: Proceedings from 16th Conference on Applied Mathematics*, Bratislava, Slovakia, 31 January–2 February 2017, Bratislava: STU, 2017, pp. 1207–1223.

PRIMAS [online]. 2018. <www.primas-project.eu>.

RCORETEAM. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2018.

SAMKOVÁ, L., HOŠPESOVÁ, A., ROUBÍČEK, F., TICHÁ, M. Badatelsky orientované vyučování matematice (Inquiry-Based Learning in Mathematics). *Scientia in educatione*, 2015, 6, pp. 91–122.

SCHOENFELD, A. H. Learning to Think Mathematically: Problem Solving, Metacognition, and Sense Making in Mathematics In: GROUWS, D. eds. *Handbook of Research on Mathematics Teaching and Learning*, New York: Macmillan, 1992.

TAN, P.-N., STEINBACH, M., KARPATNE, A., KUMAR, V. Cluster Analysis: Additional Issues and Algorithms In: TAN, P.-N., STEINBACH, M., KARPATNE, A., KUMAR, V. eds. I*ntroduction to Data Mining*, 2nd Ed. India: Pearson Education, 2019.

WARNES, G. R. et al. *Gplots: Various R Programming Tools for Plotting Data.* R Package Version 3.0.1., 2016.

WINDSCHITL, M. Framing Constructivism in Practice as the Negotiation of Dilemmas: An Analysis of the Conceptual, Pedagogical, Cultural, and Political Challenges Facing Teachers. *Review of Educational Research*, 2002, 72, pp. 131–175.

WOOD, T. An Emerging Practice of Teaching In: *The Emergence of Mathematical Meaning*, Abingdon Oxford: Routledge, 1995, pp. 211–235.

ANNEX

Table A1 Medians of responses to items according to clustering

Cluster	Item	Statement	Cluster *P* (*n*=18)			Cluster *Q* (*n*=12)		
			Me(*P*)	Quartile span		*Me*(*Q*)	Quartile span	
1 Discipline and classroom culture	9k	The students have the possibility to decide how things are done during the lesson.	2	2	2	2	2	2.25
	9q	The students behave noisily and course disorder.	2	1.25	2	1.5	1	2
	9u	The students choose which questions to do or which ideas to discuss.	2	1	2	2	2	2
	9w*	The students take long to settle down after the lesson begins.	2	1	2	1	1	1
	10n	I give a lecture.	2	2	2	1.5	1	2
2 Pedagogies and problem solving	9e	The students repeatedly practice the same method on many questions.	3	3	3	2	2	3
	9g	The students learn through doing exercises.	3	3	3.75	4	2.75	4
	9i	The students listen to what I say.	3	3	3	2	2	4
	10b	I give my students precise instructions.	3	3	3	3	2	3.25
	10i*	I summarise content and results.	3	3	3	4	3	4

Table A1 (continuation)

Cluster		Item	Statement	Cluster *P* (*n*=18)			Cluster *Q* (*n*=12)		
				***Me*(*P*)**	**Quartile span**		***Me*(*Q*)**	**Quartile span**	
3 Applications of mathematics and students’activity during the lesson	SA	9a	The students are given opportunities to explain their ideas.	3	2.25	3.75	4	3	4
		9cP	The students have the possibility to try out their own ideas.	3	2	3.75	3	3	4
		9fP	The students have discussions about the topics.	2	2	2.75	3	3	4
		9l	The students have no problems to follow the lesson.	3	2	3	3	3	3.25
		9n	The students know enough to understand the lessons.	3	2	3	3	3	3
		9o*	The students are involved in class debate or discussion.	3	2	3	3.5	3	4
		9t*	The students have an influence on what is done in the lesson.	2	2	2	3	2	3.25
	Ap	9r*	The students work on problems that are related to their real life experience.	2.5	2	3	4	2.75	4
		10a*P	I use this subject to help the students understand the world outside school.	3	2	3	4	3	4
4 Teachers’ attitudes towards students’individuality and mathematics	IS	9s	The students start with easy questions and move on to harder questions.	3.5	3	4	4	3.75	4
		9v	The students are informed about the aim of the lesson.	3	3	4	4	4	4
		10d	I show interest in every student’s learning.	3	3	4	4	4	4
		10f	I give students extra help, if they need it.	3.5	3	4	4	3	4
		10g	I continue teaching until the students understand.	3	3	4	4	3.75	4
		10j*	I help students with their learning.	3	3	3	4	3.75	4
		10l	I outline the most important points of a lesson.	4	3.25	4	4	3.75	4
	M	10cS	I enjoy teaching mathematics.	3	3	3.75	4	4	4
		10e*S	I show how mathematics is relevant to society.	3	3	3	4	4	4
		10h*S	I explain the relevance of mathematics to students’ daily lives.	3	3	3	4	3.75	4
		10kS	I really like mathematics.	4	3	4	4	4	4
		10mS	I treat mathematics as important.	4	4	4	4	4	4

Note: SA – students’ activity, Ap – applications of mathematics, IS – individual approach to students; M – teachers’ attitudes towards mathematics and its teaching. In items marked with the letter S the original wording *the/this subject* was substituted by the word *mathematics*. Items marked with the letter P were used in PISA 2006 (OECD, 2009).

Source: Authors’ construction

Searching for Correlations Between CO_2 Emissions and Selected Economic Parameters

Ladislav Rozenský[1] | *Czech University of Life Sciences Prague, Prague, Czech Republic*
Pavla Vrabcová[2] | *Czech University of Life Sciences Prague, Prague, Czech Republic*
Miroslav Hájek[3] | *Czech University of Life Sciences Prague, Prague, Czech Republic*
Tereza Veselá[4] | *Czech University of Life Sciences Prague, Prague, Czech Republic*
Petr Hukal[5] | *Czech University of Life Sciences Prague, Prague, Czech Republic*

Abstract

The Emission Trading Scheme is one of the economic instruments of environmental policy and it is used to achieve the goals of reducing greenhouse gas emissions. The Emission Trading Scheme is a common instrument of the European Union, which is mandatory for all member countries. The aim of this paper is to assess the effectiveness of the greenhouse gas emissions trading system in the Czech Republic as one of the important instruments of environmental policy. The presented research model shows that greenhouse gas emissions were only minimally affected by the GDP index level and movement in the monitored time period. The model also shows that the most significant impact on the amount of greenhouse gas production in the given time period was the consumption of renewable energy and the consumption of fossil fuels. By contrast, the price of emission allowances on the market had a minimal effect on the production of greenhouse gases.

Keywords

Economic instruments, emission trading scheme, emission allowances, renewable energy

JEL code

F64, H23, K32, O44

[1] Department of Forestry and Wood Economics, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Kamýcká 129, 165 21, Prague 6 – Suchdol, Czech Republic. E-mail: rladislav@seznam.cz.

[2] Department of Wood Processing, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Kamýcká 129, 165 21, Prague 6 – Suchdol, Czech Republic: E-mail: vrabcovap@fld.czu.cz.

[3] Department of Forestry Technologies and Construction, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Kamýcká 129, 165 21, Prague 6 – Suchdol, Czech Republic. E-mail: hajek@fld.czu.cz.

[4] Department of Forestry Technologies and Construction, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Kamýcká 129, 165 21, Prague 6 – Suchdol, Czech Republic. E-mail: sramkova.tereza@gmail.com.

[5] Department of Forestry and Wood Economics, Faculty of Forestry and Wood Sciences, Czech University of Life Sciences Prague, Kamýcká 129, 165 21, Prague 6 – Suchdol, Czech Republic. E-mail: petr@rozvoj.net.

INTRODUCTION

Greenhouse gases arise as a result of human activities (Montzka et al., 2011; Prather et al., 2012), especially due to the burning of fossil fuels (Mittal et Kumar, 2014). To reduce greenhouse gas emissions, environmental policy uses various instruments (OECD, 2007). Economic instruments work in synergy (Phalan et al., 2016) with the other greenhouse gas reduction instruments and relevant legislation (Sorrell, 2003; Directive 2003/87/EC). The individual instruments are chosen by member countries in line with the priorities of their environmental policies (Jordan et al., 2003; Aidt et al., 2004). One of basic instruments in this mix is represented by so-called environmental taxes (Baranzini et al., 2000; OECD, 2007). In the European area, carbon Emission Trading Scheme (ETS) has become a cornerstone of the design of European environmental policy (Convery et al., 2007; Braun, 2009; Hong et al., 2017; Segura et al., 2018). The purpose of ETS is above all to reduce greenhouse gas emissions and to integrate the costs for the elimination of their negative environmental impacts into the costs of their producers (Hong et al., 2017; Segura et al., 2018). Sabzevar et al. (2017) stated that the emission trading price is a base for maximizing company's profits and for controlling the amount of emissions generated.

The European Carbon Emissions Trading Scheme (EU ETS) introduced in 2005 has led to both spot and futures market trading of carbon emissions. However, despite 10 years of trading, we have no knowledge on how profitable the ETS is (Narayan et al., 2015). At present, European countries use emission allowances within the EU ETS as the main CO_2 abatement instrument, but some of them also use carbon taxes, as for example Sweden, Finland, Ireland or Denmark (Leu et Betz, 2016).

Many studies (for example Laing et al., 2014; Hintermann et al., 2015; Muûls et al., 2016) analyzed emission reductions, evolution of allowance prices, and impacts on the economic performance, competitiveness and innovation. Nordhaus (2005, 2011) focused his research mainly on comparing the effectiveness of environmental taxes and ETS, their advantages and disadvantages. Based on this research, he strongly opts for environmental taxes before the ETS and concludes that the carbon market price fluctuation and volatility in one trading period of the EU ETS are not good in terms of longer-term investment planning. As a recommendation for policy makers and regulators, a carbon tax is proposed in the context of fiscal policy as the most appropriate tool for reducing greenhouse gas emissions.

The main ETS supporters for example Fan (2017), who states that in relation to tax, the ETS can increase prosperity on the market with imperfect competition. Moreover, based on his model of strategic and competitive behavior of traders on the Central Atlantic Market, who covers power sector emissions in nine Northeast and Mid-Atlantis States, he notes that when regulators charge tax instead of allowance, the loss caused by deadweight costs in imperfect competition is higher. The Central Atlantic Market is the second largest emission allowance market in the world, which, like the EU ETS, operates on the principle of a free market for emission allowances. Very important in the ETS is the price of tradable permits (Hintermann et al. (2015) find that economic activity and growth announcements as well as oil and gas prices positively influence the prices of allowances. The emissions allowance price has varied considerably over years (Segura et al., 2018). Brink et al. (2016) points to the current market price of emission allowances, which is quite far from the projected price of EUR 20, which is planned as the target price in 2020, and does not therefore seem to meet the desired effect. Deeney et al. (2016) deal with the influence of the European Parliament on the auction price of emission allowances. In April 2013, the European Parliament was expected to pass a draft law for fixing the recognized oversupply issue in the EU ETS (Koch et al., 2014). The Commission's proposal involved postponing until 2019–2020 the release of 900 million EU emission allowances (EUAs).

Oueslati et al. (2017) deal with the specifics of emission allowances and carbon tax.They say energy taxes represent important instruments to increase economic efficiency, to achieve desired environmental outcomes, and raise public revenues. However, the implementation of energy taxes is often hampered by public concerns over their possible effects on unequal income.

Cost-saving from the use of renewable energy sources, which are also represented in our model, is dealt with by Palmer et Burtaw (2005). A renewable energy production tax credit reduces the electricity price at the expense of taxpayers, which limits its effectiveness in reducing carbon emissions, and is less cost-effective in increasing renewables than a portfolio standard. Neither policy is as cost-effective as a cap-and-trade policy for achieving carbon emission reductions (Palmer et Burtraw, 2005).

The effectiveness of environmental policy instruments is often assessed for example through the relationship between the economic growth and the growth of greenhouse gas production. There are many publications dealing with the environmental Kuznets curve (Dinda, 2004; Alam et al., 2016; Özokcu et Özdemir, 2017). Bauer et al. (2015) deal with the impact of fossil fuels policy on reducing greenhouse gas emissions in the United States climatic region. Lim et al. (2014) discussed the context of fossil fuels consumption, economic growth and greenhouse gas production in the Philippines.

The aim of this paper is to assess ETS effectiveness in the Czech Republic using the regression analysis as one of the important instruments of environmental policy.

1 MATERIAL AND METHODS

This chapter introduces research questions and goals and brings details on the individual data that is further worked on in the research section of the paper. Theories and data (CZSO, 2019; ERU, 2019; EUROSTAT, 2019) needed are included in the descriptive part, the objective of which is to assess the environmental effectiveness of emission allowances. For this purpose, the data was created to form time series, from which the charts were then compiled for a more comprehensive understanding of the problem, and the regression analysis was performed that tried to assess the remaining objectives. Another aim was to assess some further factors and how they affect greenhouse gas emissions in the countries under review.

The ETS effects on the amount of CO_2 production were analyzed in details. Because the used model affects some other factors and tools, such as the using of renewable energy sources, or the consumption of fossil fuels, the authors used the regression analysis. By our opinion a suitable research method to assess the synergistic effect of several factors on the research goals.

Greenhouse gas emissions (expressed in tons per year of CO_2 per capita) are the basic dependent variable. The data source is data from the European Statistical Office (Eurostat, 2016). The emission allowance price is an explanatory variable. The emission allowance price was chosen as a variable because the EU ETS is a fundamental obligatory regulatory element. The data was obtained from the European Energy Exchange and from the Energy Regulatory Office. Unit is the average annual emission allowance in EUR per 1 allowance.

Gross Domestic Product (GDP) expressed as a percentage of year-on-year growth for the Czech Republic and a year was obtained from the database of the Czech Statistical Office (CZSO).

The consumption of fossil fuels an explanatory variable which was chosen because the consumption of fossil fuels relates to the greenhouse gas emissions. The data was obtained from the Eurostat database and was calculated per capita and year in tones for the Czech population as of 31 December of the respective year, according to the CZSO database. The consumption of renewable energy is a control explanatory variable in our model. The production of energy from the renewable sources does not generate greenhouse gases. Their substitution for energy from the combustion of fossil fuels containing carbon has the ultimate effect of reducing the greenhouse gas production. The consumption of renewable resources is supported by the country´s policy. This policy is a common EU policy. It is based in particular on state support for the use of renewable energy (Kharlamova et al., 2018). The data was obtained from the Eurostat database and was calculated per capita and year for the population of the Czech Republic as of 31 December of the respective year according to the CZSO database in our calculation model. It is reported by converting the consumption of renewable energies into tons of oil equivalent. In the case of this variable, the theoretical expectations were negative, i.e. with the increasing consumption of renewable energies; there is a decline in greenhouse gas production.

For the data analysis in years 2005–2015 (greenhouse gas emissions, GDP in the Czech Republic in %, emission allowance price, consumption of fossil fuels and consumption of renewable energy sources), regression and correlation analysis was used. The regression analysis allows getting information about the dependence of quantitative characters (Litschmannová, 2011). The *Y* variable, whose behaviour we try to explain, is called a dependent variable (the variable explained). The *X* variable, whose behaviour explains the behaviour of the dependent *Y* variable, is called an independent variable (Hindls et al., 2002). The correlation analysis deals with interdependencies, emphasizing the strength or intensity of the relationship (Bílková et al., 2009). In most cases, the linear regression used was $\eta = \beta_0 + \beta_1 x$. We measured the intensity of dependence using the determination index (Budíková et al., 2010). If the dependency function is proven, the determination index is 1 (and vice versa if the value is 0). The Pearson correlation coefficient for the two variables *X* and *Y* was calculated. Some indicators were also analysed by using the elementary statistical analysis with selected characteristics of position, variability and concentration (median, variance, standard deviation, kurtosis and skewness). Following hypotheses are verified in this paper:

- H_0: there is no linear relationship between the *X* (GDP growth rates in %) and *Y* (GHG emission levels) variables;
- H_0: there is no linear relationship between the *X* (emission allowance price) and *Y* (level of greenhouse gas emissions) variables;
- H_0: there is no linear relationship between the *X* (consumption of fossil fuels) and *Y* (GHG emission level) variables;
- H_0: there is no linear relationship between the *X* (consumption of RES) and *Y* (level of greenhouse gas emissions) variables.

For testing the hypotheses, fixed probability of error of the first type (so-called materiality level) was chosen to be 5%. Tests of the significance of regression parameters were performed to determine if the correlation between the sample variables is strong enough to be considered as proven for the base set.

2 RESULTS

Table 1 shows the basic characteristics of the average price of emission allowance: median, variance, standard deviation, kurtosis and skewness.

Table 1 Elementary statistical analysis – average price of emission allowance in 2005–2015

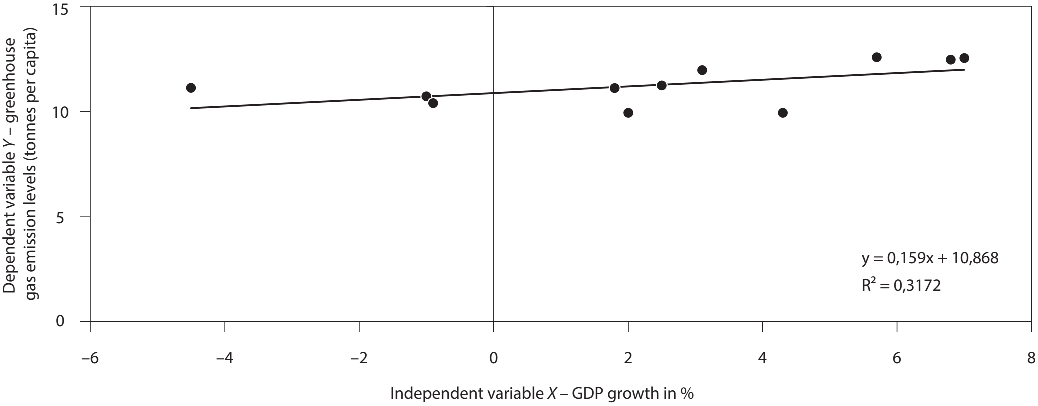
Median	12.720
Standard deviation	6.420
Variance	41.217
Kurtosis	–0.988
Skewness	0.109

Source: Own elaboration

As seen from Table 1, the average price of emission allowance shows considerable variability, which is mainly interpreted by standard deviation (= 6.42). The median, which divides the series of upwardly ranked results into two halves, came out as 12.72. Kurtosis, as a measure of the concentration of the values of random variable around the mean value, is about –0.99, which means the flatter distribution and fewer spikes than the normal distribution. Skewness, which is a measure of symmetry of the given probability distribution, came out as 0.11, indicating a positively skewed division where the values are concentrated rather to the left.

Figure 1 expresses the relationship between the independent variables, namely GDP growth rate (in %) and dependent X variable, which expresses the level of emissions in 2005–2015, including the estimation of the least squares line parameters. Expectations were that the GDP growth would increase the greenhouse gas emissions. This was the construction of a so-called empirical curve describing the observed correlation at the sample level. This curve serves as an estimate of the actual dependence (linear regression function) assumed for the entire base file.

Figure 1 Relationship between the GDP growth rates and greenhouse gas emission levels



Source: Own elaboration

In Figure 1, a theoretical line, i.e. a line drawn by the point graph, was drawn as close to all points as possible – it is the closest regression function with the equation 10.868 + 0.159x. This linear regression function was used to describe the true dependence of the monitored variables at the level of the entire base file. The regression function contained only one regressor, so we tested the zero $H_0 : \beta_1 = 0$ hypothesis against the alternative $H_1 : \beta_1 \neq 0$ hypothesis.The F-test result can be seen in Table 2.

Table 2 Overall *F*-Test design (independent variable: GDP growth rate, dependent variable: GHG emissions)

Source of variability	Sum of squares	Degrees of freedom	Dispersion (mean sum of squares)	*xOBS*	*p*-value
Model	3.188	1	3.188	4.181	0.071
Residual	6.861	6	0.762	×	×
Total	10.049	7	×	×	×

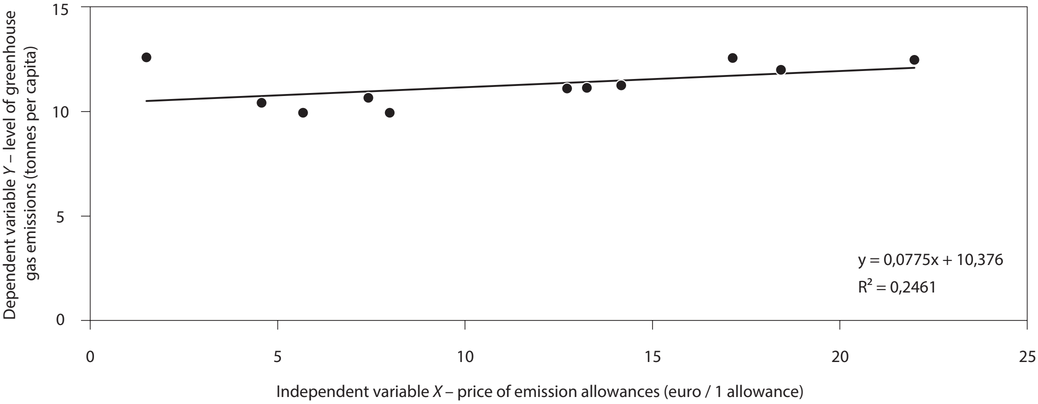
Source: Own elaboration

At the significance level of 0.05 we did not reject the zero hypothesis. Therefore, the GHG emission levels could not be estimated by using the GDP growth rate. The determination index, resp. the modified determination index, determines the quality of the model. The determination index was 0.317;

the modified determination index was 0.241. The model explains for more than 24% of the total variance of the dependent variable, so it cannot be labelled too high.

Furthermore, the relationship between the price of emission allowances and the amount of emissions produced in tonnes per capita in the Czech Republic was examined, see Figure 2.

Figure 2 Relationship between the price of emission allowances and the level of greenhouse gas emissions



Source: Own elaboration

A regression function was found with the equation 10.376 + 0.078*x*. We tested the zero $H_0 : \beta_1 = 0$ hypothesis against the alternative $H_1 : \beta_1 \neq 0$ hypothesis. The *F*-test result can be seen in Table 3.

Table 3 Construction of the total *F*-test (independent variable: emission allowance price, dependent variable: GHG emissions)

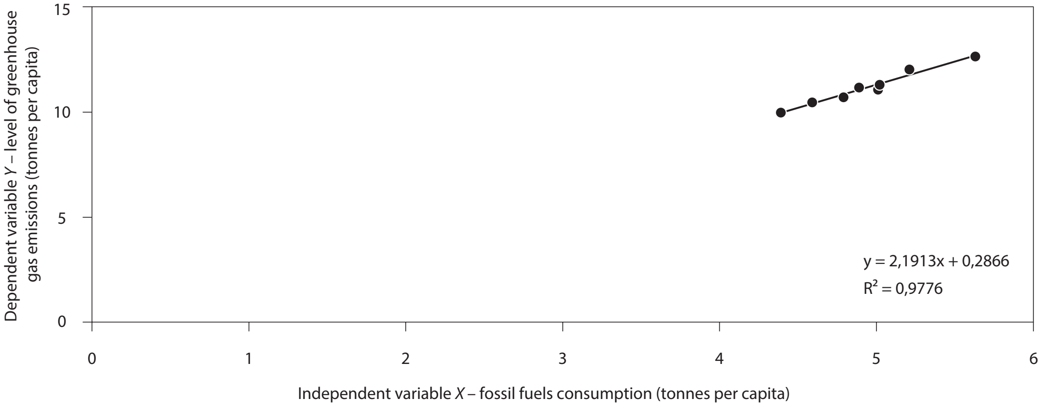
Source of variability	Sum of squares	Degrees of freedom	Dispersion (mean sum of squares)	*xOBS*	*p*-value
Model	2.473	1	2.473	2.938	0.121
Residual	7.575	9	0.842	×	×
Total	10.048	10	×	×	×

Source: Own elaboration

At the significance level of 0.05, the zero hypothesis could not be rejected, the chosen model was not statistically significant. Therefore, the level of significance of the greenhouse gas emission level could not be estimated by using the emission allowance prices. At the significance level of 0.05, we accepted the zero hypothesis, the $\beta 1$ parameter was not statistically significant. The determination index, resp. the modified determination index, determines the quality of the model. The determination index was 0.246; the modified determination index was 0.162. Therefore, the model explains for more than 16% of the total variance of the dependent variable, so it cannot be considered as good.

The relationship between the consumption of fossil fuels in the Czech Republic (tonnes per capita) and greenhouse gas emissions in years 2007–2015 is shown in Figure 3.

Figure 3 Relationship between the fossil fuels consumption and GHG emissions in 2007–2015



Source: Own elaboration

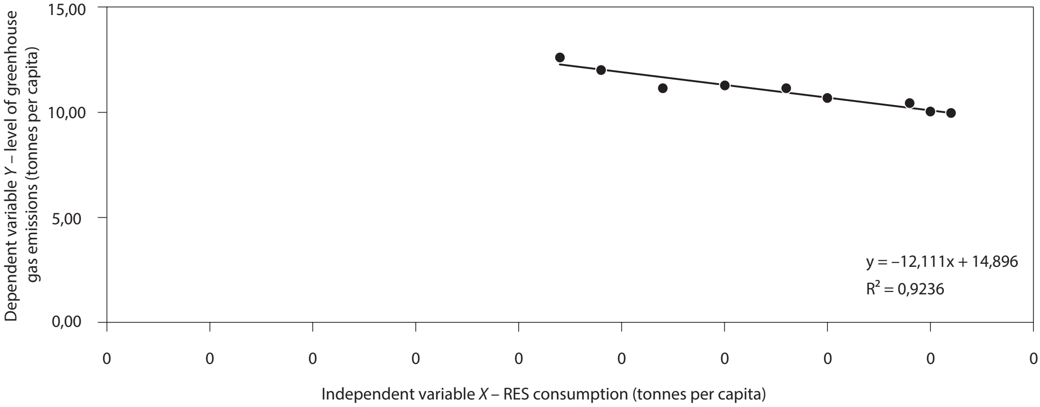
A regression function was found with the equation 0.287 + 2.191x. We tested the zero $H_0 : \beta_1 = 0$ hypothesis against the alternative $H_1 : \beta_1 \neq 0$ hypothesis. The F-test result can be seen in Table 4.

Table 4 Overall *F*-test design (independent variable: fossil fuel consumption, dependent variable: GHG emissions)

Source of variability	Sum of squares	Degrees of freedom	Dispersion (mean sum of squares)	*xOBS*	*p*-value
Model	6.148	1	6.148	305.745	4.93E–7
Residual	0.141	7	0.020	×	×
Total	6.289	8	×	×	×

Source: Own elaboration

Figure 4 Relationship between the RES consumption in the Czech Republic and GHG emissions in 2007–2015



Source: Own elaboration

At the significance level of 0.05 we rejected the zero hypothesis; the chosen model was statistically significant. Hence, greenhouse gas emission levels could be estimated using the fossil fuel consumption. The model explains for more than 97% of the total variance of the dependent variable, so it can be labelled as very good. Pearson's correlation coefficient was 0.989 (indicating a very strong positive correlation). In this case, it can be noted that the increase in the consumption of fossil fuels increases the level of greenhouse gas emissions. The last dependence between the variables is dealt with in Figure 4.

A regression function was found with the equation 14.896 – 12.111x. We tested the $H_0 : \beta_1 = 0$ zero hypothesis against the alternative $H_1 : \beta_1 \neq 0$ hypothesis.The F-test result can be seen in Table 5.

Table 5 Overall *F*-test design (independent variable: RES consumption, dependent variable: GHG emissions)

Source of variability	Sum of squares	Degrees of freedom	Dispersion (mean sum of squares)	*xOBS*	*p*-value
Model	5.808	1	5.808	84.652	3.69E-5
Residual	0.480	7	0.069	×	×
Total	6.288	8	×	×	×

Source: Own elaboration

At the significance level of 0.05 we rejected the zero hypothesis; the chosen model was statistically significant. GHG emission levels could therefore be estimated by using the renewable energy sources. The determination index was 0.924; the modified determination index was 0.913. The model explains for more than 91% of the total dispersion of the dependent variable, so it can be labelled as very good. Pearson's correlation coefficient was –0.961, which indicates a very strong negative correlation, so it can be assumed that greenhouse gas emission levels will decrease with the increasing RES consumption.

In addition to the above analyses, the dependence was examined of final energy consumption in 2010–2015 and emissions of basic pollutants (SO_x in particular) into the air in the Czech Republic (see Table 6).

Table 6 Total *F*-Test design (independent variable: final energy consumption, dependent variable: SO_x)

Source of variability	Sum of squares	Degrees of freedom	Dispersion (mean sum of squares)	*xOBS*	*p*-value
Model	379 764.076	1	379 764.076	0.336	0.593
Residual	4 525 119.965	4	1 131 279.991	×	×
Total	4 904 884.041	5	×	×	×

Source: Own elaboration

At the significance level of 0.05, we did not reject the zero hypothesis. Therefore, final energy consumption levels could not be estimated according to SO_x emissions. At the materiality level of 0.05, we did not reject the zero hypothesis, the β_1 parameter was not statistically significant. The model cannot be considered good.

3 DISCUSSION

Authors admit that relationships between all the variables observed in the article are very complex, moreover the relevant variables are far more than the article is tracked.

Long-term data from the CZSO show that, despite all measures adopted, the Czech Republic together with Estonia, the Netherlands and Luxembourg are four countries with the highest CO_2-per capita production (Eurostat, 2016). This is apparently due to more factors, e.g. by the location of the Czech Republic and by the industrial orientation of Czech economy focused on exports. The analysis shows that, despite the above-mentioned facts, the Czech Republic is able to reduce the CO_2 production in the long term. Although some authors are trying to prove the theory of GDP correlation and increased CO_2 emissions, such as Doda et al. (2013) or Cialani (2017), it seems to be somewhat simplistic. The factors that act on this phenomenon are many and interact with each other. Our model also suggested that the above correlation is minimal. The results of the regression analysis in our model indicate that the development of greenhouse gas production was linked to GDP growth only minimally in the project period. The GHG emission levels cannot be estimated by using the GDP growth rate at the significance level of 0.05. This value was insignificant in our model. This is evidenced, for example, by the fact that the Czech Republic's GDP declined from 3.1% to –4.5% of the long-term average in 2008–2009, probably due to the global economic crisis, while the CO_2 emissions decreased only from 11.96 t to 11.11 t, which is approximately only 7% per capita, and with the GDP growth of 2.9% in 2014, the drop in emissions was even 0.47 t per capita. The hypothesis that the GDP growth will increase the CO_2 emissions was not confirmed for the chosen time period. If the theory of a direct correlation between GDP and CO_2 production can be adopted, this can be besides explained by the overheated economy, depression, economic crisis and by the subsequent slow recession that took place in the period under review and consequently affected the production output and greenhouse gas emissions. The analysis showed that the EU ETS was insignificant in the Czech Republic during the monitored period. In our model, almost no correlation was found between the price of the emission allowance and the amount of CO_2 production. Our model suggests that at the stated emission market prices, it is unlikely that the correlation between the price of emission allowances and the amount of CO_2 production is likely to be demonstrated. This is apparently due to the large number of emission allowances allocated to the issuers, due to the economic crisis and the consequent low demand for emission allowances on the market, which apparently caused the low acquisition cost of this regulatory measure. According to Eurostat, there was no year-on-year increase of greenhouse gases in the EU member states, only Germany and the Czech Republic recorded a long-term decrease (Eurostat, 2016). Among other things, this was probably due to a small, renewed increase in the price of emission allowances. The increase in emissions from the other member states indicates that this price is probably unnecessarily low and does not fulfill its regulatory role properly. Czech Republic appears to be in line with the theory outlined above, still reserving greenhouse gas emission reductions. Thus we agree with Brink et al. (2016) that the projected price of 20 EUR per emission allowance for 2020 is far away from the real price and that this current price probably does not perform well. The analysis thus confirmed that the amount of fossil fuels consumption strongly affects the CO_2 emissions. Greenhouse gas emission levels can be estimated using the fossil fuel consumption – at the significance level of 0.05 we rejected the zero hypothesis, the chosen model was statistically significant. This is due to the carbon content in these fuels and their subsequent release during their combustion. It has been confirmed that with the growing consumption of renewable energy, the CO_2 production is decreasing. Therefore, the GHG emission levels can be estimated by using the renewable energy sources. As stated in the theoretical part, this is due to the substitution effect when high carbon fuels are replaced by so-called pure energy (Boyle, 2004). Statistics from the Eurostat show that the share of renewable energies in the amount consumed is increasing every year in all EU member states. The highest increase in RES consumption is attributable to wood chips. This increase is most striking in the Nordic countries (Eurostat, 2016).

A similar study was conducted by Lin et Li (2011).The authors reviewed the regression analysis of the time series of GDP development, emission allowance prices, carbon taxes, fossil fuel consumption and renewable energy sources for 5 EU member states with the established carbon tax. Their model confirmed a direct link between the GDP and the greenhouse gas production. Unlike the emission allowances, this model showed a significant impact of carbon tax, especially in Finland. It also confirmed the importance of the consumption of fossil fuels and renewable energies in the CO_2 production. The low significance of emission allowances at their low market prices was confirmed by the two studies, whose results are similar as in the Czech Republic. Lin et Li (2011) indicated a more significant role of the carbon tax and its higher environmental effectiveness. The difference in the importance of GDP growth for CO_2 production in both studies may also be due to the study period of Lin et Li (2011) when the economic recession occurred while the results of our model could be distorted by the depression and by the subsequent economic crisis.

CONCLUSIONS

Long-term scientific studies in the EU show that the ETS introduction has led from the start to a drop in CO_2 emissions from large polluters. The data from EUROSTAT, which indicate a too slow decline in CO_2 production, suggests a need of regulatory intervention by EU institutions to ensure that greenhouse gas emissions are further reduced. One possible regulatory intervention is for example a support increase in the price of emission allowances and an immediate reduction in the amount of allowances allocated free of charge in the energy sector. Another option is the introduction of carbon tax as a further tool to reduce greenhouse gas emissions across the EU. The functionality of this tool mix is proven, for example, by the experiences of the Nordic countries since the 1990s.

However, the future development in the EU territory can be predicted as the increasing importance and involvement of environmental taxes in the mix of instruments designed to reduce greenhouse gas emissions, particularly so-called carbon taxes. The Czech Republic has a study on the introduction of carbon tax, including an impact study (Ministry of Finance of the Czech Republic, 2016). However, the introduction of any tax is a very socially sensitive issue, which requires a broad social consensus, and, in the case of environmental taxes, a certain level of environmental awareness of the population. Our research responded to all research questions we asked. It is very likely that the final results could have been affected by the economic crisis occurring at all stages in our time period. This is why it would be appropriate to continue and monitor further time periods in this research project. The basic objective of the paper was to assess whether the ETS emission allowances in the Czech Republic are environmentally effective. The research did not show their environmental effectiveness during the studied period. A proposed solution could be the introduction of carbon tax, as an additional economic tool for reducing greenhouse gases and the intervention by EU authorities aimed at the emission allowance price increased to the planned level of 20 EUR. This conclusion is also reflected in the forthcoming change prepared by the European Commission, which plans to reduce number of freely allocated allowances in the following period (European Commission, 2017). This plan was announced already in 2017 and from the 3rd quarter of 2017 to the 5th month of 2018, there was a step-up increase in the price of emission allowances on the market. Explanation is the ongoing recession of economy and the effort of companies to respond and frontload with allowances for the next period.

ACKNOWLEDGMENT

This paper was supported by IGA (Czech University of Life Sciences Prague, Faculty of Forestry and Wood Sciences), Project No. B08/17 and Project No. QK1920391 financed by Ministry of Agriculture of the Czech Republic called *Diverzifikace vlivu biohospodářství na strategické dokumenty lesnicko-dřevařského sektoru jako podklad pro státní správu a návrh strategických cílů do roku 2030* (Analysis of the consequences of greenhouse gas emissions trading in the Czech forestry).

References

Act No. 695/2004 Coll. on the conditions of trading in greenhouse gas emission allowances and on the amendment of certain laws [online]. <http://aplikace.mvcr.cz/archiv2008/sbirka/2004/zakon_12.html#castka_235>.

AIDT, T. S. AND DUTTA, J. Transitional politics: emerging incentive-based instruments in environmental regulation. *Journal of Environmental Economics and management,* 2004, 47(3), pp. 458–479. DOI:10.1016/j.jeem.2003.07.002.

ALAM, M. M., MURAD, M. W., NOMAN, A. H. M., OZTURK, I. Relationships among carbon emissions, economic growth, energy consumption and population growth: Testing Environmental Kuznets Curve hypothesis for Brazil, China, India and Indonesia. *Ecological Indicators*, 2016, 70, pp. 466–479. DOI: 10.1016/j.ecolind.2016.06.043.

BARANZINI, A., GOLDEMBERG, J., SPECK, S. A future for carbon taxes. *Ecological Economics*, 2000, 32(3), pp. 395–412. DOI: 10.1016/S0921-8009(99)00122-6.

BAUER, N. et al. CO_2 emission mitigation and fossil fuel markets: dynamic and international aspects of climate policies. *Technological Forecasting and Social Change*, 2015, 90(A), pp. 243–256. DOI: 10.1016/j.techfore.2013.09.009.

BÍLKOVÁ, D., BUDINSKÝ, P., VOHÁNKA, V. *Pravděpodobnost a statistika* (Probability and statistics). Plzeň: Vydavatelství a nakladatelství Aleš Čeněk, 2009. ISBN 978-80-7380-224-0. (in Czech)

BOYLE, G. eds. *Renewable Energy.* Oxford University Press, May 2004, 456 p. ISBN-10: 0199261784.

BRAUN, M. The evolution of emissions trading in the European Union – The role of policy networks, knowledge and policy entrepreneurs. *Accounting, Organizations and Society*, 2009, 34(3–4), pp. 469–487. DOI: 10.1016/j.aos.2008.06.002.

BRINK, C., VOLLEBERGH, H. R. J., VAN DER WERF, E. Carbon pricing in the EU: evaluation of different EU ETS reform options. *Energy Policy*, 2016, 97, pp. 603–617. DOI: 10.1016/j.enpol.2016.07.023.

BUDÍKOVÁ, M., KRÁLOVÁ, M., MAROŠ, B. *Průvodce základními statistickými metodami* (Guide to basic statistical methods). Prague: Grada, 2010. ISBN 978-80-247-3243-5. (in Czech)

CIALANI, C. CO2 emissions, GDP and trade: a panel cointegration approach [online]. *International Journal of Sustainable Development*, 2017, 24(3), pp. 193–204. DOI: 10.1080/13504509.2016.1196253.

CONVERY, F. J. AND REDMOND, L. Market and price developments in the European Union emissions trading scheme. *Review of Environmental Economics and Policy*, 2007, 1(1), pp. 88–111. DOI: 10.1093/reep/rem010.

CZSO. *Hrubý domácí produkt* (Gross Domestic Product) [online]. Prague: Czech Statistical Office, 2019. <https://www.czso.cz/csu/czso/hdp_narodni_ucty>.

DAVIET, F. AND RANGANATHAN, J. *The Greenhouse Gas Protocol: The GHG Protocol for Project Accounting.* World Business Council for Sustainable Development (WBCSD), World Resources Institute, 2005.

DEENEY, P. et al. Influences from the European Parliament on EU emissions prices. *Energy Policy*, 2016, 88, pp. 561–572. DOI: 10.1016/j.enpol.2015.06.026.

DINDA, S. Environmental Kuznets curve hypothesis: a survey. *Ecological economics*, 2004, 49(4), pp. 431–455. DOI: 10.1016/j.ecolecon.2004.02.011.

Directive 2009/.../EC of the European Parliament and of the Council of ... amending Directive 2003/87/EC in order to improve and extend the scheme for greenhouse gas emission allowance trading within the Community [online]. <http://register.consilium.europa.eu/pdf/en/08/st03/st03737.en08.pdf>.

Directive 2003/87/EC of the European Parliament and of the Councilof 13 October 2003 establishing a scheme for greenhouse gas emission allowance trading within the Community and amending Council Directive 96/61/EC. Brussels, L 275/32–46.

DODA, B. et al. Emissions-GDP Relationship in Times of Growth and Decline. *Grantham Research Institute on Climate Change and the Environment Working Paper*, 2013, 116 p.

ERU. *Průměrná cena emisní povolenky pro roky 2007–2015* (Average emission permit price for the years 2007–2015) [online]. Prague: ERU, 2019. <https://www.eru.cz/cs/teplo/sdeleni/archiv>.

EUROPEAN COMMISSION. *Emission Trading Scheme (EU ETS)* [online]. European Union: European Commission, 2017. cit. 15.5.2018. <http://ec.europa.eu/environment/climat/emission/index_en.htm>.

EUROSTAT. *Air emissions accounts by NACE Rev. 2 activity* [online]. 2019. <http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=env_ac_ainah_r2&lang=en>.

FAN, Y. What policy adjustments in the EU ETS truly affected the carbon prices? [online]. *Energy Policy*, 2017, 103, pp. 145–164. DOI: 10.1016/j.enpol.2017.01.008.

HINDLS, R. *Statistika pro economy* (Statistics for economists). 8th Ed. Prague: Professional Publishing, 2007. ISBN 978-80-86946-43-6. (in Czech)

HINDLS, R., SEGER, J., HRONOVÁ, S. *Statistika pro economy* (Statistics for economists). Brno: Professional Publishing, 2002. ISBN 80-86419-26-6. (in Czech)

HINTERMANN, B., PETERSON, S., RICKELS, W. Price and Market Behavior in Phase II of the EU ETS: A Review of the Literature. *Review of Environmental Economics and Policy*, 2015, 10(1), pp. 108–128. DOI: 10.1093/reep/rev015.

HONG, Z. et al. Optimizing an emission trading scheme for local governments: A Stackelberg game model and hybrid algorithm. *International Journal of Production Economics*, 2017, 193, pp. 172–182. DOI: 10.1016/j.ijpe.2017.07.009.

JORDAN, A., WURZEL, R. K. W, ZITO, A. R. *'New' instruments of environmental governance: Patterns and path ways of change.* London: FRANK CASS PUBLISHERS, 2003.

KHARLAMOVA, N. S. AND STAVYTSKYY, A. Estimation of renewable energy sources application in the synergy with european union policy. *Vìsnik. Kiïvs'kogo Nacìonal'nogo Unìversitetu ìmenì Tarasa Ševčenka. Ekonomìka*, 2018, Vol. 3, Iss. 198, pp. 54–65. DOI: 10.17721/1728-2667.2018/198-3/7.

KOCH, N., FUSS, S., GROSJEAN, G., EDENHOFER, O. Causes of the EU ETS price drop: Recession, CDM, renewable policies or a bit of everything? New evidence. *Energy Policy*, 2014, 73, pp. 676–685. DOI: 10.1016/j.enpol.2014.06.024.

LAING, T. et al. The effects and side-effects of the EU emissions tradings cheme. *Wiley Interdisciplinary Reviews: Climate Change*, 2014, 5(4), pp. 509–519. DOI: 10.1002/wcc.283.

LEU, T. AND BETZ, R. Environmental Tax Evaluation What can be learnt so far? *International Energy Policies & Programmes Evaluation Conference,* Amsterdam, 2016, pp. 1–24. DOI: 10.1787/888932765598.

LIM, K., LIM, S., YOO, S. Oil consumption, CO_2 emission, and economic growth: Evidence from the Philippines. *Sustainability*, 2014, 6(2), pp. 967–979. DOI: 10.3390/su6020967.

LIN, B. AND LI, X. The effect of carbon tax on per capita CO_2 emissions. *Energy policy*, 2011, 39(9), pp. 5137–5146. DOI: 10.1016/j.enpol.2011.05.050.

LITSCHMANNOVÁ, M. *Vybrané kapitoly z pravděpodobnosti* (Selected chapters of probability). Ostrava: VŠB-TU, 2011. (in Czech)

MANSUR, E. T. Prices versus quantities: environmental regulation and imperfect competition. *Journal of regulatory Economics*, 2013, 44(1), pp. 80–102. DOI: 10.1007/s11149-013-9219-6.

MINISTRY OF FINANCE OF THE CZECH REPUBLIC. *Analýza k možnostem a dopadům zohlednění environmentálních prvků v sazbách spotřebních a energetických daní v České Republice* (Analysis of the possibilities and impacts of taking into account environmental elements in consumer and energy taxation rates in the Czech Republic) [online]. Prague, 2016. [cit. 20.5.2018] <https://www.mfcr.cz/cs/legislativa/materialy-na-jednani-vlady/1-ctvrtleti-17/materialy-na-jednani-vlady-dne-9-ledna-2-27155>. (in Czech)

MITTAL, M. AND KUMAR, A. Carbon nanotube (CNT) gas sensors for emissions from fossil fuel burning. *Sensors and Actuators B: Chemical*, 2014, 203, pp. 349–362. DOI: 10.1016/j.snb.2014.05.080.

MONTZKA, S. A., DLUGOKENCKY, E. J., BUTLER, J. H. Non-CO_2 greenhouse gases and climate change. *Nature*, 2011, 476(7358), pp. 43–50. DOI: 10.1038/nature10322.

MUÛLS, M. et al. *Evaluating the EU Emissions Trading System: Take it or leave it? An assessment of the data after ten years.* Tech. Rep. 21, Grantham Institute Briefing Paper, 2016.

NARAYAN, P. K. AND SHARMA, S. S. Is carbon emissions trading profitable? *Economic Modelling*, 2015, 47, pp. 84–92. DOI: 10.1016/j.econmod.2015.01.001.

NORDHAUS, W. *Life After Kyoto: Alternative Approaches to Global Warming.* National Bureau of Economic Research, 2005, 34 p. DOI: 10.3386/w11889.

OECD. *Instrument mixes for environmental policy.* OECD Publishing, 2007, 237 p. DOI: 10.1787/9789264018419-en.

OUESLATI, W. et al. Energy taxes, reforms and income inequality: An empirical cross-country analysis. *International Economics*, 2017, 150, pp. 80–95. DOI: 10.1016/j.inteco.2017.01.002.

ÖZOKCU, S. AND ÖZDEMIR, Ö. Economic growth, energy, and environmental Kuznets curve. *Renewable and Sustainable Energy Reviews*, 2017, 72, pp. 639–647. DOI: 10.1016/j.rser.2017.01.059.

PALMER, K. AND BURTRAW, D. Cost-effectiveness of renewable electricity policies. *Energy economics*, 2005, 27(6), pp. 873–894. DOI: 10.1016/j.eneco.2005.09.007.

PHALAN, B. et al. How can higher-yield farming help to spare nature? *Science*, 2016, 351(6272), pp. 450–451. DOI: 10.1126/science.aad0055.

PRATHER, M. J., HOLMES, C. D., HSU, J. Reactive greenhouse gas scenarios: Systematic exploration of uncertainties and the role of atmospheric chemistry. *Geophysical Research Letters*, 2012, 39(9), pp. 1–5. DOI: 10.1029/2012gl051440.

SABZEVAR, N. et al. Modeling competitive firms' performance under price-sensitive demand and cap-and-trade emissions constraints. *International Journal of Production Economics*, 2017, 184, pp. 193–209. DOI: 10.1016/j.ijpe.2016.10.024.

SEGURA, S. et al. Environmental versus economic performance in the EU ETS from the point of view of policy makers: A statistical analysis based on copulas. *Journal of Cleaner Production*, 2018, 176, pp. 1111–1132. DOI: 10.1016/j.jclepro.2017.11.218.

SORRELL, S. AND SIJM, J. Carbon trading in the policy mix. *Oxford review of economic policy*, 2003, 19(3), pp. 420–437. DOI: 10.1093/oxrep/19.3.420.

The Informativeness of the Technical Conversion Factor for the Price Ratio of Processing Livestock

Kris Boudt[1] | *Solvay Business School, Vrije Universiteit Brussel, Brussels, Belgium*
Hong Anh Luu[2] | *Solvay Business School, Vrije Universiteit Brussel, Brussels, Belgium*

Abstract

The technical conversion factor (TCF) is a survey-based estimate of the percentage of carcass weight obtained per unit of live weight. Practitioners and researchers have used it to predict the corresponding price ratio (PR). We use both in-sample regressions and out-of-sample forecasting analysis to test the validity of this approach in case of predicting the price effects of processing livestock in Europe. By regressing the PR on the inverse value of the corresponding TCF for a large panel of European countries and animal types, we find a significant positive relation between these variables, which also has economic value in terms of improving out-of-sample forecasting precision. This result is shown to be robust to animal type, year, and country fixed effects. The TCF therefore has predictive value about the corresponding PR.[3]

Keywords	***JEL code***
Agriculture, carcass weight, live weight, producer price, pass-through effect	*C52, Q13*

INTRODUCTION

Agricultural production is characterized by a chain of transformations from livestock to consumer products. First, a live animal is slaughtered to get primary carcass parts such as meat, offal, and skin. Then these components are processed to obtain different products such as sausage or lard (FAO, 2011). A detailed understanding of how the processing of livestock affects agricultural prices is of paramount importance for producers and consumers of agricultural products. To achieve this, economic policymakers such as the Food and Agricultural Organization of the United Nations (FAO) use surveys to collect information on the physical efficiency of the processing of livestock and the corresponding price ratios (PRs). They use the so-called technical conversion factor (TCF) to quantify the extraction productivity. In case of livestock, the TCF indicates in percent term the dressed carcass weight that can be extracted per unit

[1] Also Ghent University, Department of Economics, Ghent, Belgium and School of Business and Economics, Vrije Universiteit Amsterdam, Amsterdam, Netherlands. E-mail: Kris.Boudt@vub.be.
[2] Corresponding author: e-mail: Hong.Anh.Luu@vub.be.
[3] This article is based on contribution at the *IAOS (OECD) conference* in September 2018 in Paris, France.

of the live weight of the slaughtered animal. For example, if there are 100 kg of carcass weight obtained from 200 kg of live weight, the TCF is 0.5.

TCFs are published by the FAO using the information obtained from surveys sent to its member nations (FAOSTAT, 2009). The main objective of TCFs is "to arrive at approximate estimates of the total availability of food in each country, expressed in terms of quantity as well as in terms of calories, protein, and fat" (FAOSTAT, 2009). In recent years, TCFs have been widely applied in research and calculations in different fields. For instance, Lazarus et al. (2014) use TCFs in calculating the carbon footprint for crops and livestock, while Luan et al. (2014) employ the conversion factor in computing land requirements for food in South Africa. OECD-FAO (2015) takes them as factors in constructing prices and quantities for a variety of agricultural products, as well as for performing the Aglink-Cosimo economic model, which analyzes supply and demand of world agriculture. Smith et al. (2016) use the conversion factors to estimate the global dietary supply of nutrients. Chaudhary et al. (2016) use TCFs to obtain the weight of primary crop required for an amount of processed food in their calculations of biodiversity loss due to anthropogenic land. Finally, the FAO has been using TCFs to impute missing observations in agricultural price series (Kirkendall, 2015; Dubey et al., 2016).

In this paper, we analyze TCFs and PRs of the processing of four common types of livestock – namely cattle, pig, chicken, and sheep – in European countries over the period 1969 to 2015. We expect to find a close connection between the TCF and the PR of meat versus live weight, since the TCF corresponds to the physical reality behind the PR of livestock products. We focus on EU member states because they share the so-called Common Agriculture Policy (CAP). Focusing on one integrated geographic region allows us to avoid inter-regional shocks and differences in policy that can affect our analysis. Another reason for choosing the EU area is the availability of high-quality price data at Eurostat and the FAO for a large number of countries and types of livestock, which facilitate our study.

Our first contribution is to provide a descriptive analysis of the relation between the TCF and the producer PRs over the period 1969 to 2015. We combine various sources to create a longitudinal database of both TCFs and producer PRs. The longitudinal time series is unbalanced, because of the many cases in which the data are missing. We analyze the data through summary statistics and a scatter plot, giving a first indication of the positive relation between the TCF and PR.

Our second contribution is to propose and test a simple model to use the TCF to predict the PRs of livestock processing for a product type and country combinations. The model consists of regressing the PR (dependent variable) on the inverse of TCF (ITCF) and three other variables, namely trend (time effect), animal type, and country. We find both in-sample and out-of-sample evidence that the ITCF-based model is useful for the prediction of the PR.

The remainder of this paper is organized as follows. In Section 1, we discuss the characteristics of TCFs and PRs of processing livestock. Section 2 describes our data, while Section 3 explains the methodology. The results are presented in Sections 4 and 5, while the last Section summarizes our main conclusions.

1 THE CHARACTERISTICS OF THE TCF AND THE PR OF PROCESSING LIVESTOCK

1.1 TCF

The TCF of livestock is a measure that indicates in percent term the amount of a product extracted per unit of the originating one. These extraction rates differ across countries and time due to differences in technology, costs, and margins. It is a key statistic determining food prices, by which the FAO keeps track of and monitors the evolution of the TCF.

The FAO's analysis starts with collecting the TCF data by sending out questionnaires to member nations (FAOSTAT, 2009). Since 1960, it has produced three publications about the TCF. The first book published in 1960 presents conversion rates of products in countries around the world. It is the foundation for the second and third publications, in 1972 and in 2009, respectively, with adjusted, extended, and refined contents to increase comprehensiveness and comparability. Although these publications include several levels of transformation of different products, here we focus on only first-level conversion factors of livestock processing.

The last three columns of Table 1 present examples of TCFs of the livestock processing in European countries (members of CAP). As can be clearly seen, extraction productivity differs among animal groups. The first-stage processing of chicken and pig is characterized by higher average TCFs (around 75–77%) compared to those of cattle and sheep, which are approximately 48–54%. Next, to perceive how TCFs evolved over 53 years, we compare the statistics in 2009 with those in 1968 and 1957. We find that for most cases, the conversion rates are stable over time.

Notable exceptions include TCFs of pig in the Czech Republic and TCFs of chicken in Italy.

1.2 The PR of processing livestock

The PR of livestock processing refers to the relation between prices received by the producers when selling live animal and meat. Investigating this PR for several countries and animal types jointly is complementary to previous studies, which have either examined the consumer PRs of agricultural commodities or focused on a single animal type. For example, Tveteras and Asche (2008) and Asche et al. (2013) analyze the fishery industry, while Chavas and Holt (1991), Parker and Shonkwiler (2013) and Holt and Craig (2006) investigate the dynamics in hog-to-feed PRs.

In this paper, we research the PRs of livestock in the first transformation level, which turns the live animal into primary components such as meat, offal, fat, and skin (FAO, 2011). Particularly, we focus on the PR of meat. This choice is due to three reasons. First, among all products derived from the carcass, meat has the most important use in human daily consumption (compared to skin, bone, or offal). Second, it accounts for a major part of the carcass. As can be seen in Table 1, between 50% and 75% of the body is meat, depending on the animal type. Third, other components such as skin, bone, offal, and fat have only a negligible economic value compared to meat.

The PR of meat is obtained by dividing the carcass meat price by the live weight price, of which they are measured at the same mass (100 kg normally). Economically, the price of carcass meat should cover all transformation costs, namely the livestock purchasing cost (i.e, the live weight price), the labor and infrastructure costs needed to slaughter the animal, and the profit margin.

1.3 The relation between PR and TCF

The extraction rate (or TCF) is expected to have an inverse relation with the meat PR. When the TCF increases, the PR decreases. To clarify this argument, we consider a stylized numeric example. Assume that for processing livestock, we have a TCF of 0.5, which means with 100 kg of live weight, we can obtain 50 kg of carcass meat. Assume further that the price of 100 kg live weight is ¤100. Then the price of 50 kg carcass meat should at least cover its material costs, which is the price of 100 kg live weight. As such, 50 kg carcass meat has the minimum price of ¤100, and when expressed for the same units of weight, the PR equals at least two. In the same manner, increases in the TCF (e.g. due to higher efficiency in the processing) can be expected to lead to decreases in the PR, and vice versa in the case of a decrease. As such, on average we expect the TCF to be inversely proportional with the PR, whereby the minimum value of the PR is the inverse of the TCF. Henceforth, we call 1/TCF the inverse technical conversion factor (ITCF) and study its relation with the producer PR of processing livestock.

2 DATA

This section includes three main parts. The first part introduces our data source, the Eurostat and the FAO, and how we collect the data. The second section examines how PRs are filtered and paired with ITCFs. After that, we provide some explorative analysis about the PRs and ITCFs.

We collect the data and calculate the PR for four types of livestock – cattle, chicken, pig, and sheep – of the CAP countries. In order to obtain the longest possible time series, the prices of live weight and carcass meat are gathered and combined from Eurostat (2017) and FAOSTAT (2017). These are the prices received by farmers for livestock primary products as collected at the point of initial sale or the first marketing stage (FAOSTAT, 2018; Eurostat, 2018). The furthest data point we can get back to is 1969, while the most recent one is 2015. After matching the live weight prices and carcass meat prices

Table 1 PRs and TCFs

No	Livestock	Country	PR						TCF		
			Period	# Obs	Min	Average	Max	SD	1957	1968	2009
1	Cattle	Belgium	1971–2001	31	1.62	1.77	2.04	0.1	0.54	0.55	0.54
2	Cattle	Denmark	1991–2015	25	1.55	1.74	1.92	0.1	0.5	0.48	0.49
3	Cattle	France	1969–2002	34	1.35	1.46	1.9	0.1	n/a	0.5	0.52
4	Cattle	Greece	1995–2014	20	1.37	1.65	1.82	0.2	n/a	0.5	0.52
5	Cattle	Italy	1969–1999	31	1.56	1.69	2.03	0.1	0.53	0.54	0.55
6	Cattle	Luxembourg	1969–2015	47	1.63	1.71	1.79	0	0.56	0.54	0.54
7	Cattle	Netherlands	1969–1990	22	1.57	1.67	1.75	0	0.52	0.52	0.54
8	Chicken	Austria	1995–2015	21	1.96	2.28	2.41	0.1	n/a	0.8	0.75
9	Chicken	Denmark	1991–2015	25	1.06	1.35	1.49	0.1	0.8	n/a	0.76
10	Chicken	Italy	1969–1999	31	1.18	1.42	1.63	0.1	0.89	n/a	0.75
11	Pig	Austria	1995–2015	21	1.22	1.23	1.26	0	0.79	0.8	0.81
12	Pig	Belgium	1970–2007	38	1.09	1.16	1.21	0	0.78	0.79	0.79
13	Pig	Czech Republic	2004–2015	12	1.27	1.3	1.33	0	n/a	0.82	0.71
14	Pig	Denmark	1973–2015	43	1.23	1.36	1.41	0	0.72	0.72	0.7
15	Pig	Greece	1981–2015	35	0.75	0.98	1.18	0.1	n/a	0.77	0.76
16	Pig	Italy	1969–1999	31	1.02	1.17	1.32	0.1	0.81	0.83	0.79
17	Pig	Luxembourg	1969–2005	37	1.16	1.22	1.31	0	0.87	0.79	0.79
18	Pig	Spain	1986–2015	30	1.42	1.45	1.52	0	n/a	0.8	0.79
19	Pig	UK	1973–2005	33	0.97	1.18	1.34	0.1	n/a	0.74	0.76
20	Sheep	Austria	1995–2015	21	1.91	2.07	2.09	0.1	0.5	0.54	0.53
21	Sheep	Greece	1995–2014	20	2.92	3.57	4.02	0.2	n/a	0.5	0.49

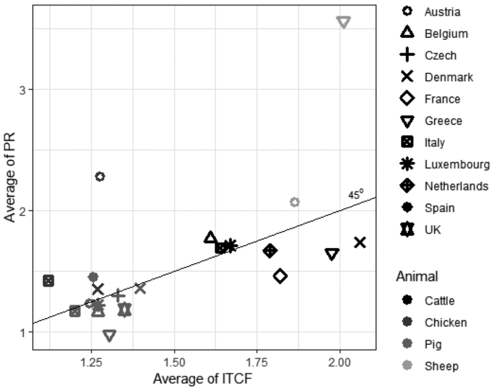
Note: "n/a" denotes that the corresponding data are not available.
Source: Eurostat (2017); FAOSTAT (1960, 1972, 2009, 2017)

and restricting the sample to the country-years that the country is effectively a member of the EU and for which there are more than ten continuous observations, we end up with a sample of 21 usable cases with 608 PR observations. There are 9 cases of pig, 7 cases of cattle, 3 cases of chicken, and 2 cases of sheep. Even though the number of chicken and sheep cases is outnumbered by cattle and pig, their observations account for more than 20% of the total database. Table 1 provides the summary statistics for this sample.

For this sample, we apply the panel unit root test (Kleiber and Lupi, 2011) to test whether the PR is stationary or not. We apply the test on the balanced panel of 21 selected cases. As can be seen in Table 1, each PR series has a different length. Therefore, we subdivide 21 cases into smaller groups based on similarity in data range and perform tests on them. For example, the subgroup with time frames 1969–2002 includes four cases: the cattle PRs of Italy (with data range from 1969 to 1999), the cattle PRs of France (1969–2002), the cattle PRs of Belgium (1971–2001), and the chicken PRs of Italy (1969–1999). A panel unit root test is performed on this subgroup. There is one exception – the pig PRs of the Czech Republic – which has a short data period of 12 years that cannot be paired with others. It is tested independently. Overall, at a 10% significance level, all tests reject the null hypothesis of a unit root and thus conclude at stationarity over the panel.

The selected PR is then paired with the corresponding ITCF, calculated by dividing one with the TCF, which is selected using the last-observation-carried-forward method. For PRs in the range from 1969 to 2008, we pair them with ITCFs obtained using information from the FAO's second publication, which was published in 1972 and provides statistics from 1968. In case these numbers are not available, we use the first publication with data from 1957. For ratios from 2009 to 2015, the ITCFs calculated from information from the third publication are used.

Figure 1 Scatter plots of the average of PRs versus the average of ITCFs



Note: This figure displays the average of PRs in relation to the corresponding average ITCFs of 21 sample cases. The black line is the 45° line.
Source: Authors´ calculations

Figure 1 displays the average of the PRs per country and animal type, in relation with the averages of corresponding ITCFs of the 21 selected cases. The 45° line indicates the reference in which the PR equals with the ITCF. Note that the observations tend to cluster for each animal type, creating disparities among groups. In particular, most chicken and pig indicators stay in lower areas compared to those of cattle and sheep. We thus find a difference in both extraction productivity and PR between animal groups. In fact, the higher TCFs of processing pigs and chickens compared to cattle explain their lower PRs (how TCFs link with PRs has been explained in Section 1.3). We also note that most observations lie along the 45° line, meaning the average PR has a close value to the ITCF. Notable exceptions are the sheep PRs of Greece with an average value of 3.57 while its ITCF is 2.92, or the chicken PRs for Austria with an average of 2.28 – substantially higher than its ITCF of 1.27.

3 METHODOLOGY

The main purpose of our research is to investigate whether the TCF is predictive of livestock price effect, or in other words, can the TCF be used to predict the corresponding livestock PR. We study this question using a panel data regression model that explains the PR using ITCFs, animal type, country, and year fixed effects. Among these variables, the ITCF is our main variable of interest. Our model is applied on a large number of countries and animal groups. It is beyond the scope of our paper to develop a model aiming at deriving causality. Instead, the model used should be interpreted as a predictive model that is useful in the operational setting of having to predict the PR using the TCF. Except for the TCF, all predictive variables used are deterministic and thus straightforward to construct. The model is estimated by ordinary least squares resulting in the best linear prediction given the set of variables used.

The empirical analysis is performed using in-sample and out-of-sample evaluation methods.

3.1 In-sample evaluation

We use nested versions of the following regression model to analyze the determinants of the PRs across observations for various animal groups (indexed by *a*), countries (indexed by *c*) and years (indexed by *t*):

$$\mathrm{PR}_{a,c,t} = \alpha + \alpha_a + \alpha_c + \gamma \mathrm{Trend}_t + \beta\, \mathrm{ITCF}_{a,c,t} + \varepsilon_{a,c,t}\,. \tag{1}$$

In Formula (1), α corresponds to the intercept of the reference category corresponding to the PR of cattle in Denmark. The terms α_a and α_c, respectively, denote the animal type and country group, while $\varepsilon_{a,c,t}$ represents the error term. The value of $ITCF_{a,c,t}$ is calculated from the TCF taken from Table 1. The deterministic trend variable $Trend_t$ takes values from 1 to 47, corresponding to the number of observations in the time series of PRs (1969–2015). In order to exploit the effect of animal group on the predicted PR, we include dummies for pig, chicken, and sheep, while there are 14 country dummies: Austria, Belgium, Czech Republic, Estonia, France, Germany, Greece, Italy, Luxembourg, the Netherlands, Romania, Slovakia, Spain, and the UK. There is therefore no dummy for the animal type cattle and the country Denmark in order to avoid multicollinearity with the intercept. We present our results using robust standard errors, computed using the Arellano (1987) standard errors, which are robust for heteroskedasticity and correlation clustering.

3.2 Out-of-sample evaluation

In order to evaluate the models' accuracy in forecasting the PR, we conduct out-of-sample forecasts using mean absolute forecast error (MAFE) as the criterion. The out-of-sample period is from 2004 to 2015. We estimate the four regression models nested in Formula (1) using an expanding estimation window. The MAFE is defined as:

$$\mathrm{MAFE} = \frac{1}{T-S}\sum_{t=S+1}^{T} |e_t|\,, \tag{2}$$

where T is the total length of the series (608 observations), S is the burn-in period corresponding to the period 1969–2003 (458 observations), and $e_t = PR_t - \widehat{PR}_t$ is the one-step-ahead forecast error. The lower the MAFE, the better is the forecasting performance of the model. We test the significance of the difference in the MAFE between models using the Diebold-Mariano test (Diebold and Mariano, 2002).

4 RESULTS

Our main result is that for all models considered we cannot reject that, in the regression models of the agricultural PR on the corresponding ITCF, there is statistically and economically significant positive coefficient for the ITCF at a 95% confidence interval.

The in-sample and out-of-sample regression results of our main models for PR are reported in Tables 2 and 4. First, we discuss the in-sample parameter estimates and the goodness-of-fit statistics of these models. Then we evaluate the out-of-sample forecasting accuracy in terms of low values for the MAFE. After that, we discuss the robustness tests for which the results are presented in Table 4.

4.1 In-sample results

Let us first study the estimation results for the single-variate regression model in column (1) of Panel A in Table 2. We find that the least squares estimate of the slope coefficient is 1.024 with a robust standard error of 0.087. The ITCF is thus statistically significant at the 95% confidence interval. It is noteworthy that this simple model can already explain 34.6% of the variation in the price-ratios. The near-one value of the ITCF means that when the ITCF increases by one unit, the PR is expected to do the same, ceteris paribus. Importantly, the estimated coefficient remains around one, when controlling in columns (2)–(4) for the effects of animal type, country, and trend. In all specifications considered, the ITCF has an estimated coefficient of around one, and it is statistically significant at 95% confidence level.

Table 2 Determinants of the PR of processing livestocks

Panel A: In-sample regression estimates

	(1)	(2)	(3)	(4)
ITCF	1.024*** (0.087)	1.000*** (0.084)	1.089*** (0.160)	1.140*** (0.162)
Trend		0.005*** (0.001)		–0.002** (0.001)
Chicken			0.582*** (0.110)	0.612*** (0.113)
Pig			0.006 (0.074)	0.023 (0.075)
Sheep			0.920*** (0.122)	0.915*** (0.122)
Austria			0.143 (0.077)	0.164* (0.078)
Belgium			0.259*** (0.051)	0.244*** (0.051)
Czech Republic			0.224*** (0.048)	0.259*** (0.050)
France			–0.146** (0.038)	–0.176*** (0.040)

Table 2 (continuation)

Panel A: In-sample regression estimates

	(1)	(2)	(3)	(4)
Greece			0.159** (0.056)	0.168** (0.056)
Italy			0 . 171 *** (0.051)	0.147** (0.051)
Luxembourg			0.242*** (0.047)	0.231*** (0.047)
Netherlands			0.095* (0.037)	0.052 (0.037)
Spain			0.458*** (0.046)	0.475*** (0.046)
UK			0.079* (0.037)	0.064 (0.035)
Constant	0.008 (0.120)	–0.078 (0.127)	–0.378 (0.120)	–0.398 (0.140)
R^2	0.346	0.361	0.727	0.729
Adjusted R^2	0.345	0.359	0.720	0.722

Panel B: Out-of-sample forecast precision results

	(1)	(2)	(3)	(4)
MAFE for forecasting PR	0.354	0.388	0.281	0.282

$*p < 0.1; **p < 0.05; ***p < 0.01$

Note: This table presents in-sample and out-of-sample results of four regression models nested in Formula (1), our main explaining factor for the dependent variable PR is ITCF.
Source: Authors´ calculations

The most parsimonious model seems to be model (3), which compared to model (4), omits the trend variable but has similar goodness of fit. Note that in the model (4), the contribution of the trend variable is small compared to others with a coefficient equal to –0.002. Meanwhile, animal type and country variables are big contributors to the R^2. The single-variate model explains 34.6% of the variation in the price-ratios, while model (3) can describe 72.7% of the PR variability. This confirms the joint predictive power of animal type and country in explaining the producer PR of livestock, in addition to the ITCF.

4.2 Forecasting precision results

The out-of-sample forecasting performance of the four regression models is evaluated using the MAFE evaluation criterion, for which the results can be found in Panel B of Table 2. In general, the lower the MAFE, the better the forecast precision provided by the model. We can see that the single-variate model in column (1) of Table 2 returns a MAFE of 0.354. Combining ITCF with animal type and country dummies improves the forecasting precision of PR predicting models, as the MAFE significantly drops to 0.281 in the model in column (3). Meanwhile, adding the trend variable deteriorates the out-of-sample forecast precision. This is indicated by the increase in the MAFE of models in columns (2) and (4), in comparison with those in columns (1) and (3), respectively. This effect is consistent with the in-sample result that the trend variable has a minor contribution in predicting the PRs.

The statistical significance of differences of models' absolute forecast errors are evaluated using the Diebold-Mariano tests. Their p-values are presented in the left panel of Table 3. With all *p*-values less than 0.05, we conclude that the absolute forecast errors of all models are significantly different from the others.

Table 3 Diebold-Mariano test results for out-of-sample evaluation

Model	Regression models with dependent variable PR				Regression models with dependent variable log(PR)			
	1	2	3	4	1	2	3	4
1	-	<0.0001	0.006	0.004	-	0.1405	0.0006	0.0003
2	-	-	<0.0032	<0.0001	-	-	0.0032	0.0001
3	-	-	-	0.039	-	-	-	0.1213
4	-	-	-	-	-	-	-	-

Note: This table presents the Diebold-Mariano tests for out-of-sample forecast precision results, with forecasting target as the PR. The test is performed on forecasting error series of model *i* and *j* (*i,j* = 1,...4). The null hypothesis is that model *i* and model *j* have a similar level of accuracy. The alternative hypothesis is two models have different levels of accuracy. With *p-value* <0.05, we reject the null hypothesis. We use – to denote that two models do not require comparison, or the result is repeated and therefore not presented.
Source: Authors´ calculations

5 ROBUSTNESS ANALYSIS

Figure 1 shows the presence of outlying values for the PRs of chicken in Austria and sheep in Greece. Those outlying values may have a large effect on the estimates. As a robustness analysis, we therefore repeat the analysis but with the log(PR) as dependent variable. It can be seen in Figure 2 that the log transformation reduces the extremes.

Table 4 Robustness tests for the log(PR) of processing livestocks

Panel A: In-sample regression estimates

	(1)	(2)	(3)	(4)
ITCF	0.594*** (0.038)	0.586*** (0.037)	0.604*** (0.086)	0.645*** (0.087)
Trend		0.002** (0.001)		–0.002*** (0.0004)

Table 4 (continuation)

Panel A: In-sample regression estimates

	(1)	(2)	(3)	(4)
Chicken			0.261*** (0.056)	0.284*** (0.058)
Pig			–0.079* (0.040)	–0.065 (0.040)
Sheep			0.366*** (0.054)	0.362*** (0.054)
Austria			0.114** (0.040)	0.131*** (0.039)
Belgium			0.115*** (0.029)	0.103*** (0.028)
Czech Republic			0.132*** (0.028)	0.160*** (0.029)
France			–0.130*** (0.022)	–0.155*** (0.024)
Greece			–0.017 (0.029)	–0.010 (0.028)
Italy			0.078** (0.029)	0.059* (0.029)
Luxembourg			0.114*** (0.026)	0.105*** (0.026)
Netherlands			0.023 (0.021)	-0.011 (0.021)
Spain			0.288*** (0.026)	0.302*** (0.026)
UK			0.018 (0.024)	0.006 (0.022)

Table 4 (continuation)

Panel A: In-sample regression estimates

	(1)	(2)	(3)	(4)
Constant	–0.496*** (0.054)	–0.527*** (0.056)	–0.595*** (0.167)	–0.611*** (0.167)
R^2	0.401	0.407	0.762	0.767
Adjusted R^2	0.400	0.405	0.756	0.761

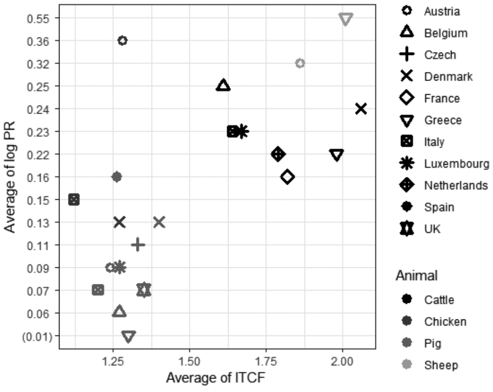
Panel B: Out-of-sample forecast precision results

	(1)	(2)	(3)	(4)
MAFE for forecasting PR	0.348	0.360	0.276	0.270

$* p < 0.1; ** p < 0.05; *** p < 0.01$

Note: This table presents results of four regression models (nested in Formula (3)) with dependent variable log(PR) and independent variables ITCF, animal type, and country dummies.
Source: Authors´ calculations

Figure 2 Scatter plots of the average log(PR)s versus the average of ITCFs



Note: This figure displays the average of log(PR)s in relation to the corresponding average ITCFs of 21 sample cases.
Source: Authors´ calculations

The regression model for the log(PR) of animal type *a* for country *c* in year *t* is:

$$\log(\mathrm{PR})_{a,c,t} = \alpha + \alpha_a + \alpha_c + \gamma \mathrm{Trend}_t + \beta\, \mathrm{ITCF}_{a,c,t} + \varepsilon_{a,c,t}\,. \qquad (3)$$

Note that the parameter β is now to be interpreted as a semi-elasticity indicating the expected percentage change in the PR when the ITCF increases by 1%. In other words, if the ITCF increases by 1 unit, we can expect the PR to increase on average by 100β%, ceteris paribus.

The results of in-sample tests for regression models of log(PR) with ITCF, animal type, and country are presented in Table 4. We find that the least squares estimates of the slope coefficients of ITCF with log(PR) are stable around 0.6 for all four nested models and statistically significant at the 95% confidence interval. If the ITCF increases from 1 to 2, the price ratio is expected to increase by 60% on average, ceteris paribus. Based on Figure 1, we can see that this effect is smaller than the unit slope coefficient found in the model with PR as the dependent variable. This smaller effect follows from the fact that taking log-transformations of PRs dampens the effect of the vertical outliers for the prices ratios of chicken in Austria and sheep in Greece. In general, we see that the estimations support our conclusions for the main regression models of PR and three independent variables.

In terms of explanatory power, we find that the single-variate model can explain 40.1% of the variation in the log(PR) and that adding animal and country variables substantially increases the goodness of fit. This is demonstrated by the change of R^2 in the model in column (3) of Table 4, from 40.1% to 76.2% compared to the model in column (1). Meanwhile, adding the variable Trend has a minor impact on the good of fitness, as its presence increases only the R^2 with 0.5 percentage points.

In Panel B of Table 4, we report the results of the out-of-sample evaluation of the forecasting precision of PRs using the log(PR)-based prediction models. Here, the single-variate model has the highest MAFE of 0.348, while the model combining all four independent variables has the lowest MAFE of 0.277. To check whether or not the forecasting errors between models are significantly different, the Diebold-Mariano tests are used. The p-values in the right panel of Table 3 show that there is no improvement in prediction power of models in the first two columns of Table 4. However, the forecasting errors are significantly lower when we combine three explaining variables: ITCF, animal type, and country. Moreover, even though the model in column (4) has the lowest MAFE, its prediction precision is not significantly better than the model in column (3). As such, among the four models, we recommend model (3) to be the most parsimonious for forecasting the PR. Note that with the same predictors, this preferred model with the log(PR) yields more accurate PR predictions (MAFE of Panel B in Table 2) than the model with the PR as dependent variable (MAFE of Panel B in Table 4). We verified that the forecasting precision is statistically significant at the 95% confidence level.

DISCUSSION AND CONCLUSION

In this paper, we study the TCF and its relation with the PR of processing livestock from live weight meat to carcass meat. Studying this relation is important for two reasons. First, understanding the close relation between the TCF and the PR is important to comprehend the passthrough between the physical efficiency of the processing of livestock and the corresponding PRs. Second, from a statistical perspective, the TCF can be used for imputation when prices are missing.

We proposed a simple model to predict the PR using the TCF, making it feasible to implement the imputation in a setting with a large number of countries and products. Such a large-scale analysis is of great importance for official institutions, as mentioned by Boudt et al. (2009). We concentrate on four major animal types (cattle, chicken, pig, and sheep) and countries belonging to the European CAP. The empirical analysis confirms that there is a statistically and economically significant relation between TCFs

ACKNOWLEDGMENTS

We are grateful to Sangita Dubey for recommending this research problem to us. We also thank two anonymous reviewers, Danny Coomans and participants at the *IAOS-OECD Conference: Better Statistics for Better Lives* (19–21 September 2018 in Paris) for their helpful comments.

References

ARELLANO, M. Practitioners' corner: Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 1987, 49(4), pp. 431–434.

ASCHE, F., OGLEND, A., TVETERAS, S. Regime shifts in the fish meal/soybean meal price ratio. *Journal of Agricultural Economics*, 2013, 64(1), pp. 97–111.

BOUDT, K., TODOROV, V., UPADHYAYA, S. Nowcasting manufacturing value added for cross-country comparison. *Statistical Journal of the IAOS*, 2009, 26(1, 2), pp.15–20.

CHAUDHARY, A., PFISTER, S., HELLWEG, S. Spatially explicit analysis of biodiversity loss due to global agriculture, pasture and forest land use from a producer and consumer perspective. *Environmental Science & Technology*, 2016, 50(7), pp. 3928–3936.

CHAVAS, J. P. AND HOLT, M. T. On nonlinear dynamics: The case of the pork cycle. *American Journal of Agricultural Economics*, 1991, 73(3), pp. 819–828.

DIEBOLD, F. X. AND MARIANO, R. S. Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 2002, 20(1), pp. 134–144.

DUBEY, S., KATZ, S., TAYYIB, S. Data and analysis to make international comparisons: Availability of existing FAOSTAT data on agricultural production, trade and prices. *Pre-conference workshop: Linkages between agricultural economics and statistics, International Conference of Agricultural Statistics (ICAS)*, FAO Headquarters, Iran Room, Rome, Italy, October 25, 2016, pp. 2–18.

EUROSTAT. *European statistics* [online]. 2017. [cit. Mar. 2017] <http://ec.europa.eu/eurostat/ data/database>.

EUROSTAT. *Selling prices of agricultural products (absolute prices)* [online]. 2018. [cit. Mar. 2018] <https://ec.europa.eu/eurostat/cache/metadata/en/apri_ap_ esms.htm>.

FAO. *Livestock statistics: Concepts, definitions and classifications.* Technical report, Food and Agriculture Organization of the United Nations, 2011.

FAOSTAT. *Technical conversion factors for agricultural commodities.* Food and Agriculture Organization of the United Nations, Statistics Division, 1960.

FAOSTAT. *Technical conversion factors for agricultural commodities.* Food and Agriculture Organization of the United Nations, Statistics Division, 1972.

FAOSTAT. *Technical conversion factors for agricultural commodities.* Technical report, Food and Agriculture Organization of the United Nations, Statistics Division, 2009.

FAOSTAT. *Food and agriculture data* [online]. 2017. [cit. April 2017] <http://www.fao.org/waicent/faostat/agricult/prodpric-e.htm>.

FAOSTAT. *Producer Prices – Annual* [online]. 2018. [cit. Mar. 2018] <http://www.fao.org/faostat/en/#data/PP>.

HOLT, M. T. AND CRAIG, L. A. Nonlinear dynamics and structural change in the US hog-corn cycle: A time-varying star approach. *American Journal of Agricultural Economics*, 2006, 88(1), pp. 215–233.

KIRKENDALL, N. *Data and research to improve the US food availability system and estimates of food loss.* Workshop report, National Academy of Sciences, 2015, pp. 67–149.

KLEIBER, C. AND LUPI, C. *Panel Unit Root Testing with R* [online]. R package, 2011. <http://cran.r-project.org/web/views>.

LAZARUS, E. et al. *Working guidebook to the national footprint accounts: 2014 Ed.* Global Footprint Network, 2014.

LUAN, Y., CUI, X., FERRAT, M., NATH, R. Dynamics of arable land requirements for food in South Africa: From 1961 to 2007. *South African Journal of Science*, 2014, 110(1–2), pp. 1–8.

OECD-FAO. *Aglink-Cosimo Model Documentation – A partial equilibrium model of world agricultural markets.* The Secretary-General of the OECD, 2015.

PARKER, P. S. AND SHONKWILER, J. On the centenary of the German hog cycle: New findings. *European Review of Agricultural Economics*, 2013, 41(1), pp. 47–61.

SMITH, M. R., MICHA, R., GOLDEN, C. D., MOZAFFARIAN, D., MYERS, S. S. Global expanded nutrient supply (genus) model: A new method for estimating the global dietary supply of nutrients. *PloS One*, 2016, 11(1).

TVETERAS, S. AND ASCHE, F. International fish trade and exchange rates: An application to the trade with salmon and fishmeal. *Applied Economics*, 2008, 40(13), pp. 1745–1755.

Boomerang Effect of Quality Control on the Compilation of Financial Accounts and Flow of Funds: the Experience of Banco de Portugal

Susana Santos[1] | *Banco de Portugal, Lisbon, Portugal*
António Agostinho[2] | *Banco de Portugal, Lisbon, Portugal*

Abstract

Financial Accounts are fundamental to monitor financial stability by quantifying the impact of financial decisions of a host of economic agents. In Portugal, the compilation of these statistics is a responsibility of Banco de Portugal. One of the main purposes of the Statistics Department of Banco de Portugal is to ensure this statistical production with high quality standards, aiming at fully meeting user's needs, by developing a wide set of quality control procedures. In this context, Banco de Portugal developed a multidisciplinary team with experts from financial accounts and from the different underlying primary statistics. Within this format, that can be viewed as a "boomerang effect", all team members are co-responsible for producing national financial accounts, on a bottom-up approach, thus improving both the quality of these statistics, as well as the quality of primary statistics. This is the result of a systematic iterative process of data cross-check and reconciliation which may represent an opportunity to validate the soundness of microdata, on a top-down approach.[3]

Keywords	*JEL code*
Quality control, financial accounts, multidisciplinary team, data cross-check, flow of funds	*G20, L15*

INTRODUCTION

In a context of an increasingly complex economic and financial reality, the National Financial Accounts (hereinafter referred as "financial accounts") are fundamental to monitor financial stability by quantifying the impact of financial decisions of the economic agents. National financial accounts provide an overall

[1] Banco de Portugal, Statistics Department, Statistics Audit Unit, R. Francisco Ribeiro 2, 1150-165 Lisbon, Portugal. E-mail: smsantos@bportugal.pt.

[2] Banco de Portugal, Statistics Department, Statistics Audit Unit, R. Francisco Ribeiro 2, 1150-165 Lisbon, Portugal. E-mail: afagostinho@bportugal.pt.

[3] This article is based on contribution at the *European Conference on Quality in Official Statistics (Q2018)* in June 2018 in Krakow, Poland.

view of the financial interlinkages between institutional sectors helping in the identification of sector vulnerabilities, imbalances and potential over-exposures to certain financial instruments.

The statistical function of central banks is changing and it is important to develop solutions that contribute to enhance the effectiveness and efficiency of its statistical system. In this context, the quality of the financial accounts statistics is a priority for Banco de Portugal, which is the competent statistical authority in this domain. To follow this purpose, the Statistics Department developed a multidisciplinary team with experts from financial accounts and from the different underlying primary statistics. This new approach established a cooperative work, with a positive impact on the quality and consistency among the various statistics produced in Banco de Portugal, and it can be viewed as a "boomerang effect".

1 METHODOLOGICAL FRAMEWORK

Financial accounts are one of the components of the national accounts that records two kinds of information, flows and stocks, between the different institutional sectors of the economy and between these sectors and the "rest of the world".

These statistics are prepared in accordance with the guidelines set out in the European System of National and Regional Accounts (ESA 2010) – Regulation (EU) No 549/2013 of the European Parliament and of the Council of 21 May 2013.

Accordingly to ESA 2010, "flows refer to actions and effects of events that take place within a given period of time, while stocks refer to positions at a point of time". Stocks (also referred as positions or outstanding amounts) are the holdings of assets and/or liabilities at a given point of time, recorded at the end of each accounting period.

Institutional sectors are economic agents, or "institutional units", with the same economic role, grouped according to the sectorial classification rules of ESA 2010, based on the type of producer, function and main activity: non-financial corporations, financial corporations, general government, households and non-profit institutions serving households and, rest of the world.

Financial accounts are broken down by financial instruments, such as: monetary gold and special drawing rights; currency and deposits; debt securities; loans; equity and investment fund shares or units; insurance, pensions and standardised guarantee schemes; financial derivatives and employee stock options; and other accounts receivable and payable.

The accounting principle underlying the national accounts is a quadruple-entry principle, i.e. each operation must be entered twice by the two parties involved.

The financial accounts are considered derived statistics as they are based on a vast array of other primary statistics, including, in the case of Portugal, balance of payments and international investment position statistics, monetary and financial statistics, central balance sheet statistics, securities statistics and central credit register statistics.

Although the main data sources are internal to Banco de Portugal, external data sources are also used, such as the information provided by the Portuguese Insurance and Pension Funds Supervisory Authority, the Portuguese Treasury and Debt Management Agency and, the Portuguese National Statistical Institute (Instituto Nacional de Estatística – INE). Actually, INE is responsible for compiling the national non-financial account, while Banco de Portugal takes the responsibility for the compilation of the national financial account (Banco de Portugal already produced a cluster of statistics that is necessary for compiling financial accounts), following a protocol signed in 1998 between Banco de Portugal and INE. This protocol provides for the establishment of mechanisms of cooperation, mutual consultation and methodological discussion on the compilation of national accounts, in particular regarding the harmonised implementation of the European System of National and Regional Accounts. This interaction leads to better quality in the two types of accounts.

Due to this aggregation of multiple sources of information, financial accounts provide a picture of the impact of financial decisions among the different economic agents. These statistics provide an overall view of the financial interlinkages between institutional sectors, helping in the identification of sector vulnerabilities, imbalances and potential over-exposures to certain financial instruments. In Portugal, this kind of analysis turned out to be very useful in a context of the global financial crisis, because it enables an overview of the degree of intermediation of the financial sector and of the structure of private sector wealth. With these statistics it is possible to measure the relationships and interconnections between the different institutional sectors of the economy and to monitor their exposure to different risks.

To better understand economic sectors' interlinkages and to assess how intersectoral financial linkages have changed, flow of funds is a powerful analytical tool. This type of analysis allows to detail the data by counterpart sector and type of financial instrument, identifying specific economic behaviours. It enables to analyse intersectoral relationships among the resident sectors of an economy and between these and the rest of the world.

At this point, flow of funds are a subset of the financial accounts, as it allows to establish the net transactions between the different institutional sectors. This data gives the user an overall picture of the whole economy, since financial accounts, by being at the end of the cycle, is the only system where all sectors of the economy are put together in an integrated system.

2 A MULTIDISCIPLINARY TEAM

One of the main purposes of the Statistics Department of Banco de Portugal is to ensure the production of high quality statistics and to provide a more efficient data quality management in statistical systems, developing a wide set of quality control procedures.

Following this purpose of high quality standards, the Portuguese solution to compile national financial accounts was to develop a multidisciplinary team, with experts from financial accounts and from the different underlying primary statistics.

This multidisciplinary team, that involves the different statistical domains, was created by the end of 2009, with national financial accounts experts, permanently allocated to financial accounts' tasks, and two experts of each underlying primary statistics (one effective and one substitute). It is chaired by the National Financial Accounts Head of Division of Statistics Department.

This new organizational model of compiling financial accounts can be easily transposed to any kind of organization where the final goal is to improve quality and consistency. This can be seen as a project organisation where management structures coexist in the form of a matrix management structure, instead of a traditionally hierarchy management organisation. Despite all its advantages, this kind of organisational model is nevertheless more demanding in terms of coordination.

This multidisciplinary team turns out to be very efficient as all members are actively engaged in collectively contributing to the end-product, producing a high quality output. It is a collective effort that benefits from the expertise of the technicians of each primary statistics in analysing the specific data of their domain. For instance, experts from the Central Balance Sheet Statistics Unit provide not only primary data but are also specifically responsible for the compilation of the non-financial corporations sector account, and are more generally co-responsible for national financial accounts (Matos, 2016).

The responsibility of the compilation of financial accounts is shared by all team members and distributed as follows:

- The compilation of each institutional sector is provided by the statistical area that is responsible for the majority of primary data. For instance, the compilation of financial accounts of general government is allocated to the General Government Statistics Unit; and the compilation regarding the financial sector is a responsibility of the Monetary and Financial Statistics Division;

- The securities statistics data is provided by the Securities Statistics Unit, which has detailed information concerning the issues, issuers and holdings of the securities in the resident country (this information can be easily compared with the one sent by monetary institutions to the Monetary and Financial Statistics Division);
- Methodological definitions and procedures are a responsibility of the Methodological Statistics Unit in cooperation with the Financial Accounts Unit;
- The final management of financial accounts, namely the aggregation of all statistical institutional sectors data and the disclosure of national financial accounts outputs to final users, is a responsibility of the Financial Account Statistics Unit.

Since managing such a multidisciplinary team is not an easy task, Banco de Portugal has been adopting a stepwise approach, since 2009. It has been a "work in progress" system, as it turned out to be very useful in developing new ways of improving quality, not only for the final statistics output of financial accounts, but also for primary statistics.

For instance, when ESA 2010 was implemented, financial accounts faced the need of implementing a new information system compliant with the new recommendations. This was also the opportunity for improving the financial accounts compilation system. Instead of developing a new system within the Unit, all the members of this multidisciplinary team were involved. The benefits were clear: the system was defined with a minuteness detail because each team member developed their procedures in the new system attending the needs of the new guidelines. On another hand, the consistency between primary statistics and this ones, as well as with the previous and the final output of financial accounts was preserved and guaranteed. In this case, all the details of the primary data are analyzed and checked with information of another statistical domains.

This multidisciplinary team has faced, during the later years, several improvements, not only concerning the system underlying the compilation procedures, but also in the management of resources. The final goal is always to improve quality and increase process efficiency.

Besides the quality improvements, there are also several costs due to the complexity involved with the coordination and management of such a team. First of all, this kind of work organisation must have a very good planning calendar, and hierarchic managers and multidisciplinary team managers must agree over the allocation of resources. One of the main problems that this kind of organisation structure may face is the risk that the team members receive conflicting tasks. To avoid this kind of conflicts, priorities must be agreed and all team members must be aware of their roles. Their activities must be settled in each team member annual planning, for both matrix management and hierarchy management, and should be captured and reflected in their performance evaluation. Managing people with more than one reporting line is a big challenge and it is very important to clarify who has the responsibility to evaluate the performance of each team member for which task.

3 BOOMERANG EFFECT

This new method of compiling financial accounts, which can be easily described as a bottom-up and top-down approach, can generate many benefits, ensuring a high quality of the financial accounts outputs, as well as a better quality of primary statistics. Thus, experts gain a global insight of how their data affect other statistics and are able to take interrelated and synergic combined final decisions.

The main result of this boomerang effect is to take advantage of the interaction and cooperation between the different statistical areas, to ensure the quality and establish different levels of responsibility in the compilation of financial accounts. This is achieved by separating the data processing activities from the activities of analysing and exploring the information. However, this multidisciplinary team shares the responsibility for the entire production cycle of the compilation of financial accounts.

This approach encourages the cooperative work between the different areas of the statistics department, and promotes a more efficient contribution of the primary data to the financial accounts compilation. It also avoids duplicating the tasks of compiling data for the primary statistics in one moment, and after that compiling the same information for the compilation of financial accounts purpose. On another hand, the primary statistics benefit from the concerted data produced by the compilation process of financial accounts. This boomerang effect is an opportunity to implement not only internal quality control procedures, but also to ensure consistency between statistics produced.

This results on a systematic iterative process of data cross-check and reconciliation which may represent an opportunity to validate the soundness of micro data, on a top-down approach. It promotes the consistency of the financial accounts between the institutional sectors, because, for all instruments, the assets of one sector must be equal to the liabilities of the counterpart sector. Thus, the validation of the final output of financial accounts must fulfil horizontal and vertical consistency.

Horizontal consistency is an internal validation that ensures inter-sector consistency for the different types of information, while vertical consistency certifies that financial accounts outputs are consistent with final data of the primary statistics, despite the discrepancies that may exist due to different methodological processes.

The multidisciplinary team can serve the purpose of different statistical domains. First of all, primary statistics feed the system with data that is an output of their own compilation process, which have already met the first level of quality control tests within their respective production cycle.

It is important to refer that primary statistics are the owners of granular information concerning the institutional sector that they are responsible for. This granular data is often stored into different micro databases, which are a powerful tool with a high statistical potential. For instance, Monetary and Financial Statistics comprise the Balance Sheet Information on Financial Corporations that has granular information on assets and liabilities of the sector; the Balance of Payments and International Investments Position system has micro data on the assets and liabilities of the rest of the world sector; the Central Credit Register contains granular information on credit exposure and loans to all sectors of the economy; the Securities Statistics Integrated System is a security-by-security and investor-by-investor database of securities holdings and issues; the Central Balance Sheet Database contains accounting and financial information of non-financial corporations.

On the other hand, the Financial Account Unit can input into the system the information they need with a high quality standard, as the information is already confirmed and validated by the respective primary statistic's owners. This process provides more complete and detailed statistics, aiming at fully meeting user's needs, with high quality standards of the final output.

Additionally, the potential problems and inconsistencies among primary statistics are analysed before the final compilation of the financial accounts and all the institutional sectors take combined decisions aiming at the internal and external consistency of the final results.

Ultimately, this joint coordination effort requires also an alignment of the revisions policy for statistical domains involved.

CONCLUSION

"Good statistics are a precondition to good policy-making" (Matos and Nunes, 2017), and the way Banco de Portugal achieved this goal in National Financial Accounts was through the creation of a multidisciplinary team that has been a success in the compilation of these statistics.

Although demanding in terms of management, this new method has proved to improve the consistency between statistics as well as the quality of primary and final financial accounts statistics disseminated. Users' needs are thus more easily met, allowing for greater integration and consistency between the different statistical products.

The success of this multidisciplinary team work is confirmed by a more efficient production process and a higher quality output.

It can be viewed as a boomerang effect, as the final output of financial accounts is also likely to provoke a number of second-order consequences, namely the better quality and coherence of primary statistics, and raise awareness of primary statistics compilers (also part of the financial accounts compilation team) to what needs to be done as preparatory work for producing consistent statistics.

The other side of the coin of this matrix organisational structure relates to the challenges in terms of planning and management. However, the Portuguese experience provides evidence that such costs are clearly outweighed by the benefits.

References

CADETE DE MATOS, J. *Innovative solutions in compiling financial accounts.* 2016.

CADETE DE MATOS, J. AND NUNES, L. *Upgrading Financial Accounts with Central Balance Sheet Data – What's in it for central bank's policy?* 2017.

European System of National and Regional Accounts (ESA 2010). Regulation (EU) No 549/2013 of the European Parliament and of the Council of 21 May 2013.

LIMA, F. AND MONTEIRO, O. *An integrated analysis of the Portuguese economy: the financial and the real economy.* Paper presented at the 58th World Statistics Congress of the International Statistical Institute held in Dublin, Ireland, 2011.

OECD. *Understanding Financial Accounts* [online]. Paris: OECD Publishing, 2017. <http://dx.doi.org/10.1787/9789264281288-en>.

Pattern Normalization – a New Tool for Dynamic Comparisons

Iwona Müller-Frączek[1] | *Nicolaus Copernicus University, Toruń, Poland*

Abstract

The article presents a new method of normalization – normalization with respect to pattern (or pattern normalization in short). It has properties expected for this type of transformation: preserves skewness, kurtosis and the Pearson correlation coefficients. Although pattern normalization uses only observations from the current unit of time, it can be used in dynamic research. An additional advantage of new normalization is the ability to reflect different analysis environments. The effects of pattern normalization are illustrated by an empirical example. Indicators monitoring the implementation of the Europe 2020 Strategy are used. Normalizations are carried out for two reference groups: the entire EU and countries that joined the EU in 2004. The results for two years are compared. The example of Poland shows that the "dynamic image" of the country is affected by the use of pattern normalization itself as well as by the choice of the environment. In this context pattern normalization is similar to dynamic standardization, and different than dynamic scaling.

Keywords	***JEL code***
Normalization, standardization, transformation of variables	*C43, C19, C38*

INTRODUCTION

We understand normalization as procedure of pre-treatment of data in order to allow for their mutual comparison and further analysis. Such a procedure is used, for example, in a study of a complex phenomenon, i.e. a qualitative phenomenon that is characterized by a collection of quantitative variables. Without losing generality, we assume that this is the phenomenon observed for objects in space, such as socio-economic development of countries. In this case, normalization deprives variables of their units and unifies their ranges. After normalization we can compare variables separately or construct a composite indicator. The composite indicator is one-dimensional image of multidimensional phenomenon (compare Saisana and Saltelli, 2011; Saltelli, 2007).

There are many normalization formulas (see Jajuga and Walesiak, 2000; Milligan and Cooper, 1988; Młodak, 2006; Steinley, 2004). Most often they are given for a static analysis, i.e. for a fixed point in time. Normalization problems appear when we want to compare a given phenomenon at several time points. In this case diagnostic variables should also be comparable over time.

To achieve this effect we can use two approaches. In first of them, we exploit all values of variable (both in space and time) to determine the parameters needed for normalization (compare Nardo et al., 2005). We can call this approach the stochastic one, because we treat observations for a given time point as randomly selected sample of population. But it is rather controversial in regional comparisons where

[1] Faculty of Economic Sciences and Management, Gagarina 13 a, 87-100 Toruń, Poland. E-mail: muller@umk.pl, phone (+48)566114718.

we work with the whole population of objects in space, but not with a sample (compare Zeliaś, 2002). In addition, a practical disadvantage of this solution is the need to recalculate all results with the appearance of observations for next unit of time.

In the second approach, parameters needed for normalization do not result directly from variable distributions. They are taken in advance, the same for all the units of time (also future). This solution is used, for example, in the very popular Human Development Index (HDI), as well as in a newer proposal i.e. the Adjusted Mazziotta-Pareto Index, called AMPI in short (compare Mazziotta and Pareto, 2015, 2016).

The article proposes another way of solution to dynamic problems. We introduce a new method of feature normalization – normalization with respect to the pattern (or pattern normalization for short). The method is consistent with the static approach (only current observation are taken), but it can be used to compare objects at different time points. The method meets requirements of normalization that are suggested in literature (compare e.g. Jajuga and Walesiak, 2000; Młodak, 2006). It preserves skewness and kurtosis. Moreover, the absolute values of the Pearson correlation coefficients are not changed after normalization.

An additional advantage of pattern normalization is the possibility to reflect different environments in research. This is the same as in standardization and on the contrary to scaling (or min-max normalization) used in the mentioned HDI or AMPI. This property is illustrated by an empirical example. Indicators monitoring the implementation of the Europe 2020 Strategy (see European Commission, 2010) are normalized in two environments, one is the whole European Union, and the second – a group of countries that joined the EU in 2004. The example of Poland shows differences for both environments.

The article is divided into 6 parts. Section 1 introduces the normalization with respect to pattern. Section 2 presents properties of the pattern normalization. Section 3 discusses advantages of new proposal. Section 4 illustrates theoretical consideration. The article ends with conclusions.

1 DEFINITION OF PATTERN TRANSFORMATION

Consider a set of $n \in N$ objects in space. For these objects, we analyze a phenomenon which is not directly measurable and it is composed of many aspects (a complex phenomenon). Various aspects of this phenomenon are characterized by measurable diagnostic variables, that is, variables for which a connection with a certain aspect of the complex phenomenon is not in doubt and the direction of this relationship can be determined (a stimulant is a diagnostic variable that has a positive impact on the analyzed complex phenomenon, while a destimulant negative).[2] An example of a complex phenomenon is socio-economic development of the European Union countries, and diagnostic variables for this phenomenon are, among others, the indicators monitoring implementation of the Europe 2020 Strategy (considered in Section 4).

The analyzed objects can be ordered due to individual diagnostic variables, i.e. in relation to particular aspects of the complex phenomenon. To order objects due to all aspects of this phenomenon, we can construct a synthetic variable (a composite indicator).[3] One of the stages of such construction is normalization of variables.

For a given unit of time consider one diagnostic variable $x = (x_1, x_2, \dots , x_n) \in R^n$. This variable is a stimulant (then we write $x \in S$, where S denotes the set of stimulants) or a destimulant ($x \in D$ respectively). We choose a pattern – the most beneficial of all values of the variable x. This name was inspired by the Hellwig's paper (Hellwig, 1968). The pattern is unique for all objects and is described by the formula:

[2] Other types of variables are not considered. If they must be used in the study, they should be transformed into stimulants.

[3] In this case, the diagnostic variables must meet additional statistical requirements such as sufficient variability or weak correlation. This is beyond the scope of the article, for more details we refer, for example, to Zeliaś (2002).

$$x^{+} = \begin{cases} \max\limits_{i=1,\dots,n} x_i & \text{if} \quad x \in S, \\ \min\limits_{i=1,\dots,n} x_i & \text{if} \quad x \in D. \end{cases} \tag{1}$$

After specifying the pattern x^{+} we can consider a new variable u instead of the variable x given by:

$$u_i = \frac{|x_i - x^{+}|}{\sum_{j=1}^{n} |x_i - x^{+}|} = \begin{cases} \frac{x^{+} - x_i}{\sum_{j=1}^{n}(x^{+} - x_j)} & \text{if} \quad x \in S, \\ \frac{x_i - x^{+}}{\sum_{j=1}^{n}(x_j - x^{+})} & \text{if} \quad x \in D. \end{cases} \quad \text{for} \quad i = 1, \dots, n. \tag{2}$$

The Formula (2) determines a transformation of initial variable $x = (x_1, x_2, \dots, x_n)$ into a new variable $u = (u_1, u_2, \dots, u_n)$. After this transformation (transformation with respect to pattern) the new variable describes the same aspect of complex phenomenon as x describes. So u is a diagnostic variable of this phenomenon.

2 STATIC PROPERTIES OF PATTERN TRANSFORMATION

2.1 Basic properties

1. All diagnostic variables after pattern transformation are unitless, non-negative and limited to interval [0,1]. Because of that, the new set of diagnostic variables contains comparable elements.
2. The lower is the value u_i the better is the situation of the i-th object. It means that the variable after the pattern transformation becomes destimulant irrespective of its initial nature. So, the pattern transformation unifies the nature of the diagnostic variables.
3. Transforming of variables does not affect the ordering of objects.

2.2 Extreme values after pattern normalization

1. The variable u can take the zero value only for the pattern object:

$$u_i = 0 \Leftrightarrow x_i = x^{+} \quad \text{for} \quad i = 1, \dots, n. \tag{3}$$

2. Since the pattern is chosen among the values of the variable x, the zero value is taken:

$$\min_{i=1,\dots,n} u_i = 0. \tag{4}$$

3. The value u_i equals 1 when all objects except the i-th one are patterns:

$$u_i = 1 \Leftrightarrow \forall_{j \neq i} \; x_i = x^{+} \quad \text{for} \quad i = 1, \dots, n. \tag{5}$$

4. The maximum value of u depends on the nature of variable x:

$$\max_{i=1,\dots,n} u_i = \begin{cases} \frac{\max\limits_{i=1,\dots,n} x_i - \min\limits_{i=1,\dots,n} x_i}{\sum_{j=1}^{n}(\max\limits_{i} x_i - x_j)} & \text{if} \quad x \in S, \\ \frac{\max\limits_{i=1,\dots,n} x_i - \min\limits_{i=1,\dots,n} x_i}{\sum_{j=1}^{n}(x_j - \min\limits_{i=1 \quad n} x_i)} & \text{if} \quad x \in D. \end{cases} \tag{6}$$

2.3 Descriptive characteristics of transformed variables

1. The mean value of u depends only on the number of objects:

$$\bar{u} \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} u_i = \frac{1}{n} \,. \tag{7}$$

2. The variance of u is described by:

$$S^2(u) \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})^2 = \frac{S^2(x)}{n^2\,(x^+ - \bar{x})^2} \,. \tag{8}$$

3. The standard deviation of u depends on the nature of variable x and it is expressed by:

$$S(u) \overset{def}{=} \sqrt{S^2(u)} = \begin{cases} \frac{S(x)}{n(x^+ - \bar{x})} & \text{if} \quad x \in S, \\ \frac{S(x)}{n(\bar{x} - x^+)} & \text{if} \quad x \in D. \end{cases} \tag{9}$$

4. The coefficient of variation of u is given by:

$$CV(u) \overset{def}{=} \frac{S(u)}{\bar{u}} = \begin{cases} \frac{S(x)}{x^+ - \bar{x}} & \text{if} \quad x \in S, \\ \frac{S(x)}{x \,\bar{-}\, x^+} & \text{if} \quad x \in D. \end{cases} \tag{10}$$

5. The 3rd central moment of u is expressed by:

$$\mu_3(u) \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})^3 = \frac{\mu_3(x)}{n^3\,(\bar{x} - x^+)^3} \,. \tag{11}$$

6. The absolute value of the coefficient of skewness is preserved:

$$\mathrm{A}(u) = \frac{\mu_3(u)}{\mathrm{S}^3\,(u)} = \begin{cases} -\mathrm{A}(x) & \text{if} \quad x \in S, \\ \mathrm{A}(x) & \text{if} \quad x \in D. \end{cases} \tag{12}$$

7. The 4th central moment of u is given by:

$$\mu_4(u) \overset{def}{=} \frac{1}{n} \sum_{i=1}^{n} (u_i - \bar{u})^4 = \frac{\mu_4(x)}{n^4\,(x^+ - \bar{x})^4} \,. \tag{13}$$

8. The kurtosis of u does not change after the pattern transformation:

$$\mathrm{K}(u) \overset{def}{=} \frac{\mu_4(u)}{\mathrm{S}^4(u)} - 3 = \mathrm{K}(x). \tag{14}$$

2.4 Linear relation between variables after transformation

Denote by $u_1 = (u_{11}, u_{12}, \dots , u_{1n})$ and $u_2 = (u_{21}, u_{22}, \dots , u_{2n})$ two diagnostic variables after pattern transformation.

1. The covariance between u_1 and u_2 equals:

$$\operatorname{cov}(u_1, u_2) \overset{def}{=} \frac{1}{n}\sum_{i=1}^{n} (u_{1i} - \overline{u}_1)\,(u_{2i} - \overline{u}_2) = \begin{cases} \dfrac{\operatorname{cov}(x_1, x_2)}{n^2\,(x_1^+ - \overline{x}_1)\,(x_2^+ - \overline{x}_2)} & \text{if } x_1, x_2 \in S \text{ or } x_1, x_2 \in D, \\ \dfrac{-\operatorname{cov}(x_1, x_2)}{n^2\,(x_1^+ - \overline{x}_1)\,(x_2^+ - \overline{x}_2)} & \text{otherwise.} \end{cases} \tag{15}$$

2. The absolute value of the Pearson correlation coefficient is preserved:

$$\rho(u_1, u_2) \overset{def}{=} \frac{\operatorname{cov}(u_1, u_2)}{\mathrm{S}(u_1)\cdot \mathrm{S}(u_2)} = \begin{cases} \rho(x_1, x_2) & \text{if } x_1, x_2 \in S \text{ or } x_1, x_2 \in D, \\ -\rho(x_1, x_2) & \text{otherwise.} \end{cases} \tag{16}$$

3 DISCUSSION ON PATTERN TRANSFORMATION

The transformation described by Formula (2) can be called normalization, because it makes variables comparable (1) and has expected properties. First, it preserves two important characteristics of variable distribution – skewness (6) and kurtosis (8). Second, this conversion does not disrupt linear relation between variables – the absolute value of the Pearson correlation coefficient does not change (2).

Unlike other methods the pattern normalization is not just a technical procedure, it has clear interpretation. u_i specifies the share of distance between the i-th object and the pattern in the total distance of all objects from the pattern. We can say that by pattern normalizing we get a relative assessment of the objects situations.

The values of the variable u characterize the positions of objects in the whole system of objects. The value u_i is influenced by all the values of the variable x, so it is important in which environment (a reference group) the normalization is carried out. This is particularly important when analyzing changes of u over time. For one reference group, normalized values for i-th object can increase, and for another group they can decrease. Such situations are presented in an empirical example described in Section 4.

Similar property occurs for standardization:

$$z_i = \frac{x_i - \bar{x}_i}{\mathrm{S}(x)}, \quad \text{for} \quad i = 1, \dots , n, \tag{17}$$

where the reference group is represented by the arithmetic mean $\bar{x}$ and standard deviation $\mathrm{S}(x)$, calculated on the basis of the values for all objects. However, it is different for scaling (or min-max normalization):

$$s_i = \frac{x_i - \min\limits_i x_i}{\max\limits_i x_i - \min\limits_i x_i}, \quad \text{for} \quad i = 1, \dots , n. \tag{18}$$

In this case only the maximum and minimum values represent the environment and influence the values of the variables after transformation.

A major advantage of normalization with respect to pattern appears in dynamic approach. In the case of other types of normalization, if we transform the variable for each time unit separately (i.e. we use a static approach for each unit of time), the results are not comparable over time. To achieve comparability over time, two ways are possible. Firstly, the parameters needed for normalization (e.g. average, deviation, extreme values) can be determined on the basis of all observations (in space and time). Secondly, some reference values for these parameters can be established, that are common to all objects and all (also future) units of time (this can be done on the basis of expert knowledge).

In the case of pattern normalization, we obtain comparability over time using a static approach, because for each unit of time, we distribute the same "mass" (equal to 1) between the same number of objects. For a given object if value of a normalized variable increases, it means that this object increases its share in the total distance from the pattern. So in comparison to other objects, it moves away from "the best" object, so its relative situation is getting worse. Although current data are the sole data used to convert variables, after normalization variables are naturally comparable over time.

The property mentioned above is very advantageous when creating dynamic synthetic variables (composite indicators). The results obtained for a certain time interval are permanent and do not require recalculation after the appearance of observations for the next time period.

4 EMPIRICAL EXAMPLE

To illustrate the effects of pattern normalization, indicators monitoring implementation of the Europe 2020 Strategy (European Commission, 2010) are used. Data come from the statistical office of Poland (Statistics Poland, 2018). 4 stimulants and 7 destimulants are transformed. They are:

x_1 – Gross domestic expenditure on R&D (*% of GDP*; $x_1 \in$ S),
x_2 – Early leavers from education and training (%, $x_2 \in$ D),
x_3 – Tertiary educational attainment of persons aged 30–34 (%; $x_3 \in$ S),
x_4 – Greenhouse gas emissions (*1990 = 100*; $x_4 \in$ D),
x_5 – Share of renewables in gross final energy consumption (%; $x_5 \in$ S),
x_6 – Consumption of primary energy (*kg of oil equivalent per 1 000 EUR of GDP*; $x_6 \in$ D),
x_7 – Employment rate of persons aged 20–64 (%; $x_7 \in$ S),
x_8 – Share of people at risk-of-poverty or social exclusion (%; $x_8 \in$ D),
x_9 – People living in households with very low work intensity (%; $x_9 \in$ D),
x_{10} – People at risk-of-poverty rate (after social transfers) (%; $x_{10} \in$ D),
x_{11} – Severely materially deprived people (%; $x_{11} \in$ D).

The pattern normalization is carried out for two years: 2010 and 2015, as well as in two environments: the entire European Union (abbr. EU28) and 10 countries that joined the EU in 2004 (abbr. EU10). Table 1 and Table 2 show the characteristics of indicators before and after normalization.

Table 1 Characteristics of indicators before (abbr. raw) and after (abbr. norm) normalization in both environments (reference groups) EU28 and EU10 – year 2010

Indicator		Reference group	Max	Min	Mean	Standard deviation	Skewness	Kurtosis
$x_1 \in$ S	raw	EU28	3.730	0.450	1.514	0.893	0.759	–0.285
	norm		0.053	0.000	0.036	0.014	–0.759	–0.285
	raw	EU10	2.060	0.450	0.991	0.497	0.913	–0.389
	norm		0.151	0.000	0.100	0.047	–0.913	–0.389

Table 1 (continuation)

Indicator		Reference group	Max	Min	Mean	Standard deviation	Skewness	Kurtosis
$x_2 \in \mathrm{D}$	raw	EU28	28.300	4.700	12.168	6.306	1.176	0.878
	norm		0.113	0.000	0.036	0.030	1.176	0.878
	raw	EU10	23.800	4.700	9.910	5.589	1.286	1.073
	norm		0.367	0.000	0.100	0.107	1.286	1.073
$x_3 \in \mathrm{S}$	raw	EU28	50.100	18.300	34.325	9.894	–0.135	–1.506
	norm		0.072	0.000	0.036	0.022	0.135	–1.506
	raw	EU10	45.300	20.400	32.220	8.743	0.057	–1.403
	norm		0.190	0.000	0.100	0.067	–0.057	–1.403
$x_4 \in \mathrm{D}$	raw	EU28	163.770	43.200	90.637	27.493	0.271	0.072
	norm		0.091	0.000	0.036	0.021	0.271	0.072
	raw	EU10	163.770	43.200	83.126	36.835	0.953	–0.185
	norm		0.302	0.000	0.100	0.092	0.953	–0.185
$x_5 \in \mathrm{S}$	raw	EU28	47.200	1.000	15.857	10.765	0.889	0.485
	norm		0.053	0.000	0.036	0.012	–0.889	0.485
	raw	EU10	30.400	1.000	14.370	8.632	0.348	–0.882
	norm		0.183	0.000	0.100	0.054	–0.348	–0.882
$x_6 \in \mathrm{D}$	raw	EU28	464.900	82.400	191.943	93.795	1.269	1.134
	norm		0.125	0.000	0.036	0.031	1.269	1.134
	raw	EU10	417.900	142.000	250.140	75.287	0.512	0.340
	norm		0.255	0.000	0.100	0.070	0.512	0.340
$x_7 \in \mathrm{S}$	raw	EU28	78.100	59.900	68.136	5.334	0.239	–1.188
	norm		0.065	0.000	0.036	0.019	–0.239	–1.188
	raw	EU10	75.000	59.900	66.000	4.496	0.489	–0.576
	norm		0.168	0.000	0.100	0.050	–0.489	–0.576
$x_8 \in \mathrm{D}$	raw	EU28	49.200	14.400	24.575	8.202	1.222	1.221
	norm		0.122	0.000	0.036	0.029	1.222	1.221
	raw	EU10	38.200	14.400	25.070	6.994	0.409	–0.806
	norm		0.223	0.000	0.100	0.066	0.409	–0.806

Table 1 (continuation)

Indicator		Reference group	Max	Min	Mean	Standard deviation	Skewness	Kurtosis
$x_9 \in \mathrm{D}$	raw	EU28	22.900	4.900	9.657	3.442	1.913	5.314
	norm		0.135	0.000	0.036	0.026	1.913	5.314
	raw	EU10	12.600	4.900	8.570	2.269	0.322	–0.736
	norm		0.210	0.000	0.100	0.062	0.322	–0.736
$x_{10} \in \mathrm{D}$	raw	EU28	21.600	9.000	15.957	3.455	0.045	–0.981
	norm		0.065	0.000	0.036	0.018	0.045	–0.981
	raw	EU10	20.900	9.000	15.190	3.610	0.087	–0.896
	norm		0.192	0.000	0.100	0.058	0.087	–0.896
$x_{11} \in \mathrm{D}$	raw	EU28	45.700	0.500	10.621	10.038	1.879	3.383
	norm		0.159	0.000	0.036	0.035	1.879	3.383
	raw	EU10	27.600	5.900	13.350	7.040	0.727	–0.721
	norm		0.291	0.000	0.100	0.094	0.727	–0.721

Note: After pattern normalization, the minimum value is always zero, while the mean is always 1/*n*, i.e. 0.036 for EU28 and 0.1 for EU10 (compare proprieties 2.2.2, 2.3.1).
Source: Own calculation

Table 2 Characteristics of indicators before (abbr. raw) and after (abbr. norm) normalization in both environments (reference groups) EU28 and EU10 – year 2015

Indicator		Reference group	Max	Min	Mean	Standard deviation	Skewness	Kurtosis
$x_1 \in \mathrm{S}$	raw	EU28	3.270	0.480	1.610	0.823	0.606	–0.864
	norm		0.060	0.000	0.036	0.018	–0.606	–0.864
	raw	EU10	2.200	0.480	1.208	0.523	0.490	–0.757
	norm		0.173	0.000	0.100	0.053	–0.490	–0.757
$x_2 \in \mathrm{D}$	raw	EU28	20.000	2.700	9.821	4.397	0.919	0.236
	norm		0.087	0.000	0.036	0.022	0.919	0.236
	raw	EU10	19.800	5.000	8.760	4.500	1.331	0.837
	norm		0.394	0.000	0.100	0.120	1.331	0.837
$x_3 \in \mathrm{S}$	raw	EU28	57.600	25.300	40.496	9.077	–0.066	–1.069
	norm		0.067	0.000	0.036	0.019	0.066	–1.069
	raw	EU10	57.600	27.800	40.610	9.915	0.246	–1.095

Table 2 (continuation)

Indicator		Reference group	Max	Min	Mean	Standard deviation	Skewness	Kurtosis
$x_3 \in S$	norm	EU10	0.175	0.000	0.100	0.058	–0.246	–1.095
$x_4 \in D$	raw	EU28	144.450	41.990	80.583	24.071	0.489	0.135
	norm		0.095	0.000	0.036	0.022	0.489	0.135
	raw	EU10	144.450	41.990	73.372	30.371	1.057	0.356
	norm		0.326	0.000	0.100	0.097	1.057	0.356
$x_5 \in S$	raw	EU28	53.900	5.000	19.811	11.697	0.944	0.560
	norm		0.051	0.000	0.036	0.012	–0.944	0.560
	raw	EU10	37.600	5.000	18.270	9.491	0.615	–0.603
	norm		0.169	0.000	0.100	0.049	–0.615	–0.603
$x_6 \in D$	raw	EU28	448.500	62.000	165.468	85.366	1.538	2.563
	norm		0.133	0.000	0.036	0.029	1.538	2.563
	raw	EU10	358.000	90.500	208.490	67.861	0.400	0.482
	norm		0.227	0.000	0.100	0.058	0.400	0.482
$x_7 \in S$	raw	EU28	80.500	54.900	69.936	5.796	–0.469	0.056
	norm		0.087	0.000	0.036	0.020	0.469	0.056
	raw	EU10	76.500	67.700	70.630	3.160	0.636	–1.170
	norm		0.150	0.000	0.100	0.054	–0.636	–1.170
$x_8 \in D$	raw	EU28	41.300	14.000	24.318	6.743	0.675	–0.204
	norm		0.094	0.000	0.036	0.023	0.675	–0.204
	raw	EU10	30.900	14.000	23.890	5.240	–0.369	–0.983
	norm		0.171	0.000	0.100	0.053	–0.369	–0.983
$x_9 \in D$	raw	EU28	19.200	5.700	10.343	3.256	0.945	0.343
	norm		0.104	0.000	0.036	0.025	0.945	0.343
	raw	EU10	10.900	6.600	8.130	1.375	0.633	–0.861
	norm		0.281	0.000	0.100	0.090	0.633	–0.861
$x_{10} \in D$	raw	EU28	25.400	9.700	17.061	3.940	0.208	–0.876
	norm		0.076	0.000	0.036	0.019	0.208	–0.876

Table 2 (continuation)

Indicator		Reference group	Max	Min	Mean	Standard deviation	Skewness	Kurtosis
$x_{10} \in D$	raw	EU10	22.500	9.700	16.760	4.080	0.003	–1.037
	norm		0.181	0.000	0.100	0.058	0.003	–1.037
$x_{11} \in D$	raw	EU28	45.700	0.500	10.621	10.038	1.879	3.383
	norm		0.159	0.000	0.036	0.035	1.879	3.383
	raw	EU10	27.600	5.900	13.350	7.040	0.727	–0.721
	norm		0.291	0.000	0.100	0.094	0.727	–0.721

Note: After pattern normalization, the minimum value is always zero, while the mean is always 1/n, i.e. 0.036 for EU28 and 0.1 for EU10 (compare proprieties 2.2.2, 2.3.1).
Source: Own calculation

Poland is selected as an example. Table 3 compares the results of normalization for both years. An influence of normalization itself and normalization environment on the dynamic image of the country are examined. That is, for a given object (Poland) we analyze what happens to the normalized value of indicator if the raw value improves (or gets worse). These aspects are important when comparing the pattern normalization with scaling (min-max normalization). Scaling, the most popular method of dynamic normalization, in this context can be called neutral. It does not affect the dynamic image of objects, moreover, the environment of scaling does not matter.

Table 3 Raw and normalized indicators for Poland – changes over time

Indicator		2010	2015	2015 to 2010
$x_1 \in S$	raw	0.720	1.000	+
	EU28	0.049	0.049	–
	EU10	0.125	0.121	+
	rank	5	4	+
$x_2 \in D$	raw	5.400	5.300	+
	EU28	0.003	0.013	–
	EU10	0.013	0.008	+
	rank	4	3	+
$x_3 \in S$	raw	34.800	43.400	+
	EU28	0.035	0.030	+
	EU10	0.080	0.084	–
	rank	6	6	0

Table 3 (continuation)

Indicator		2010	2015	2015 to 2010
$x_4 \in \mathrm{D}$	raw	87.170	82.760	+
	EU28	0.033	0.038	–
	EU10	0.110	0.130	–
	rank	7	7	0
$x_5 \in \mathrm{S}$	raw	9.300	11.800	+
	EU28	0.043	0.044	–
	EU10	0.132	0.133	–
	rank	4	3	+
$x_6 \in \mathrm{D}$	raw	278.300	227.300	+
	EU28	0.064	0.057	+
	EU10	0.126	0.116	+
	rank	8	8	0
$x_7 \in \mathrm{S}$	raw	64.300	67.800	+
	EU28	0.049	0.043	+
	EU10	0.119	0.148	–
	rank	3	2	+
$x_8 \in \mathrm{D}$	raw	27.800	23.400	+
	EU28	0.047	0.033	+
	EU10	0.126	0.095	+
	rank	7	5	+
$x_9 \in \mathrm{D}$	raw	7.300	6.900	+
	EU28	0.018	0.009	+
	EU10	0.065	0.020	+
	rank	4	3	+
$x_{10} \in \mathrm{D}$	raw	17.600	17.600	0
	EU28	0.044	0.038	+
	EU10	0.139	0.112	+

Table 3 (continuation)

Indicator		**2010**	**2015**	**2015 to 2010**
$x_{10} \in D$	rank	8	7	+
$x_{11} \in D$	raw	14.200	8.100	+
	EU28	0.048	0.029	+
	EU10	0.111	0.059	+
	rank	7	4	+

Note: + improvement, – deterioration, 0 no changes.
Source: Own calculation

In the analyzed period in Poland, raw values of all indicators except x_{10} improve, i.e. values of stimulants increase, values of destimulants decrease. Changes would be the same after dynamic scaling, but after pattern normalization the changes over time are not so uniform.

From this point of view, the indicators monitoring the implementation of the Europe 2020 Strategy can be divided into three groups. In the first one there are x_6, x_8, x_9, x_{11}. In their case, pattern normalization does not change the dynamics of variables. Raw indicators are improved as well as normalized indicators (for both environments).

The second group are indicators for which normalization changes the "dynamic image" of Poland, but the normalization environment is irrelevant. This group includes x_{10}. The raw value of this indicator does not change, but after normalization it improves in both environments. This means that Poland's objective situation has not improved, but the relative one has (because the situation of other countries in this period has deteriorated). The next in this group are x_4 and x_5. For them, the impact of normalization is more evident. Although the raw values of these indicators improve, the situation of Poland in both considered environments has got worse. All three indicators after normalization increase.

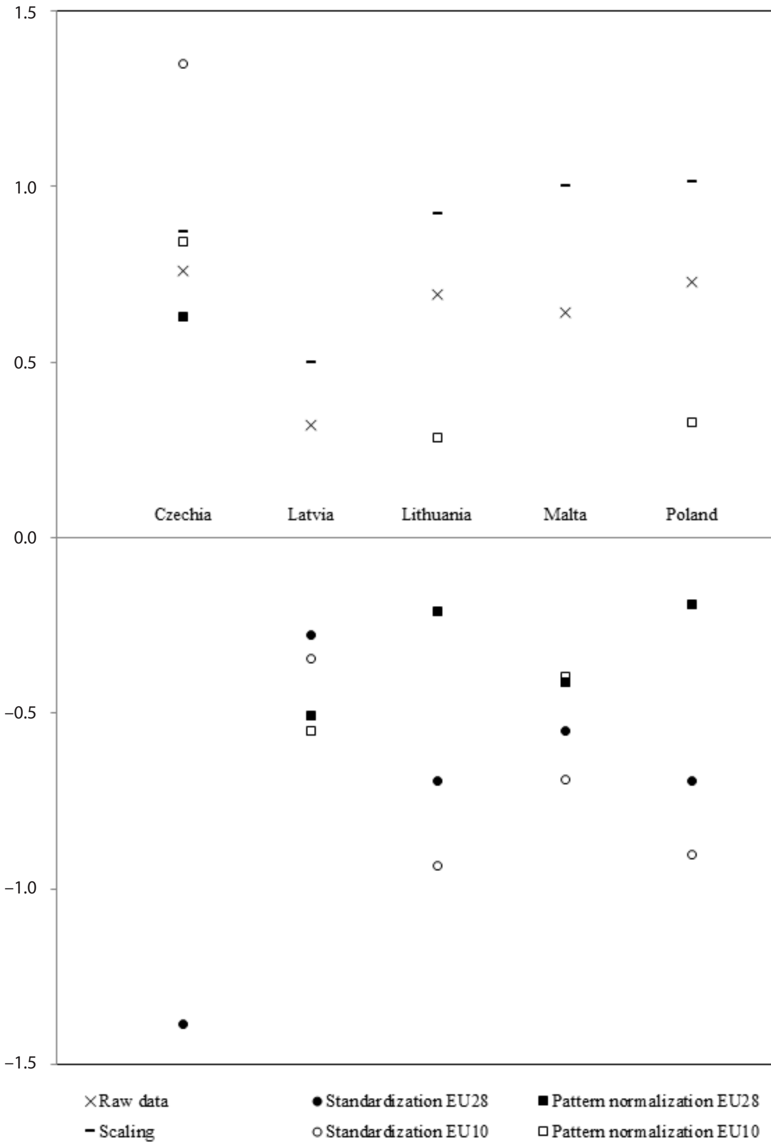
The last group are indicators for which the normalization environment is important. For x_3 and x_7, Poland has improved against the background of a bigger environment, and it has declined against a smaller one. For x_1, x_2 the situation is reversed.

It is interesting to confront the above considerations with the analysis of changes in Poland's position in the ranking. For some indicators the relative improvement is so great that the position of Poland in the ranking also improves (e.g. x_8), however the relative change can be insufficient to improve the position (e.g. x_6). There is also a situation (x_5) in which the direction of changes in the normalized variables and positions is reversed.

Next, the pattern normalization is compared with the most popular methods of normalization: standardization (17) and scaling (18). The direction of changes in the values of the normalized variable in 2015 as compared to 2010 is analyzed. Dynamic standardization and dynamic scaling is performed based on data from both years. For selected countries Figure 1 shows a relative increase in the value of variable $x_1 \in S$ in 6 versions: without normalization, after scaling (in this case, the reference group does not matter), after standardization and pattern normalization for both E28 and E10 environments.

For all presented countries, the raw values of the variable x_1 increase, and thus its values after scaling increase as well. This differs the scaling from the pattern normalization and the standardization. For the last two normalizations the direction of changes in transformed values does not necessarily coincide with the direction of changes in raw data. Moreover, for a certain country, the normalized value for one reference group may increase, and for another group it may decrease.

Figure 1 Relative increments of variable x_1 before and after normalization



Note: To make the graph more transparent, the increments are transformed with the cube root. If a country is located above the axis, its situation in 2015 improved compared to 2010, that is, the value of the stimulants increased (x, z, s), and the value of the destimulant decreased (u).
Source: Own calculation

CONCLUSIONS

The article presents a new transformation of diagnostic variables, that plays a double role in analyses of complex phenomenon: it unifies the nature of variables and makes variables comparable. The transformation is called normalization with respect to the pattern (or pattern normalization in short). The pattern normalization has properties expected for this type of transformation.

The values of variables after normalization with respect to pattern characterize the relative situation of the objects, i.e. the situation on the background of the environment in which the research is carried out. Changing the environment can change the research results. This feature is both an advantage and the biggest disadvantage of the proposed method. The pattern normalization can only be used in research in which the context of the environment is important. "Objective" changes may be distorted during this transformation.

A main advantage of new normalization is the possibility of use in dynamic analysis (i.e. for different time units). However, it is not necessary to re-calculate results with the appearance of observations for next unit of time, as, for example, in the case of dynamic standardization.

The effects of normalization with respect to pattern are illustrated by an empirical example. Indicators monitoring the implementation of the Europe 2020 Strategy are normalized. Normalizations are carried out for two environments: the entire EU and countries that joined the EU in 2004. The results for two years are compared. The example of Poland shows that the "dynamic image" of the country is affected by the use of normalization itself as well as by the choice of the environment in normalization.

Pattern normalization can be used in common construction of composite indicators instead of other methods of normalization. A possible applications are shown in Müller-Frączek (2017, 2018).

The proposed construction can have various modifications. First of all, we can change the mass distributed between objects (for example to *n*). We can also change the measure of distance or the method of choosing the pattern.

References

EUROPEAN COMMISSION. *Europe 2020. A strategy for smart, sustainable and inclusive growth* [online]. Brussels: European Commission, 2010. [cit. 20.12.2018] <https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2010:2020:FIN:EN:PDF>.

HELLWIG, Z. Zastosowanie metody taksonomicznej do typologicznego podziału krajów ze względu na poziom ich rozwoju i strukturę kwalifikowanych kadr. *Przeglęad Statystyczny*, 1968, 15(4), pp. 307–327.

JAJUGA, K. AND WALESIAK, M. Standardisation of Data Set under Different Measurement Scales. In: DECKER, W. AND GAUL, W. eds. *Classification and Information Processing at the Turn of the Millennium*, Berlin, Heidelberg: Springer, 2000.

MAZZIOTTA, M. AND PARETO, A. Comparing Two Non-Compensatory Composite Indices to Measure Changes over Time: a Case Study [online]. *Statistika: Statistics and Economy Journal*, 2015, 95(2), pp. 44–53.

MAZZIOTTA, M., PARETO, A. On a generalized non-compensatory composite index for measuring socio-economic phenomena. *Social Indicators Research*, 2016, 127(3), pp. 983–1003.

MILLIGAN, G. AND COOPER, M. A study of standardization of variables in cluster analysis. *Journal of Classification*, 1988, 5(2), pp. 181–204.

MŁODAK, A. Multilateral normalizations of diagnostic features. *Statistics in Transition*, 2006, 7(5), pp. 1125–1139.

MÜLLER-FRĄCZEK, I. Propozycja miary syntetycznej. *Przegląd Statystyczny*, 2017, 64(4), pp. 413–428.

MÜLLER-FRĄCZEK, I. Dynamic measure of development. In: PAPIEŻ, M. AND ŚMIECH, S. eds. *The 12th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena. Conference Proceedings*, Cracow: Foundation of the Cracow University of Economics, 2018.

NARDO, M., SAISANA, M., SALTELLI, A., TARANTOLA, S., HOFFMAN, A., GIOVANNINI, E. *Handbook on Constructing Composite Indicators.* OECD publishing, 2005.

SAISANA, M. AND SALTELLI, A. Rankings and Ratings: Instructions for use. *Hague Journal on the Rule of Law*, 2011, 3.2, pp. 247–268.

STATISTICS POLAND. *Europe 2020 indicators* [online]. Warsaw, 2018. [cit. 20.12.2018] <https://stat.gov.pl/en/international-statistics/international-comparisons/tables-about-countries-by-subject/europe-2020-indicators>.

ANNEX

Proof of (3):

$$u_i = 0 \Leftrightarrow \frac{|x_i - x^+|}{\sum_{j=1}^{n}|x_j - x^+|} = 0 \Leftrightarrow |x_i - x^+| = 0 \Leftrightarrow x_i = x^+.$$

Proof of (5):

$$u_i = 1 \Leftrightarrow \frac{|x_i - x^+|}{\sum_{j=1}^{n}|x_j - x^+|} = 1 \Leftrightarrow |x_i - x^+| = \sum_{j=1}^{n} |x_j - x^+| \Leftrightarrow \forall_{j \neq i}\ x_j = x^+.$$

Proof of (6):

$$\text{If } x \in S\text{, then: } \max_i u = \frac{\min_i(x^+ - x_i)}{\sum_{j=1}^{n}(x^+ - x_j)} = \frac{x^+ - \min_i x_i}{\sum_{j=1}^{n}(x^+ = x_j)} = \frac{\max_i x_i - \min_i x_i}{\sum_{j=1}^{n}(\max_i x_i - x_j)}.$$

$$\text{If } x \in D\text{, then: } \max_i u = \frac{\max_i(x_i - x^+)}{\sum_{j=1}^{n}(x_j - x^+)} = \frac{\max_i x_i - x^+}{\sum_{j=1}^{n}(x_j - x^+)} = \frac{\max_i x_i - \min_i x_i}{\sum_{j=1}^{n}(x_j - \max_i x_i)}.$$

Proof of (7):

$$\bar{u} = \frac{1}{n} \sum_{i=1}^{n} \frac{|x_i - x^+|}{\sum_{j=1}^{n}|x_j - x^+|} = \frac{1}{n} \frac{\sum_{i=1}^{n}|x_i - x^+|}{\sum_{j=1}^{n}|x_j - x^+|} = \frac{1}{n}.$$

Proof of (8): Assume that $x \in S$, but the proof is similar when $x \in D$.

$$S^2(u) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x^+ - x_i}{\sum_{j=1}^{n}(x^+ - x_j)} - \frac{1}{n} \right)^2 = \frac{1}{n^3} \sum_{i=1}^{n} \left(\frac{x^+ - x_i}{x^+ - \frac{1}{n}\sum_{j=1}^{n} x_j} - 1 \right)^2 = \frac{1}{n^3} \sum_{i=1}^{n} \left(\frac{x^+ - x_i}{x^+ - \bar{x}} - 1 \right)^2$$

$$= \frac{1}{n^3} \sum_{i=1}^{n} \left(\frac{\bar{x} - x_i}{x^+ - \bar{x}} \right)^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(\bar{x} - x_i)^2}{n^2(x^+ - \bar{x})^2} = \frac{S^2(x)}{n^2(x^+ - x)^2}.$$

Proof of (11): Assume that $x \in S$, but the proof is similar when $x \in D$.

$$\mu_3(u) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x^+ - x_i}{\sum_{j=1}^{n}(x^+ - x_j)} - \frac{1}{n} \right)^3 = \frac{1}{n^4} \sum_{i=1}^{n} \left(\frac{x^+ - x_i}{x^+ - \frac{1}{n}\sum_{j=1}^{n} x_j} - 1 \right)^3$$

$$= \frac{1}{n^4} \sum_{i=1}^{n} \left(\frac{x^+ - x_i}{x^+ - \bar{x}} - 1 \right)^3 = \frac{1}{n^4} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\bar{x} - x^+} \right)^3 = \frac{\mu_3(x)}{n^3(\bar{x} - x^+)^3}.$$

Proof of (13): Assume that $x \in S$, but the proof is similar when $x \in D$.

$$\mu_4(u) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x^+ - x_i}{\sum_{j=1}^{n}(x^+ - x_j)} - \frac{1}{n} \right)^4 = \frac{1}{n^5} \sum_{i=1}^{n} \left(\frac{x^+ - x_i}{x^+ - \frac{1}{n}\sum_{j=1}^{n} x_j} - 1 \right)^4$$

$$= \frac{1}{n^5} \sum_{i=1}^{n} \left(\frac{x^+ - x_i}{x^+ - \bar{x}} - 1 \right)^4 = \frac{1}{n^5} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{x^+ - \bar{x}_i} \right)^4 = \frac{\mu_4(x)}{n^4 (x^+ - \bar{x})^4} .$$

Proof of (15): Assume that x_1 and x_2 are stimulants. The proof in other cases is similar.

$$\text{cov}(u_1, u_2) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x^+ - x_{1i}}{\sum_{j=1}^{n}(x_1^+ - x_{1j})} - \frac{1}{n} \right) \left(\frac{x_2^+ - x_{2i}}{\sum_{j=1}^{n}(x_2^+ - x_{2j})} - \frac{1}{n} \right)$$

$$= \frac{1}{n^3} \sum_{i=1}^{n} \left(\frac{x_1^+ - x_{1i}}{x_1^+ - \frac{1}{n}\sum_{j=1}^{n} x_{1i}} - 1 \right) \left(\frac{x_2^+ - x_{2i}}{x_2^+ - \frac{1}{n}\sum_{j=1}^{n} x_{2i}} - 1 \right)$$

$$= \frac{1}{n^3} \sum_{i=1}^{n} \left(\frac{\overline{x}_1 - x_{1i}}{x_1^+ - \overline{x}_1} \cdot \frac{\overline{x}_2 - x_{2i}}{x_2^+ - \overline{x}_2} \right) = \frac{\frac{1}{n} \sum_{i=1}^{n} (\overline{x}_1 - x_{1i})(\overline{x}_2 - x_{2i})}{n^2 (x_1^+ - \overline{x}_1)(x_2^+ - \overline{x}_2)} = \frac{\text{cov}(x_1, x_2)}{n^2 (x_1^+ - \overline{x}_1)(x_2^+ - \overline{x}_2)} .$$

Empirically Supported Methodological Critique of Double Entry in Dyadic Data Analysis

Imre Dobos[1] | *Budapest University of Technology and Economics, Budapest, Hungary*
Andrea Gelei[2] | *Corvinus University of Budapest, Budapest, Hungary*

Abstract

Analyzing dyadic phenomena (e.g. trust, power, and satisfaction) gains importance not only in sociology and psychology, but also in economics and management. The aim of the paper is to examine the mathematical foundation of Dyadic Data Analysis (DDA). On one hand, we critique the database development of DDA for exchangeable cases, and develop an algorithm for transforming such a data set into distinguishable cases. On the other hand, we question the usefulness of a widely used data development technique of DDA, the so-called double entry. We reason that this technique does not necessarily lead to additional information. In contrast, it might lead to information losses. We develop approximations for correlations and regression models of DDA. These are also empirically tested using a database of 89 dyads. The obtained results back our theoretical reasoning, most of the approximations give satisfying results. This support our main proposition that mathematical foundation of DDA needs further research.

Keywords	***JEL code***
Correlational analysis, regression analysis, dyadic data analysis, double entry technique	*C10, C39, C49*

INTRODUCTION

The problem of analyzing dyadic data is well known from paired samples. The basic question is whether or not a given variable in two dependent samples has the same shape of distribution, expected value, and standard deviation. These questions are important but are also supplemented by new research challenges, since researchers in sociology, psychology, economics and management are increasingly interested in complex research issues that make it necessary to apply multivariate analytical techniques in dyadic settings (Kenny et al., 2006). The traditional technique of paired sample analysis is inappropriate for answering such research questions. (e.g. How the level of perceived trust of the partners in a business relationship influences the partners' willingness to take risk in joint future innovation projects.) Instead,

[1] Budapest University of Technology and Economics, Magyar Tudósok Körútja 2, 1111 Budapest, Hungary. Corresponding author: email: dobos@kgt.bme.hu.

[2] Corvinus University of Economics, Fővám tér 8, 1093 Budapest, Hungary.

the use of dyadic data analysis (DDA) is suggested (Griffin and Gonzalez, 1995; Gonzalez and Griffin, 1999; 2000; Gonzales, 2010; Kenny et al., 2006; Burk et al., 2007; Kenny, 2015). According to literature, the processing of paired samples, also called dyadic databases, using traditional statistical methods may lead to a number of error types[3] (Gonzalez and Griffin, 2000). Dyadic data analysis is a specific, interrelated set of statistical techniques that aim at overcoming these errors.

Moreover, let us point out limitation of traditional inductive statistics, namely it assumes the representativeness of the sample. In an explicit form, this usually appears in a way that data observed can be regarded as the independent sample with identical distribution, or it can be assumed that it approaches identical distribution because of the small sample and different placements of weight. In the examination of relational trust and similar social problems, the representativeness of the sample is out of the question. Often, the population is not known by the researcher, the respondents are the ones who just participate in the given study, which means that the analysis is basically descriptive. In this case, inductive statistics makes little sense. The essence of the dyadic approach is that it regards each relationship unique and intends to put the consequences of the unique context in the center of analysis. Hence, this approach does not pose any requirements about generalization regarding total population either (Gelei and Sugár, 2017).

Previously dyadic data analysis to a trust-related management problem has already been applied (Gelei and Dobos, 2016). Later, DDA and the classical statistical techniques have been compared (Gelei and Sugár, 2017). This comparison concluded, despite the substantial methodological differences between classical statistics and DDA, that the empirical results were not significantly different. This finding has motivated us to look into the mathematical fundamentals of dyadic data analysis. The results of this investigation make up this methodological article. We discuss key concepts of dyadic data analysis, focusing on the suggested database development technique, called double entry (Gonzalez and Griffin, 2000; Ledermann and Kenny, 2015). We discuss the so-called exchangeable case, the related homogeneity analysis, the core correlations of DDA and its regression equations (Gonzalez and Griffin, 1995, 1999; Ledermann et al., 2011). These fundamentals are relevant for more elaborate and complex analytical techniques, such as the curve-of-factors model (McArdle, 1988; Whittaker et al., 2014), structural equation modeling (Peugh et al., 2013; Deng and Yan, 2015), and situations dealing with longitudinal dyadic data (Planalp et al., 2017). Our objective is to critically discuss this relatively new statistical methodology.

In dyadic data analysis, the first analytical step is the so-called homogeneity analysis. Here, one deals with the problem of assessing interdependence in a dyad for a single variable (Gonzalez and Griffin, 2000). The key question is whether the informants in a dyad have symmetric or asymmetric positions (e.g., physician and patient). First, we talk about exchangeable cases, in contrast to distinguishable ones. The homogeneity analysis is different from the classical analysis, in which the core issue is to evaluate the similarity of the distributions of the variables in two databases. Instead of using the ANOVA framework, DDA suggests applying a technique for database development called double entry (Gonzalez and Griffin, 2000). This technique has crucial importance not only for DDA but also for our critique. Therefore, in the following sections, we discuss basic concepts and techniques of dyadic data analysis, including double entry and homogeneity analysis. As a next step, we attempt to refine the concept of dyadic correlations and calculate them using the initial, raw database, which does not necessitate the use of the double-entry technique. (This database reflects the timely development of pairwise sampling; the first pair in the survey is fixed in the database as the first dyad, and so on.) Finally, dyadic regression models are investigated. We conclude that the suggested technique of double entry and the statistical constructs using these models do not necessarily lead to additional information. In contrast, these techniques might lead to information

[3] These error types are the following: (1) error of assumed independence; (2) data omission error; (3) error between levels; and (4) error of the levels of analysis (Gonzalez and Griffin, 2000).

losses. We develop statistical approximations for these dyadic constructs by applying classical statistics on the initial, raw database.

Each of the suggested statistical constructs are tested using a database that has been developed in one of our previous field studies using pairwise sampling. The research hypothesis of this previous study was as follows: In a business relationship characterized by mutually high levels of trustworthiness perceived by the counterparts, the willingness to be involved in risky situations is higher than in relationships in which actors do not mutually believe that their partners are highly trustworthy. In order to test our hypothesis, we developed a questionnaire, where respondents had to answer the followings:

- Evaluate the perceived levels of trustworthiness of their actual pair (1–7 Likert scale);
- Evaluate the level of different information sharing situations listed (ranking);
- Trust in the relationship: the willingness to share specific information with the actual partner in the pair (yes = 1 or no = 0).

We organized workshops for purchasing and logistics managers, where theyir formed concrete pairs and filled out the questionnaire. Data gathering was so carried out in the physical presence of respondents, but in an anonym way. Concrete answers were neither visible nor accessible to the participants in order to avoid biases in responses. We gathered 89 pairs of questionnaires, with 178 dyadic data points. A more detailed description of the field study and its dataset development is presented in the work of Gelei and Dobos (2016). For this article we used the variable of perceived levels of trustworthiness for calculating the newly developed correlation constructions based on the initial, raw dataset. For testing our suggested regression models we used trust as the dependent variable while independent variables were the perceived levels of trustworthiness in the pairs. We used SPSS 22 and Microsoft Excel throughout this article for statistical calculations.

The results of our empirical test show that in most cases, the suggested approximations can give really good results and support our suggestion not to use the difficult technique of double entry and the statistical constructs based on it.

1 FUNDAMENTALS OF THE CRITIQUE OF DYADIC DATA ANALYSIS

These form the analytical unit for statistical analysis. A very simple question arises, when developing dyadic datasets from such data pairs: In what order to fix the two answers of a pair? In a distinguishable case it is obvious since positions in any pair are given (e.g. doctor and patient). An initial or raw database is shown in Table 1.

Table 1 Dyadic data analysis with three dyads in the database

Variables **Observations**	***1. variable (X)***	
	1. data (X_1)	***2. data (X_2)***
1. dyad	x_{11}	x_{12}
2. dyad	x_{21}	x_{22}
3. dyad	x_{31}	x_{32}

Source: Own construction

In the so-called exchangeable case however these positions are not predefined, and can change. In such cases *n* number of such data pairs can lead to a number of 2^n number databases. In case data pairs are interpreted as paired sample, different datasets can lead to different results during analysis. Therefore, the question arises, which one should be use? This is the first problem we investigate. We suggest a method which transforms any exchangeable data set into a distinguishable one. As a next step another issue related to dyadic dataset development is discussed, the so-called double entry. Our focal problem is, whether this doubling leads to any information surplus or not.

1.1 Issues of data set development in DDA

A key innovation in dyadic data analysis is the double entry of data obtained through field research using pairwise sampling. The essential idea is to create two vectors from all the aligned data pairs by changing the order in which the data are entered into the database. Changing this order creates two variables from one. The original and the newly created variables are denoted as *X* and *X'*; see the example in Table 2. This table shows that the number of observations belonging to variables *X* and *X'* is twice the number of dyads, which is the number of pairs in the database. Dyadic data analysis requires this transformation to create and use vectors instead of matrices (tables) for further statistical analysis.

Table 2 Symbolic representation for double entry and the pairwise data setup

Observations	***Variables***	
	X	***X'***
1. pair (initial order)	x_{11}	x_{12}
1. pair (changed order)	x_{12}	x_{11}
2. pair (initial order)	x_{21}	x_{22}
2. pair (changed order)	x_{22}	x_{21}
3. pair (initial order)	x_{31}	x_{32}
3. pair (changed order)	x_{32}	x_{31}
4. pair (initial order)	x_{41}	x_{42}
4. pair (changed order)	x_{42}	x_{41}

Source: Own construction

1.2 The so-called exchangeable case and homogeneity analysis

In DDA, there are two types of analytical situations called cases, including the exchangeable and the distinguishable cases (Gonzalez and Griffin, 2000). In the exchangeable case, the informants in a given dyad (or pair) cannot be distinguished in advance, in contrast to the distinguishable case, in which the informants in any given dyad have specific systemic characteristics or positions that are known well in advance of the analysis (e.g., one informant in a pair is the husband, the other is the wife). In this article, the analysis starts with the exchangeable case.

As mentioned previously, in the distinguishable case, the two people in a pair are in asymmetric positions, in contrast to the exchangeable case, in which the positions of the two informants in any pair are symmetric, i.e., they are identical. Suppose we have three dyads or pairs in the database, as shown in Table 1. This table reflects the sequential data collection in field research: the first dyad (or pair) was the first one questioned, the second dyad was questioned next, and so on.

Since we have exchangeable cases, we can transform this initial database by simply changing the order of the data related to a given variable in a dyad.

Table 3 Changing the order of the data related to a given variable in a dyad to develop a new database for dyadic data analysis

Variables **Observations**	***1. variable (X)***	
	1. data (X'_1)	***2. data (X'_2)***
1. dyad	x_{12}	x_{11}
2. dyad	x_{21}	x_{22}
3. dyad	x_{31}	x_{32}

Source: Own construction

Since we have three dyads, this process could be continued an additional six time periods, leading to $2^3 = 8$ potential databases. Generally, we can state that having *n* dyads for analysis offers 2^n slightly different databases.

Table 4 Potential databases with three dyads in the survey

	Database 1	**Database 2**	**Database 3**	**Database 4**	**Database 5**	**Database 6**	**Database 7**	**Database 8**
1. dyad	(x_{11}, x_{12})	(x_{11}, x_{12})	(x_{11}, x_{12})	(x_{11}, x_{12})	(x_{12}, x_{11})	(x_{12}, x_{11})	(x_{12}, x_{11})	(x_{12}, x_{11})
2. dyad	(x_{21}, x_{22})	(x_{21}, x_{22})	(x_{22}, x_{21})	(x_{22}, x_{21})	(x_{21}, x_{22})	(x_{21}, x_{22})	(x_{22}, x_{21})	(x_{22}, x_{21})
3. dyad	(x_{31}, x_{32})	(x_{32}, x_{31})	(x_{31}, x_{32})	(x_{32}, x_{31})	(x_{31}, x_{32})	(x_{32}, x_{31})	(x_{31}, x_{32})	(x_{32}, x_{31})

Source: Own construction

To obtain valid and reliable results, the statistical analysis applied to dyadic data must be unaffected by the choice of database. Let us test this first! For this purpose, we used the database with 89 purchasing and logistics manager dyads.

We chose the first dataset for our calculation randomly from all the potential databases, while the second was developed from this initial one by carrying out the following change systematically: the data with lower value in any given dyad/pair were recorded systematically in data position 1 in the dyad. Since we can assume that the two datasets are interdependent, we applied a test developed for paired samples, namely, the *t*-test. Table 5 shows that the results obtained using the two databases are significantly different. The first database led to the acceptance of the null hypothesis; the means of the two informants of the pairs in the given database do not differ significantly. In contrast, using the second, modified database resulted in the rejection of this null hypothesis. The objective was to highlight the problem related to the choice of database which might lead markedly different results.

So, the order of the data in the databases might really pose problems in exchangeable cases. According to dyadic data analysis, a potential solution to this challenge could be the technique of double entry. However, this solution does not really solve the problem; it only doubles the size of the database. As discussed above, the number of potential databases is 2^n, since we have *n* dyads available for analysis. Any statistical method applied for analyzing dyadic data must be completely independent from the order

Table 5 Testing the means of the two databases

	Paired differences					t-test	Freedom	Level of significance (two-tailed)
	Mean	Standard deviation	Standard error	Confidence interval (95%) for the differences				
				Lower	Upper			
Database 1	0.07865	1.79788	0.19058	−0.30008	0.45738	0.413	88	0.681
Database 2	1.13483	1.39146	0.14749	0.84172	1.42795	7.694	88	0.000

Source: Own construction

of the data in the database, i.e. the database is transformed into a distinguishable one. Let us note that such methods can rely on a technique that operates on the absolute values of the sum and/or difference of the data in a given dyad; these are the same for all dyads.

Considering this let us introduce two new variables, z_{i1} and z_{i2} from the raw database, as follows:

$$z_{i1} = ½\ (x_{i1} + x_{i2}), \text{ and } z_{i2} = ½\ |x_{i1} - x_{i2}|,$$

where x_{i1} and x_{i2} are the answers of the first and second informants of the dyads to a specific question. The first new variable (z_{i1}) can be interpreted as the variable measuring the aggregated effect, while the second (z_{i2}) measures the differences between these answers. We emphasize that the benefit of these new variables subsists in their indifference to the order in which the data are entered into the database. One can easily recover the initial, raw data from these new variables:

a) If $x_{i1} \geq x_{i2}$, then $x_{i1} = z_{i1} + z_{i2}$ and $x_{i2} = z_{i1} - z_{i2}$.
b) If $x_{i1} < x_{i2}$, then $x_{i2} = z_{i1} + z_{i2}$ and $x_{i1} = z_{i1} - z_{i2}$.

1.3 Homogeneity analysis of exchangeable cases – the pairwise intraclass correlation

In exchangeable cases, one should begin dyadic data analysis with homogeneity analysis, which is carried out using the pairwise interclass correlation (Kenny et al., 2006) based on double entry. As stated above, the null hypothesis is that the informants of the dyads give homogeneous answers. In this section, we demonstrate that homogeneity can be tested in a simpler way, based on the initial database that does not require the technique of double entry. We develop and suggest a formula that approximates the suggested pairwise interclass correlation of DDA. After presenting our theoretical argument, we test the developed formula using our database and calculate the homogeneity in the case of randomly chosen variables with both the pairwise interclass correlation and our suggested approximation.

1.3.1 Theoretical argument

As described above, the technique of double entry transforms the initial database with n dyads (vectors) into another database with $2n$ dyads, as in Table 2. Suppose we fix the order of the informants within the dyads. This means that the previously mentioned issue of exchangeability is not a problem. We denote the values of the variables in the initial database as (x_1, x_2) and (y_1, y_2). The values of the same variables obtained with double entry are denoted as (X, X') and (Y, Y'). The values (X, X') and (Y, Y') are the transformed values of (x_1, x_2) and (y_1, y_2). Therefore, assuming the data can be rearranged, we obtain:

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, X' = \begin{bmatrix} x_2 \\ x_1 \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \text{ and } Y' = \begin{bmatrix} y_2 \\ y_1 \end{bmatrix}.$$

This equation reflects that the new variables can be derived from the initial ones by arranging the two vectors for a specific observation, one on top of the other, in reverse order. Remember, our key question is whether applying double entry is beneficial or not; do we obtain additional information with this method that is useful for further statistical analysis?

To avoid biases, we assume that the vectors represent the population. This way, we can facilitate calculation and use the number of vectors in variance-covariance calculations. First, we calculate the means for both variables and databases:

$$E\,(X) = E\,(X') = \frac{E\,(x_1) + E\,(x_2)}{2}, \text{ and } E\,(Y) = E\,(Y') = \frac{E\,(y_1) + E\,(y_2)}{2},$$

which can be determined easily. These equations indicate that the means of the new variables obtained through double entry are equal to the means of the original elements. We can formulate this differently; the mean of all the answers corresponding to a variable is the same as the mean of vectors *X* and *Y*, which stems from the technique of double entry.

Calculating the variance requires slightly more patience, but it is not very complicated either:

$$var\,(X) = var\,(X') = \frac{var\,(x_1) + var\,(x_2)}{2} + \left(\frac{E\,(x_1) + E\,(x_2)}{2}\right)^2, \text{ and} \tag{1}$$

$$var\,(Y) = var\,(Y') = \frac{var\,(y_1) + var\,(y_2)}{2} + \left(\frac{E\,(y_1) + E\,(y_2)}{2}\right)^2. \tag{2}$$

In addition, the covariance is calculated as follows:

$$cov\,(X, X') = cov\,(x_1, x_2) - \left(\frac{E\,(x_1) + E\,(x_2)}{2}\right)^2, \text{ and} \tag{3}$$

$$cov\,(Y, Y') = cov\,(y_1, y_2) - \left(\frac{E\,(y_1) + E\,(y_2)}{2}\right)^2. \tag{4}$$

Moreover,

$$cov\,(X, Y) = cov\,(X', Y') = \frac{cov\,(x_1, y_1) + cov\,(x_2, y_2)}{2} + \frac{[E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{4}, \text{ and} \tag{5}$$

$$cov\,(X, Y') = cov\,(X', Y) = \frac{cov\,(x_1, y_2) + cov\,(x_2, y_1)}{2} - \frac{[E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{4}. \tag{6}$$

Let us note here that the double entry actually decreases the amount of useful information, since the mean, variance and covariance of the new variables are, in several cases, the same. Because of the symmetries mentioned, related indices of variables (x_1, x_2) and (y_1, y_2) cannot be calculated from the new variables (*X*, *X'*) and (*Y*, *Y'*) without knowing a construction algorithm of the last variables. This finding reflects a unidirectional logical relationship between the two databases; variables (x_1, x_2) and (y_1, y_2) unambiguously determine (*X*, *X'*) and (*Y*, *Y'*), while the reverse does not hold. The loss of information is due to this asymmetry.

This fact has the consequence that we can find a relation between the new and old covariances only in a few cases. These cases are the following using (3) and (4):

$$cov\,(X, X') \leq cov\,(x_1, x_2) \text{ and } cov\,(Y, Y') \leq cov\,(y_1, y_2)\,.$$

Additionally, the variance is a special case of the covariance from (1) and (2):

$$var\,(X) \geq \frac{var\,(x_1) + var\,(x_2)}{2} \geq \sqrt{var\,(x_1)} \cdot \sqrt{var\,(x_2)}\text{ , and}$$

$$var\,(Y) \geq \frac{var\,(y_1) + var\,(y_2)}{2} \geq \sqrt{var\,(y_1)} \cdot \sqrt{var\,(y_2)}\,.$$

We suppose that the informants in any dyad give nearly the same answers, i.e. the expected values are nearly the same. This can be expressed as follows:

$$max\,\{|E\,(x_1) - E\,(x_2)|;\,|E\,(y_1) - E\,(y_2)|\} \leq \varepsilon\,,$$

where ε is an arbitrarily small positive number. In this way, using the initial data, we obtain the following approximations for the new variables, which we obtained by using double entry:

$$var\,(X) = var\,(X') \sim \frac{var\,(x_1) + var\,(x_2)}{2},$$

$$var\,(Y) = var\,(Y') \sim \frac{var\,(y_1) + var\,(y_2)}{2},$$

$$cov\,(X, X') \sim cov\,(x_1, x_2)\,,$$

$$cov\,(Y, Y') \sim cov\,(y_1, y_2)\,,$$

$$cov\,(X, Y) = cov\,(X', Y') \sim \frac{cov\,(x_1, y_1) + cov\,(x_2, y_2)}{2}\text{, and}$$

$$cov\,(X, Y') = cov\,(X', Y) \sim \frac{cov\,(x_1, y_2) + cov\,(x_2, y_1)}{2}.$$

These relations can be confirmed using elementary statistical methods, so we do not present their detailed derivation. We can state that the variance of variable X is larger than the product of the variances of the two vectors. This also might lead to a loss of information.

Since in case of the two covariances – *cov* (X,Y) and cov (X,Y') – the product of the expected values on the right-hand side can be either positive or negative, we cannot estimate the relation between the covariances. However, we can state that:

$$cov\,(X, Y) + cov\,(X, Y') = cov\,(X, Y + Y') = \frac{cov\,(x_1, y_1) + cov\,(x_1, y_2) + cov\,(x_2, y_1) + cov\,(x_2, y_2)}{2} = \frac{cov\,(x_1 + x_2, y_1 + y_2)}{2},$$

which results from the application of variance-covariance algebra.

We express the two correlations using the following formulas using (3), (1) and (4), (2):

$$r\,(X, X') = \frac{\sqrt{var\,(x_1)} \cdot \sqrt{var\,(x_2)} \cdot r\,(x_1, x_2) - (\frac{E\,(x_1) - E\,(x_2)}{2})^2}{\frac{var\,(x_1) + var\,(x_2)}{2} + (\frac{E\,(x_1) - E\,(x_2)}{2})^2}\text{, and}$$

$$r\,(Y,\,Y') = \frac{\sqrt{var\,(y_1)} \cdot \sqrt{var\,(y_2)} \cdot r\,(y_1, y_2) - (\frac{E\,(y_1) - E\,(y_2)}{2})^2}{\frac{var\,(y_1) + var\,(y_2)}{2} + (\frac{E\,(y_1) - E\,(y_2)}{2})^2}.$$

Recall that the variances of the two new variable pairs (*X'* and *Y'*) are the same as variances of *X* and *Y*. Therefore, the correlations can be approximated as follows in case of positive correlations:

$$r\,(X,\,X') = \frac{\sqrt{var\,(x_1)} \cdot \sqrt{var\,(x_2)}}{\frac{var\,(x_1) + var\,(x_2)}{2}} \cdot r\,(x_1, x_2) \leq r\,(x_1, x_2), \text{ and} \tag{7}$$

$$r\,(Y,\,Y') = \frac{\sqrt{var\,(y_1)} \cdot \sqrt{var\,(y_2)}}{\frac{var\,(y_1) + var\,(y_2)}{2}} \cdot r\,(y_1, y_2) \leq r\,(y_1, y_2).$$

This equation implies that the homogeneity analysis of dyadic data analysis can be carried out not only using the ANOVA tables but also using the initial database. There is no need to introduce new variables by applying double entry. Using correlations $r\,(x_1, x_2\,)$ and $r\,(y_1, y_2\,)$,we can analyze whether the answers of the two informants in a given dyad correspond to each other or not, i.e., whether a linear relationship between them exists or not. The suggested method and the same calculations are also relevant for distinguishable cases.

1.3.2 Testing homogeneity with DDA and the suggested approximation

Above, we presented a new formula that can replace pairwise intraclass correlation so double entry can be omitted. In this way, statistical analysis becomes easier yet remains reliable. Recall that this formula is given as (7).

This formula not only indicates the possibility of leaving out double entry but also reveals that it will result in higher values than the original dyadic correlation based on double entry. This finding may also indicate information loss due to double entry.

We tested the homogeneity using both formulas. The original pairwise intraclass correlation index is 0.490537. The reduced, simplified correlation index is 0.490877. This supports our statement that the difficulties raised by double entry are not outweighed by its potential positive effect.

2 A CRITICAL DISCUSSION OF CORRELATIONS OF DYADIC DATA ANALYSIS

Dyadic data analysis has introduced five types of correlations (Griffin and Gonzalez, 1995, 1999, 2004), excluding the pairwise interclass correlation discussed above:

1. Overall within-partner correlation;
2. Cross-intraclass correlation;
3. Mean-level correlation (correlation between dyad means);
4. Individual-level correlation;
5. Dyad-level correlation.

This section critically discusses these correlations and presents approximations for them based on a similar logic to that applied before. First, we theoretically discuss these correlations, and develop the approximations. Next, we test them using the database developed previously.

2.1 The overall within-partner and the cross-intraclass correlations

The overall within-partner correlation $r(X,\,Y)$ is specified in dyadic data analysis by the following equation using (5), (1) and (2):

$$r\,(X,\,Y) = \frac{\frac{1}{2}\cdot[\sqrt{var\,(x_1)}\cdot\sqrt{var\,(y_1)}\cdot r\,(x_1,y_1)+\sqrt{var\,(x_2)}\cdot\sqrt{var\,(y_2)}\cdot r\,(x_2,y_2)]+\frac{[E\,(x_1)-E\,(x_2)]\cdot[E\,(y_1)-E\,(y_2)]}{4}}{\sqrt{\frac{var\,(x_1)+var\,(x_2)}{2}+\left(\frac{E\,(x_1)-E\,(x_2)}{2}\right)^2}\cdot\sqrt{\frac{var\,(y_1)+var\,(y_2)}{2}+\left(\frac{E\,(y_1)-E\,(y_2)}{2}\right)^2}}\,.$$

The covariance in the formula's numerator measures the direction of the stochastic relationship between the answers of the two informants within a given dyad. In this way, this covariance can be interpreted as an 'internal' or 'individual' correlation.

Let us suppose again that both expected values and variances are approximately the same. Then,

$$max\,\{|var\,(x_1)-var\,(x_2)|;\,|var\,(y_1)-var\,(y_2)|\}\le\eta\,,$$

where η is an arbitrarily small number. For the above positive correlation, we can formulate the following approximation:

$$r\,(X,\,Y)\sim\frac{\frac{1}{2}\cdot[\sqrt{var\,(x_1)}\cdot\sqrt{var\,(y_1)}\cdot r\,(x_1,y_1)+\sqrt{var\,(x_2)}\cdot\sqrt{var\,(y_2)}\cdot r\,(x_2,y_2)]}{\sqrt{\frac{var\,(x_1)+var\,(x_2)}{2}}\cdot\sqrt{\frac{var\,(y_1)+var\,(y_2)}{2}}}$$
$$\le\frac{1}{2}\cdot[r\,(x_1,y_1)+r\,(x_2,y_2)].$$

The cross-intraclass correlation is defined as follows using (6), (1) and (2):

$$r\,(X,\,Y') = \frac{\frac{1}{2}\cdot[\sqrt{var\,(x_1)}\cdot\sqrt{var\,(y_2)}\cdot r\,(x_1,y_2)+\sqrt{var\,(x_2)}\cdot\sqrt{var\,(y_1)}\cdot r\,(x_2,y_1)]+\frac{[E\,(x_1)-E\,(x_2)]\cdot[E\,(y_1)-E\,(y_2)]}{4}}{\sqrt{\frac{var\,(x_1)+var\,(x_2)}{2}+\left(\frac{E\,(x_1)-E\,(x_2)}{2}\right)^2}\cdot\sqrt{\frac{var\,(y_1)+var\,(y_2)}{2}+\left(\frac{E\,(y_1)-E\,(y_2)}{2}\right)^2}}\,.$$

The covariance of the initial dataset reflects the relationship between the answers given by the two informants of a specific dyad to two different questions. Based on the previous argument, this covariance is approximated as follows:

$$r\,(X,\,Y')\sim\frac{\frac{1}{2}\cdot[\sqrt{var\,(x_1)}\cdot\sqrt{var\,(y_2)}\cdot r\,(x_1,y_2)+\sqrt{var\,(x_2)}\cdot\sqrt{var\,(y_1)}\cdot r\,(x_2,y_1)]}{\sqrt{\frac{var\,(x_1)+var\,(x_2)}{2}}\cdot\sqrt{\frac{var\,(y_1)+var\,(y_2)}{2}}}$$
$$\le\frac{1}{2}\cdot[r\,(x_1,y_2)+r\,(x_2,y_1)].$$

2.2 Mean-level correlation

The mean-level correlation (also called the correlation between dyad means) is specified by Griffin and Gonzalez (1995) as follows:

$$r_m\,(X,X',\,Y,\,Y') = \frac{r\,(X,\,Y)+r\,(X,\,Y')}{\sqrt{1+r\,(X,X')}\cdot\sqrt{1+r\,(Y,\,Y')}}\,. \tag{9}$$

The Formula (9) can be rewritten in terms of variances and covariances. After small transformations, and using elementary covariance algebra, we obtain:

$$r_m\,(X,X',\,Y,\,Y') = \frac{cov\,(X,\,Y+Y')}{\sqrt{cov\,(X,X+X')}\cdot\sqrt{cov\,(Y,\,Y+Y')}}\,.$$

After calculating the covariance, this expression can be rewritten in terms of the raw data:

$$r_m\,(X, X', Y, Y') = \frac{\frac{1}{2} \cdot cov\,(x_1 + x_2, y_1, y_2)}{\sqrt{\frac{1}{2} \cdot var\,(x_1 + x_2)} \cdot \sqrt{\frac{1}{2} \cdot var\,(y_1 + y_2)}} = r\,(x_1 + x_2, y_1, y_2)\,.$$

This means that the dyad-level correlation is a classical correlation that interprets the correlation between two newly introduced variables as the sum of the dyad-level values. Interestingly, when using new data, r_m (X, X', Y, Y') does not correspond to the traditional Pearson correlation because the covariance in the numerator suggests the formula $\sqrt{var\,(X)} \cdot \sqrt{var\,(Y + Y')}$ instead of the covariance. If somebody takes the trouble to calculate the classical correlation, he/she will conclude:

$$r\,(X, Y + Y') = \frac{cov\,(X, Y + Y')}{\sqrt{var\,(X)} \cdot \sqrt{var\,(Y + Y')}} = \frac{\frac{1}{2} \cdot cov\,(x_1 + x_2, y_1, y_2)}{\sqrt{\frac{var\,(x_1) + var\,(x_2)}{2}} + \sqrt{(\frac{E\,(x_1) - E\,(x_2)}{2})^2} \cdot \sqrt{\frac{1}{2} \cdot var\,(y_1 + y_2)}}\,.$$

This is not the same as the previous correlation, $r\,(x_1 + x_2, y_1 + y_2)$, but it is very close to it.

2.3 Individual-level correlation

The most problematic correlation coefficients in dyadic data analysis are the individual- and dyad-level correlation coefficients. The individual-level correlation is suggested to calculate:

$$r_i\,(X, X', Y, Y') = \frac{r\,(X, Y) - r\,(X, Y')}{\sqrt{1 - r\,(X, X')} \cdot \sqrt{1 - r\,(Y, Y')}}\,. \qquad (10)$$

The Formula (10) can also be rewritten in terms of the variance and covariance:

$$r_i\,(X, X', Y, Y') = \frac{cov\,(X, Y - Y')}{\sqrt{cov\,(X, X - X')} \cdot \sqrt{cov\,(Y, Y - Y')}}\,.$$

Before proceeding with the transformation, we present the traditional Pearson correlation, which is widely available in the statistical literature:

$$r\,(X, Y - Y') = \frac{cov\,(X, Y - Y')}{\sqrt{var\,(X)} \cdot \sqrt{var\,(Y - Y')}} = \frac{\frac{1}{2} \cdot cov\,(x_1 - x_2, y_1 - y_2) + \frac{[E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{4}}{\sqrt{\frac{var\,(x_1) + var\,(x_2)}{2} + (\frac{E\,(x_1) - E\,(x_2)}{2})^2} \cdot \sqrt{var\,(y_1 - y_2) + E\,(y_1) - E\,(y_2))^2}}\,.$$

Now, we continue the process of reducing the correlation to a formula using initial, raw data. The above expression is similar to the mean-level correlation discussed above; the difference is in the reversed signs. As a next step, we substitute our initial data into the above formula and

$$r_i\,(X, X', Y, Y') = \frac{\sqrt{var\,(x_1 - x_2)} \cdot \sqrt{var\,(y_1 - y_2)} \cdot r\,(x_1 - x_2, y_1 - y_2) + [E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{\sqrt{var\,(x_1 - x_2) + [E\,(x_1) - E\,(x_2)]^2} \cdot \sqrt{var\,(y_1 - y_2) + [E\,(y_1) - E\,(y_2)]^2}}\,.$$

This formula indicates that the upper limit of the individual-level correlation is the correlation between variables, which is the correlation between the differences in the answers of the partners in any given dyad.

We can approximate this positive correlation by supposing that the expected values of the answers given by the two informants of any dyad or pair to the two questions/variables are equal:

$$r_i\,(X, X', Y, Y') = \frac{\sqrt{var\,(x_1 - x_2)} \cdot \sqrt{var\,(y_1 - y_2)} \cdot r\,(x_1 - x_2, y_1 - y_2) + [E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{\sqrt{var\,(x_1 - x_2) + [E\,(x_1) - E\,(x_2)]^2} \cdot \sqrt{var\,(y_1 - y_2) + [E\,(y_1) - E\,(y_2)]^2}}$$
$$\leq [r\,(x_1 - x_2, y_1 - y_2)].$$

This proves that this correlation actually measures the difference in the individual effect between dyads.

2.4 Dyad-level correlation

Lastly, we discuss the dyad-level correlation, which is described by the following formula:

$$r_d\,(X, X', Y, Y') = \frac{r\,(X, Y')}{\sqrt{r\,(X, X')} \cdot \sqrt{r\,(Y, Y')}}\,.$$

Let us remark that this is not a strict correlation in traditional statistical terms, since the variables under the square root might have negative values. This happens when the informants of a dyad give opposite answers to a question. Now, we set aside this problem and suppose that the expression under the square root is non-negative. The above formula can be transformed using the definition of correlation as follows using (1)–(6):

$$r_d\,(X, X', Y, Y') = \frac{\frac{1}{2} \cdot [\sqrt{var\,(x_1)} \cdot \sqrt{var\,(y_2)} \cdot r\,(x_1, y_2) + \sqrt{var\,(x_2)} \cdot \sqrt{var\,(y_1)} \cdot r\,(x_2, y_1)] - \frac{[E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{4}}{\sqrt{cov\,(x_1, x_2) - (\frac{E\,(x_1) - E\,(x_2)}{2})^2} \cdot \sqrt{cov\,(y_1, y_2) - (\frac{E\,(y_1) - E\,(y_2)}{2})^2}}\,.$$

We can see that if:

$$cov\,(x_1, x_2) - \left(\frac{E\,(x_1) - E\,(x_2)}{2}\right)^2 < 0 \text{ and/or}$$
$$cov\,(y_1, y_2) - \left(\frac{E\,(y_1) - E\,(y_2)}{2}\right)^2 < 0,$$

then this type of correlation cannot be produced. This reflects that dyad-level correlation is similar to cross-intraclass correlation, as discussed in the context of homogeneity analysis.

When analyzing the covariance in the numerator of our expression, we can see that the correct correlation here is the cross-intraclass correlation $r\,(X, Y')$. This result can also be obtained by supposing the members of the dyads give similar answers to the questions. In such cases, the covariance becomes close to the variance because the expected values and standard deviations are close to each other.

2.5 Testing our suggested approximation formulas

In the previous sections, we critically analyzed five correlations, which were developed by dyadic data analysis. In the next table, we summarize these correlations, giving both the formulas developed by DDA and our suggested approximations.

These correlations were analyzed locally, and approximations were developed. We summarize our results in Table 7.

Table 6 DDA correlations with double entry and using the initial database

Type of correlation	**(*X*, *X'*), (*Y*, *Y'*)** **(double entry)**	**(x_1, x_2), (y_1, y_2)** **(initial data)**
Cross- intraclass correlation	$r\,(X, Y') = \dfrac{cov\,(X, Y')}{\sqrt{var\,(X)} \cdot \sqrt{var\,(Y)}}\,,$	$r\,(X, Y') = \dfrac{\frac{cov\,(x_1 \cdot y_2) + cov\,(x_2 \cdot y_1)}{2} - \frac{[E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{4}}{\sqrt{\frac{var\,(x_1) + var\,(x_2)}{2} + (\frac{E\,(x_1) - E\,(x_2)}{2})^2} \cdot \sqrt{\frac{var\,(y_1) + var\,(y_2)}{2} + (\frac{E\,(y_1) - E\,(y_2)}{2})^2}}\,.$
Overall within-partner correlation	$r\,(X, Y) = \dfrac{cov\,(X, Y)}{\sqrt{var\,(X)} \cdot \sqrt{var\,(Y)}}\,,$	$r\,(X, Y) = \dfrac{\frac{cov\,(x_1 \cdot y_1) + cov\,(x_2 \cdot y_2)}{2} + \frac{[E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{4}}{\sqrt{\frac{var\,(x_1) + var\,(x_2)}{2} + (\frac{E\,(x_1) - E\,(x_2)}{2})^2} \cdot \sqrt{\frac{var\,(y_1) + var\,(y_2)}{2} + (\frac{E\,(y_1) - E\,(y_2)}{2})^2}}\,.$
Mean-level correlation	$r_m\,(X, X', Y, Y') = \dfrac{r\,(X, Y) + r\,(X, Y')}{\sqrt{1 + r\,(X, X')} \cdot \sqrt{1 + r\,(Y, Y')}}\,,$	$r_m\,(X, X', Y, Y') = r\,(x_1 + x_2, y_1 + y_2).$
Individual-level correlation	$r_i\,(X, X', Y, Y') = \dfrac{r\,(X, Y) - r\,(X, Y')}{\sqrt{1 - r\,(X, X')} \cdot \sqrt{1 - r\,(Y, Y')}}\,.$	$r_i\,(X, X', Y, Y') = \dfrac{cov\,(x_1 - x_2, y_1 - y_2) + [E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{\sqrt{var\,(x_1 - x_2) + [E\,(x_1) - E\,(x_2)]^2} \cdot \sqrt{var\,(y_1 - y_2) + [E\,(y_1) - E\,(y_2)]^2}}\,.$
Dyad-level correlation	$r_d\,(X, X', Y, Y') = \dfrac{r\,(X, Y')}{\sqrt{r\,(X, X')} \cdot \sqrt{r\,(Y, Y')}}\,,$	$r_d\,(X, X', Y, Y') = \dfrac{\frac{cov\,(x_1, y_2) + cov\,(x_2, y_1)}{2} - \frac{[E\,(x_1) - E\,(x_2)] \cdot [E\,(y_1) - E\,(y_2)]}{4}}{\sqrt{cov\,(x_1, x_2) - (\frac{E\,(x_1) - E\,(x_2)}{2})^2} \cdot \sqrt{cov\,(y_1, y_2) - (\frac{E\,(y_1) - E\,(y_2)}{2})^2}}\,.$

Source: Own construction

Table 7 Suggested approximations of the correlations specified by DDA using the initial, raw data

Types of DDA correlations	Suggested approximations
Overall within-partner correlation	$r\,(X, Y) \sim \dfrac{1}{2} \cdot [r\,(x_1, y_1) + r\,(x_2, y_2)]$
Cross-intraclass correlation	$r\,(X, Y') \sim \dfrac{1}{2} \cdot [r\,(x_1, y_2) + r\,(x_2, y_1)]$
Mean-level correlation	$r_m\,(X, X', Y, Y') = r\,(x_1 + x_2, y_1 + y_2)$
Individual-level correlation	$r_i\,(X, X', Y, Y') \sim r\,(x_1 - x_2, y_1 - y_2)$

Source: Own construction

We have theoretically elaborated the different correlation types and developed the formulas presented above. These formulas enable to avoid the application of double entry and to approximate correlations using the initial/raw data. Using our database and the same variables as before, we have calculated these correlations applying both the suggested traditional DDA formulas based on double entry and our developed expressions based on the initial data. Our objective is to test whether our suggested approximations lead to good results. If this is the case, the technique of double entry does not necessarily lead to additional information for statistical analysis. In Table 8, we have summarized the results of our empirical tests.

Table 8 Summary of the results of testing the correlation coefficients using the formulas of dyadic data analysis (based on double entry) and the suggested approximations (with initial database)

Types of DDA correlations	Values calculated using the database developed through double entry	Values calculated using the initial dataset
Overall within-partner correlation	0.291	0.293
Cross-intraclass correlation	0.588	0.589
Mean-level correlation	0.617	0.617
Individual-level correlation	0.522	0.522
Dyad-level correlation	0.689	0.293

Source: Own construction

Our suggested approximations resulted in good agreement with the original correlation indices of DDA, except for the dyad-level correlation. This result also supports our statement that the database development technique of double entry does not always yield significant benefit for statistical analysis.

3 DYADIC REGFRESSION MODELS

The core question of regression models is the effect that the independent variable has on the dependent variable. It is assumed that it is easier to specify independent variables in classical statistics compared to dyadic data analysis because DDA takes into account not only individual-level but also dyad-level effects. Therefore, regression analysis of dyadic data necessitates incorporating several factors, even if we have only one independent and one dependent variable. These factors are as follows (Gonzalez, 2010):

- Actor effect,
- Partner effect,
- Mutual effect.

The model of the intraclass correlation coefficient (ICC) incorporates only the actor and partner effects, while the actor-partner interdependence model (APIM) takes into consideration the mutual effect as well.

3.1 Theoretical discussion

In this section, we discuss the ICC model. First, we introduce the model. The objective is to analyze critically whether this linear model is capable of describing complex relationships between its dyadic variables. We know that the ICC model aims at describing only the actor and partner effects.

The model is formulated mathematically as follows (Gonzalez and Griffin, 2000):

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X' + \varepsilon \,,$$

where *X* and *X'* are the independent variables that we obtained using double entry, *Y* is the dependent variable, ε is the error vector, and β_0, β_1 and β_2 are the regression coefficients.

This model can also be expressed in terms of the initial database as follows:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \beta_0 \cdot \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \beta_1 \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \beta_2 \cdot \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix},$$

where vector 1 is a vector, in which all elements are equal to 1, and ε_1 and ε_2 are the error vectors. We unfold this estimation to examine its elements:

$$y_1 = \beta_0 \cdot 1 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \varepsilon_1 \text{ and}$$

$$y_2 = \beta_0 \cdot 1 + \beta_1 \cdot x_2 + \beta_2 \cdot x_1 + \varepsilon_2 \,.$$

We see that regression parameters in the second equation are the same as those in the first. This means that the estimate based on a database developed using the double-entry technique loosely approximates the value of any variable given by the second member of the dyad as y_2.

Based on the above argument, the following formulas lead to a better estimate:

$$y_1 = \beta_{01} \cdot 1 + \beta_{11} \cdot x_1 + \beta_{21} \cdot x_2 + \varepsilon_{11} \text{ and}$$

$$y_2 = \beta_{02} \cdot 1 + \beta_{12} \cdot x_1 + \beta_{22} \cdot x_2 + \varepsilon_{21} \,.$$

Here, we must estimate six coefficients instead of three. The main complication is that the previous two estimation equations are transformed into two independent equations that are not linked by any joint coefficients; ε_{11} and ε_{21} are the error vectors. Although we discuss only exchangeable cases in the paper, the proposed estimations can also be useful for distinguishable ones.[4]

One can also see that the estimate suggested above leads to a smaller error and parameters can capture linear relationships more precisely (given, of course, that both models use the same estimation method)[5].

We assume that parameters ($\beta_{01}, \beta_{11}, \beta_{21}$) and ($\beta_{02}, \beta_{12}, \beta_{22}$) optimize our estimation functions that are defined as the least square functions, i.e. f_1 ($\beta_{01}, \beta_{11}, \beta_{12}$) and f_2 ($\beta_{02}, \beta_{12}, \beta_{22}$). Rao et al. (2008) and Grosz (2011) describe the solution procedure in their works. In this case, the estimation function of the first model –which is obtained using the same methodology, namely, the least-squares procedure– leads to

$$f_1 (\beta_0, \beta_1, \beta_2) + f_2 (\beta_0, \beta_2, \beta_1).$$

Because f_1 ($\beta_{01}, \beta_{11}, \beta_{12}$) and f_2 ($\beta_{02}, \beta_{12}, \beta_{22}$) utilize optimal coefficients, the following hold:

$$f_1 (\beta_{01}, \beta_{11}, \beta_{12}) \leq f_1 (\beta_0, \beta_1, \beta_2) \text{ and}$$

$$f_2 (\beta_{02}, \beta_{12}, \beta_{22}) \leq f_2 (\beta_0, \beta_2, \beta_1)_,$$

which means that:

$$f_1 (\beta_{01}, \beta_{11}, \beta_{12}) + f_2 (\beta_{02}, \beta_{12}, \beta_{22}) \leq f_1 (\beta_0, \beta_1, \beta_2) + f_2 (\beta_0, \beta_2, \beta_1) = f (\beta_0, \beta_1, \beta_2)\ .$$

We proved that the modified linear model using the initial dataset offers a better estimate than the original estimate suggested by DDA. We continue our discussion with the APIM model.

The APIM model differs from the ICC model with respect to the mutual effect. This model not only maps the interrelations between the partners of the dyads (the actor and partner effects) but also incorporates into the model the interrelations among different dyads. The mathematical formula is:

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X' + \beta_3 \cdot \langle X \cdot X' \rangle + \varepsilon\ ,$$

where β_0, β_1 and β_2 are defined as in the case of the ICC model, and ε again denotes the error. The only difference between the two formulas is that the mutual effect is incorporated into the model using the expression $\beta_3 \cdot \langle X \cdot X' \rangle$.

In this case, vector $\langle X \cdot X' \rangle$ is a new variable reflecting the joint, mutual effect of the partners in the same dyad on one partner's (called the actor) Y variable (or answer).

Again, we can express the model using the initial dataset in the following way:

$$y_1 = \beta_0 \cdot 1 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle\ + \varepsilon_1 \text{ and}$$

[4] Let us note that the basic objective of this paper is to critically analyze the statistical consequences of double entry, so we apply the classic regression models (Gonzalez and Griffin, 2000). Both ICC and APIM have been further developed. Discussion of these extended models is not in the focus of our paper (N.N., 2019.)

[5] Let us use the least-squares procedure for the estimation. In this case, the two equations we obtained are independent. The estimation functions obtained by the least-squares procedure are quadratic functions in the case of the first equation f_1, while for the second equation f_2 the parameters minimize the estimation functions, so we obtain the following inequalities: $f_1() \leq f_1()$ and $f_2() \leq f_2()$. Since the parameters of the least-squares procedure maximize R^2, the two equations will result in a slightly better estimate. We can make a similar argument in the case of maximum likelihood estimation.

$$y_2 = \beta_0 \cdot 1 + \beta_1 \cdot x_2 + \beta_2 \cdot x_1 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_2 .$$

Expression $\langle x_1 \cdot x_2 \rangle$ denotes the vector that is created by multiplying the elements of vectors x_1 and x_2.

In this case, the following new functions are suggested:

$$y_1 = \beta_{01} \cdot 1 + \beta_{11} \cdot x_1 + \beta_{21} \cdot x_2 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_{11} \text{ and}$$

$$y_2 = \beta_{02} \cdot 1 + \beta_{12} \cdot x_1 + \beta_{22} \cdot x_2 + \beta_3 \cdot \langle x_1 \cdot x_2 \rangle + \varepsilon_{21} .$$

The considerations discussed in relation to the ICC model are relevant here as well. Consequently, the estimation functions suggested above are superior.

3.2 Testing the suggested estimation functions for the ICC model

We have tested the suggested estimation functions for the ICC model and carried out calculations using the DDA functions. *Y* is the dependent variable, and *X* and *X'* are the independent variables.

Using the ICC model, the value of *R* was 0.588 (Table 9). The model and the coefficient of variable *X* were significant, but the coefficient of *X'* was not.

Table 9 Results of the ICC model

R	R^2	Adjusted R^2	Standard error
0.588	.346	.338	1.220

Independent variables: X, X′

ANOVA table

Model	Sum of squares	df	Mean of sum of squares	F	Sig.
Regression	137.819	2	68.910	46.276	.000
Residual	260.591	175	1.489		
Sum	398.410	177			

Dependent variable: Y
Independent variables: X, X′

Coefficients

Model	Non standardized coefficients		Standardized coefficient	t	Sig.
	B	Std. error	Beta		
Constant	.761	.096		7.938	.000
X	.495	.059	.586	8.357	.000
X′	.004	.059	.004	.063	.950

Dependent variable: *Y*
Source: Own construction

As a next step, we applied the model to the initial database, as suggested previously:

$$y_1 = \beta_{01} \cdot 1 + \beta_{11} \cdot x_1 + \beta_{21} \cdot x_2 + \varepsilon_{11} \text{ and}$$

$$y_2 = \beta_{02} \cdot 1 + \beta_{12} \cdot x_1 + \beta_{22} \cdot x_2 + \varepsilon_{21} \, .$$

Recall that here we must estimate six coefficients, instead of the three for the original ICC model, and we must use two independent, separate estimation functions. Here, ε_{11} and ε_{21} are the errors.

We have calculated the two regression models. The results of Model 1 are summarized in Table 10, and the results of Model 2 are summarized in Table 11.

3.2.1 Model using the initial or raw dataset

Table 10 Results of the regression model between y1 and x1, x2, respectively – Model 1

Variables

Model	Independent variables
1	x_1, x_2

Dependent variable: y_1

Model	R	R^2	Adjusted R^2	Standard error
1	.583	.339	.324	1.170

Independent variables: x_1, x_2

ANOVA table

Model		Sum of squares	df	Mean of sum of squares	F	Sig.
1	Regression	60.481	2	30.241	22.096	.000
	Residual	117.699	86	1.369		
	Sum	178.180	88			

Dependent variable: y_1
Independent variables: x_1, x_2

Coefficients

Model		Non standardized coefficients		Standardized coefficient	t	Sig.
		B	Std. error	Beta		
1	Constant	.738	.130		5.672	.000
	x_1	.438	.079	.560	5.569	.000
	x_2	.035	.082	.043	.429	.669

Dependent variable: y_1
Source: Own construction

3.2.2 Model using the initial or raw database

Table 11 Results of the regression model between y_2 and x_1, x_2 – Model 2

Variables

Model	Independent variables
2	x_1, x_2

Dependent variable: y_2

Model	R	R^2	Adjusted R^2	Standard error
2	.599	.358	.343	1.282

Independent variables: x_1, x_2

ANOVA table

Model		Sum of squares	df	Mean of sum of squares	F	Sig.
2	Regression	78.907	2	39.453	24.010	.000
	Residual	141.318	86	1.643		
	Sum	220.225	88			

Dependent variable: y_2
Independent variables: x_1, x_2

Coefficients

Model		Non standardized coefficients		Standardized coefficient	t	Sig.
		B	Std. error	Beta		
2	Constant	.793	.143		5.562	.000
	x_1	–.028	.086	–.033	–.328	.744
	x_2	.558	.090	.614	6.192	.000

Dependent variable: y_2
Source: Own construction

The calculations presented above support our theoretical argument; in the case of ICC, we have obtained very similar results with the two suggested models that leave out the technique of double entry and use the initial database for analysis.

SUMMARY – CONCLUSION

Dyadic phenomena have become highly important not only in sociology and psychology but in a networked economy for economics and management studies. The paper critically discussed a relatively new statistical methodology that was developed for analyzing such dyadic problems, called dyadic data analysis. We had two objectives with our work. On the one hand, we critiqued the database development of DDA related to exchangeable cases and suggested an algorithm for solving the problem transforming such a data set into a distinguishable one. On the other hand, we concentrated on double entry and its statistical consequences for classical dyadic correlations and regression analysis.

We concluded that an exchangeable case can be traced back to a distinguishable one with a relatively simple algorithm. Because of the symmetry of the partners' roles in any dyad, the number of potential databases is the exponential function of the number of dyads in the initial dataset. Therefore, in exchangeable cases, one has to look for a consensus in the way data are treated. We suggested applying a transformation of the initial data that eliminates this symmetry, such as summing and/or calculating the absolute values of the data differences.

The second focal issue of the paper was the double-entry technique. We analyzed whether this technique adds value through developing a richer information base or leads to information losses. Our examination revealed that double entry does not supply additional information compared to the initial database. Rather, it might lead to information losses, consequently making the statistical analysis less reliable.

We discussed the different correlation constructs of DDA, clarified their statistical content, and succeeded in tracing them back to the classical Pearson correlation. These correlation constructs also do not require the use of double entry. We reduced these correlations to a formula that uses the initial database to approximate them. This formula was carried out not only for the correlation constructs of DDA but also for its regression models. Statistical discussion revealed that in the ICC and APIM models, the double-entry method might make the estimates less reliable. The suggested regression models that use the simple initial database can achieve better estimation.

After we developed the new correlations and regression equations using the initial database, we carried out empirical analysis as well. We tested the suggested approximations for all correlation constructs and the ICC regression model with an empirical database. This database was developed in a previous field study using a trust-related questionnaire with pairwise sampling. In respect of the correlations we had mainly supporting results. Except for the dyad-level correlation, our suggested formulas resulted in good approximations. Results of the suggested two ICC regression models led to a slightly higher R^2, however differences were quite small. Empirical results support our theoretical argument in this respect too. We have to emphasize though, that other empirical databases might lead to different results, so further empirical research is needed in this respect.

ACKNOWLEDGMENT

The project is supported by the Hungarian Scientific Research Fund (OTKA), Project No. K 115542.

References

BURK, J. W., STEGLICH, C. E. G., SNIJDERS, T. A. B. Beyond dyadic interdependence: actor-oriented models for co-evolving social networks and individual behaviors. *International Journal of Behavioral Development*, 2007, Vol. 31, No. 4. pp. 397–404.

DENG, L. AND KE-HAI, Y. Multiple-group analysis for structural equation modeling with dependent samples. *Structural Equation Modeling: A Multidisciplinary Journal*, 2015, 22(4), pp. 552–567.

GELEI, A. AND DOBOS, I. Mutual trustworthiness as a governance mechanism in business relationships – A dyadic data analysis. *Acta Oeconomica*, 2016, 66(4), pp. 661–684.

GELEI, A. AND SUGÁR, A. The challenge of researching dyadic phenomena – the comparison of dyadic data analysis and traditional statistical methods. *Hungarian Statistical Review*, 2017, Special Number 21, pp. 78–100.

GONZALEZ, R. AND GRIFFIN, D. The correlational analysis of dyad-level data in the distinguishable case. *Personal Relationships*, 1999, 6(4), pp. 449–469.

GONZALEZ, R. AND GRIFFIN, D. On the Statistics of Interdependence: Treating Dyadic Data with Respect. In: ICKES, W. AND DUCK, S. eds. *The Social Psychology of Personal Relationships*, John Wiley and Sons, Ltd., 2000, pp. 181–213.

GONZALEZ, R. *Dyadic Data Analysis* [online]. University of Michigan. 2010. [cit. 2.5.2011] <http://www.cfs.purdue.edu/CFF/documents/Families_and_Health/purdue.pdf>.

GRIFFIN, D. AND GONZALEZ, R. Correlational Analysis of Dyad-Level Data in the Exchangeable Case. *Psychological Bulletin*, 1995, 118(3), pp. 430–439.

GROSZ, J. Identification of Influential Points in a Linear Regression Model [online]. *Statistika: Statistics and Economy Journal*, 2011, No. 1, pp. 71–77.

KENNY, D. A. *Dyadic Analysis* [online]. 2015. [cit. 26.1.2019] <http://davidakenny.net/dyad.htm>.

KENNY, D. A., KASHY, D. A., COOK, W. L. *Dyadic data Analysis*. New York, London: The Guilford Press, 2006.

LEDERMANN, T., MACHO, S., KENNY, D. A. Assessing mediation in dyadic data using the actor-partner interdependence model. *Structural Equation Modeling: A Multidisciplinary Journal*, 2011, 18(4), pp. 595–612.

LEDERMANN, T. AND KENNY, D. A. A toolbox with programs to restructure and describe dyadic data. *Journal of Social and Personal Relationships*, 2015, Vol. 32(8), pp. 997–1011. DOI: 10.1177/0265407514555273

MCARDLE, J. J. Dynamic but structural equation modeling of repeated measures data. In: NESSELROADE, J. R. AND CATTEL, R. B. eds. *Handbook of multivariate experimental psychology*, 2nd Ed. New York: Plenum, 1988, pp. 561–614.

N.N. *Dyadic Data Analysis* [online]. [cit. 11.3.2019] <https://www.mailman.columbia.edu/research/population-health-methods/dyadic-data-analysis>.

PEUGH, J. L., DILILLO, D., PANUZIO, J. Analyzing mixed-dyadic data using structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 2013, 20(2), pp. 314–337.

PLANALP, E. M., DU, H., BRAUNGART-RIEKER, J. M., WANG, L. Growth Curve Modeling to Studying Change: A Comparison of Approaches Using Longitudinal Dyadic Data With Distinguishable Dyads. *Structural Equation Modeling: A Multidisciplinary Journal*, 2017, 24(1), pp. 129–147.

RAO, R. C., TOUTENBURG, H., HEUMANN, C. *Linear Models and Generalizations: Least Squares and Alternatives*. Berlin: Springer, 2008.

WHITTAKER, T. A., BERETVAS, S. N., FALBO, T. Dyadic Curve-of-Factors Model: An Introduction and Illustration of a Model for Longitudinal Nonexchangeable Dyadic Data. *Structural Equation Modeling: A Multidisciplinary Journal*, 2014, 21(2), pp. 303–317.

Some Future Challenges for Czech Official Statistics

Edvard Outrata[1] | *Former President of the CZSO, Prague, Czech Republic*

Abstract

At its centenary, Czech official statistics is facing several challenges. Some are well known, others are emerging in a new and rapidly changing world. The Czech Statistical Office and other parts of Czech official statistics have a good record of coping and are well prepared. The time ahead, however, may be harder than the past. The acceleration of change in the world at large emphasises several challenges in front of official statistics, see relevance and education, as well as a number of technical developments like big data and changing means of communication. The main issue, however, will probably be the need to protect the independence of official statistics from vested interests in a world where automatisms replace specific decisions and where statistics are used as controls.

Keywords

Czech official statistics, challenges, dangers

JEL code

Z00

INTRODUCTION

We are celebrating a century of Czechoslovak official statistics these days, as well as 120 years of official statistics in the Kingdom of Bohemia. While the development of statistics in what is now the Czech Republic has, of course, a longer tradition both in the contributions of other institutions and in the context of imperial statistics in Vienna, these two dates are salient, because they mark the institutionalisation of official statistics as an autonomous entity in our country´s civil service and start a tradition of professionalism and trustworthiness that its present avatar, the Czech Statistical Office, builds upon.

The growth and professional development of official statistics in the 19th and 20th centuries were spurred by the self-confidence that enlightenment brought to the state and its civil service, namely that it is both right and possible to govern by taking decisions on the basis of verified facts (as opposed to hunches and feelings). This was later called "evidence-based decision-making". The concentration of official statistics in separate institutions created, in turn, further demand for the development of statistical theory and methods, by whose application statistics grew in quantity and quality. The modern state now requires a comprehensive description of the country and its people in all relevant aspects. It uses this picture and its dynamics constantly and routinely for the growing number of decisions it must take in the process of governing in a multiplying number of areas and in expanding detail. Official statistics, supplying data to evidence-based decision-making, has become a foundation stone of the process of governing.

As a result, advanced statistical offices are not run-of-the-mill civil service institutions. Specialised statistical work requires high proficiency both in up-to-date statistical theory and in the management

[1] Member of the Czech Statistical Council. E-mail: e.outrata@centrum.cz.

of country-wide activities. More than other state institutions, its effectiveness requires a high level of credibility both in protecting data provided by respondents in confidence, and in publishing true and reliable results. This credibility is required not just as a moral principle, but particularly, also as a condition of its proper functioning. Respondents could not be trusted to provide true information, if they did not trust the institution; nor would statistics from such data be of any use.

Czech official statistics has managed to maintain a high professional standard throughout its history. The fact that this standard was re-established early on after the fall of communism goes to the credit of the founders of Czech official statistics and their teams before the second world war, as well as to the professionalism of the staff and some of the leadership of the institution, who did not allow the ethos of the profession to disappear under the pressures of fifty years of totalitarian rule.

1 ACCELERATION OF CHANGE

The world, whose accurate description is the objective of official statistics, has always been changing. The rapidity of these changes has been continuously increasing at least since the period of enlightenment, during which the idea of beneficial change was promoted, based on growing knowledge and verified information. Official statistics was thus both an object and one of many subjects of this change and has been developing in the process; and keeping pace with these changes has always been a necessity and an important challenge. However, since about 30 years ago the number, rapidity and profoundness of changes in the world have advanced to the point where new types of challenges have developed and official statistics must cope with them. I see four categories of such changes, each posing distinct new issues, namely technological change. changes in human perception, shifting relevance, and changes in the process of governance.

Progress in information technology will probably continue to revolutionise all aspects of official statistics, data collection, data processing and dissemination. It will continue to spur methodological development (for example in the use and processing of big data).

As things stand at this point in time, in my opinion the following challenges facing the Czech Statistical Office and official statistics in general seem to merit consideration.

1.1 Relevance

One of the eternal challenges of official statistics is to provide its users the information they really require in a changing world. This problem is compounded by the need to maintain time series of indicators, which poses a limit on methodological improvement and serves as a conservative break on change. The need to harmonise statistics over large areas in order to allow horizontal comparisons and supranational statistical time series has been with us for a long time, but the need has grown more urgent with the integration of the European market. Reducing a statistic to only those components which can be harmonised can be dangerous, because it can limit the information the statistic provides. An example was the exclusion of housing prices from the harmonised CPI, which was necessary because of the differences in taxation schemes and living style patterns in different nations, but which reduced the relevance of the CPI as a practical measure of inflation for some important purposes.

The body of official statistics is fully anchored in the industrial age and is still largely calibrated to serve the needs of this age. Its product is at its most reliable where the measured phenomenon can be counted like beans. This is largely the case in population statistics and in statistics describing primary and secondary industries (or, more precisely, the legal market in primary and secondary industry products). It is also these areas where sampling methods are most effective in saving resources while maintaining a minimal and well measured loss of precision.

The product of services is much harder to measure in this fashion, so it often has to be roughly estimated. While this was acceptable in the past, the growing importance of the tertiary sector is making this situation awkward and may start to affect the reliability of some statistics (like the GDP). Our efforts in combining rough estimates in sectors that cannot be precisely measured with detailed precision in traditional sectors may sometimes seem similar to those of the man in the old joke, who was searching for a lost penny not in the dark where he had lost it, but under the lamppost where he could see much better. In a similar way (only more so), the unofficial economy and production that does not reach the market (like housework) can only be estimated very roughly.

More important than these and similar well known and recognised problems in the body of official statistics are new problems of relevance which will crop up as society develops. Official statistics will be required to measure new features of life that were not relevant before. The recent expansion of data production and availability may make this possible, but at the same time, opens the question of the intrinsic value of data and its dynamics. Information has become a valuable asset in its own right, so methods should be developed that measure its state and dynamics, as is the case with other assets. Answers to these requirements will probably lead to further demands for new statistics. Developments in this area may lead to new important challenges for official statistics.

More down to earth is the fate of the gross domestic product. As is the case with other popular, and therefore sensitive, indicators (CPI, unemployment), GDP is being interpreted and used in a much broader sense than its methodology justifies. In popular understanding it has all but replaced the gross national income as the ultimate measure of economic performance and is even generally used as a measure of wealth, prosperity and general public happiness. The misappropriation of GDP for these purposes is, of course, the result of a growing call for an indicator that would somehow quantify the vague feeling people have of some country or some time period being happier and nicer than another one. This feeling does not necessarily correspond to measures of economic performance, so the use of such measures (usually GDP) is clearly wrong. Attempts to define another, true, measure of happiness which could then be used to compare not only countries horizontally, but also country dynamics, have had only limited success. Whether statistics can ultimately satisfy this demand at all, is uncertain, even doubtful, but the demand for it is always with us and will remain a challenge for official statistics. It is an important challenge, if only to allow the GDP to drop its transcendental burden and return exclusively to its true and important, purely economic, sense.

The greatest number of challenges that confront official statistics in the area of relevance address, however, its ability to satisfy demands from users whose needs reflect changes in the world around them and who need to adjust their function to these new facts. The form and scope of these is impossible to predict. Yet, these requirements will grow with time and will demand constant re-evaluation of the body of statistics that are being produced. In order to be prepared to cope with these challenges, official statistics must continue to develop its profession and to open itself to ideas and initiatives from users and academia.

1.2 Education

The Czech Statistical Office has maintained close and fruitful links with the Faculty of Informatics and Statistics of the University of Economics and with other relevant university faculties, thus keeping up to date with statistical theory and knowledge, providing academic institutions with practical know-how and participating in the training of young statisticians. Its challenge in the area of education, however, extends further.

Official statistics has always been confronted with how many interpretations of its important results are counter-intuitive. Some of these apparent paradoxes are classical and appear as a warning in statistical textbooks; there are also several good monographs devoted to this phenomenon. The untrained human brain seems to be relatively bad at understanding the contribution of probability theory, and so untrained intuition leads often to mistaken conclusions.

While the requirement to counteract these misunderstandings has always been with us, this need has become more imperative in the last few decades. This is so because of at least three factors. Firstly, statistics have probably become more widely used in general political argumentation, thus giving more chance to mistaken or mendacious misinterpretations. Secondly, methodological development and harmonisation of statistics have led in many areas (unemployment, inflation) to the parallel publication of indicators describing the same phenomenon using different methodologies. While this is, of course, a positive development allowing to choose the most fitting model best suited to each immediate requirement, the parallel publication of figures describing the same phenomenon but differing in value is often confusing for the general reader, who might not understand the methodological differences. And thirdly, the general spread of data and information enabled by the recent advancement in information technology has given space to organised misinformation and mendacity, which can be supported by deliberate statistical falsifications constructed to feign scientific evidence that does not exist in reality.

The only reliable protection against toxic impacts of false information is raising the standard of general education in the relevant fields. This is clearly a major challenge for the powers that be in democratic countries in their fight against "fake news" and general misinformation. Official statistics should address this challenge as its own, and in cooperation with schools and other educational institutions look for ways of teaching the general public the basics of statistics. Reducing ignorance is the only way to counter the twin dangers of general misbelief in published statistics, and facile belief in all figures that appear to be "scientific".

The record of the Czech Statistical Office in the area of educating the public and popularising official statistics is actually very good. Apart from its approved statistical programme, it publishes many studies and monographs of topical subjects and, by far not least, publishes an attractive monthly magazine for the general public, which has won awards for its superb quality among similar publications by Czech public offices. General media have also helped by publishing some methodological detail of statistics they use. (While these details are often tedious, their regular appearance reinforces a general understanding of the dangers of misinterpretation.) All that acknowledged, there is still a long way to go before the immunity of the public against the misleading use of statistical data can be taken for granted.

1.3 Data protection and record linkage

Record linkage between surveys, and particularly between surveys and administrative data, even where technically and methodologically possible, was often rejected in the past century, because of its implications for data security. Official statistics and other public offices that collected information for official use adhered to strict promises required by specific laws not to allow the use of the data they had collected for any other purposes than those defined in the laws. This guarantee was seen as an important barrier to deliberate or accidental misuse of collected information. It also established full responsibility for this protection in each case with one public office only. Toward the end of the 20th century, with the growing amount of information collected by government offices, it became obvious that this arrangement is in many instances very wasteful, because it often required the same information to be submitted in different format by the same respondents to several government institutions. That led to legitimate protests by respondents and the public at large, who correctly believed that information once given to the government should be made available to all legitimate users by the government, which then should not come around asking for it again with another form under the heading of another department. The public position on this issue, while somewhat schizophrenic in the minds of official statisticians ("don´t ask me for the same information a second time, but never allow it to be used for any other purpose!") was very strong and, obviously, fully justifiable at a time of continuously increasing response burden.

In the end the obvious solution was found based on the arrangement where official statistics received in principle access to all government data while guaranteeing protection at least at the same level as provided by the other parties. The legal objections were also overcome due to the fact that protection of data in the Czech Statistical Office was already at least as strict as elsewhere and was seen to be so.

Recent technological development allows to build on the availability of data and to search for the application of big data methods to extract further information from data that already exist in the public or government domains. The development and application of such methods in the statistical programme is an interesting new methodological challenge for official statistics in the future.

1.4 Big data

New technological advances have been expanding the amounts of information that is available in the retail and production processes. This raises the possibility in the future for official statistics to get this information directly from the original source. If successful, this approach should eliminate the costs of reporting and thus do away with much of the response burden. The design and implementation of such systems is a major challenge for future official statistics and could in due course radically change the whole pattern of data collection.

The challenge of big data lies no longer in its technical aspect, but rather in overcoming problems of cooperation between official statistics and private business. It raises the problems of data protection to a new level and may bring along methodological concerns where it will need to combine information from different types of sources.

Czech official statistics has already started to address these issues in the collection of price data in retail, where the advantages are most obvious and where the data protection issue is least important. Broader applications will be much harder and constitute a major challenge for the future.

1.5 Register quality

Czech official statistics have long been challenged by the inadequate quality of registers maintained by other government offices. While this is basically a technical issue, it is very frustrating that efforts to minimise response burden and perform record linkages, which have finally become legal after much additional effort, are hampered by register quality and, generally, by difficulty in building common teams with other government institutions to make the registers fit for running important surveys. The prime example here is the decennial census which, as it turns out, again cannot be fully run from existing popular registers because of incompatibility or even poor quality, and so will again require the additional effort and cost that this engenders from both official statistics staff and the general population. The census is, of course, not the only instance. A great challenge for Czech official statistics remains, therefore, to facilitate and implement a working system of reliable registers together with other government departments.

1.6 Automatisms and changes in the decision-making process

There are several factors in the present world that cause the decision-making process to be more complicated than it used to be only a few decades ago. Firstly, amounts of data available to the decision makers and relevant to the issue at hand have increased significantly. Secondly, the number and sophistication of vested interests have multiplied, increasing the complexity of decision-making processes they try to influence. Thirdly, globalisation on the one hand and democracy on the other, while opening the world to many new endeavours and possibilities, require in many areas more rules and regulations, which in turn complicate the decision-making process. Fourthly, technological progress keeps extending these endeavours and possibilities, leading to further regulations in newly invented areas. Fifthly, the growing propensity to litigation seems to make matters worse yet. All these factors combine to make deciding much more onerous than before, both in the number of decisions required, and in the complexity of each of them.

In this situation it is reasonable to reduce the number of necessary decisions by grouping some of them, designing a rule that covers a series of decisions and allowing the rule to apply automatically, instead of deciding each time separately. Such a general decision can, of course, be changed once the situation changes (or exceptions may be approved), but such a change is elaborate and time consuming, so will often be avoided or postponed where and when possible.

This idea is actually very old and serves its purpose (avoidance of lengthy negotiations and litigation where the pattern of arguments is the same and where a general rule can be negotiated in the first place). One of the oldest and established examples is the indexing of payments to some measure of inflation, e.g. of rents or pensions to the consumer price index. Necessary negotiations no longer occur every year but are performed once for a longer period and then followed automatically. The parties to the negotiation effectively delegate their annual decision to an indicator produced according to a pre-defined methodology (often not fully understood by the parties but trusted to be neutral to their interests). The indicator is usually produced by official statistics.

This approach is becoming very common in the developed world and particularly in the European Union in many areas (e.g. the Maastricht criteria). Thus, the product of official statistics is in many cases becoming the arbiter of conflicting interests, rather than only a measure of some societal phenomenon. Almost by stealth, the character of the statistical product is being transformed.

The danger of this change has been recognised in the 1970´s, when Goodhart´s Law was formulated by several authors at about the same time. It may best be expressed as follows: *"an indicator used as a control ceases to be a reliable statistic"*. This had already been obvious to managers in the planned economy of communist countries, who found that the "gross material product" indicator suffered unexplainable distortions once it became the main control of factory performance used for the allocation of quarterly salary premiums. (Without the experience of communist economy, it took the authors of Goodhart´s Law a bit longer to formulate it.) The reason why Goodhart´s Law applies is clear: once the indicator has become a control, vested interests will find a way to satisfy the formal definition of the indicator with the least necessary cost. This is nearly always possible, because nearly all statistics are designed as models on the assumption of neutrality. In other words, once there is a strong enough incentive to find ways around the control, such a way will be found. As a result, the statistic might lose much of its relevance.

Official statistics tries to defend these indicators by systematic improvement of their definition (as can be seen in the history of GDP) or by inventing new indicators when old ones are compromised, all with mixed and diminishing success. The deterioration of the relevance of the affected statistics due to Goodhart´s Law is a growing challenge for official statistics in the future.

The greater danger following from the replacement of individual decisions by automatisms, however, is a shift in the object of manipulation by vested interests. Today in most instances, the pressure of lobby groups is still being applied to the actual decision makers, politicians or civil servants. This did not change with the early instances of automatisms, as (e.g.) in our example above, nobody would try to influence the method of computing the CPI instead of putting pressure on the parties to the general indexing rule. This, however, seems to be changing with the multiplication of fixed automatisms today. The more such rules are being established where the statistic itself decides money flows or other advantages without the direct participation of the original decision makers, the more practical it will be for vested interests to invest in manipulating the statistic-cum-control directly. While in some instances this pressure may result in simply circumventing the control by satisfying the indicator by other means (thereby "only" compromising the relevance of the indicator), in other cases it may be directed at the definition of the indicator. Official statistics will probably be the target of illicit pressure of this kind more often than in the past. Standing up to this pressure will be, in my opinion, the greatest challenge official statistics will meet in the near future.

2 RESPONSES

The Czech Statistical Office as the main representative of official statistics in the Czech Republic is mostly well prepared to meet the challenges of today and, hopefully, of the near future. Its long tradition of reliable service over more than a century has built a position of trust in the nation, which bodes well for the future.

Some of the challenges listed above have been recognised for some time and are being addressed already, albeit with mixed success. Many of them require more cooperation with other parts of government or public institutions. This is probably the area where improvement is most urgent. The full use of existing administrative data in the system for statistical purposes is still being hampered by incompatibilities of different kinds (particularly of registers), which are incomprehensible in the age of sharing information on the internet, social networks, blockchain systems and big data methods. Unfortunately, this problem cannot be resolved by official statistics alone, but official statistics must persevere in its efforts to overcome the "Chinese walls" between departments.

The problem of relevance in this rapidly changing world is common to the global statistical community and must mostly be pursued together with official statistics in the European Union and other developed nations. The Czech Statistical Office is already an integral part of this work.

While in some areas the Czech Statistical Office is a leader in popularising the function and role of official statistics, much still remains to be done in educating the general public to understand statistics, their strengths and limitations. In this time of ubiquitous proliferation of false information and fake news, the ability of the public to discern and reject misinformation disguised as serious statistics may become very important. The Czech Statistical Office should become a prime mover in providing this education.

By definition, the challenges that we are least prepared for are those we did not anticipate. In the rapidly developing world we live in, one must expect new challenges to appear regularly. It is therefore imperative that the Czech Statistical Office should actively monitor not only developments that have reached the point of implementation, but rather that it should be aware of developing ideas as soon as they emerge. In this area the Czech Statistical Office maintains fruitful contacts with statistical faculties at universities and some other statistical institutions. However, many new developments relevant for the future direction of statistics appear outside the statistical community in various governmental or private research and development bodies as well as other faculties of academia. To prepare for unexpected challenges, official statistics should extend their contacts with such institutions.

Most important and yet not very visible is the defence of the professional independence of official statistics and the quality and unbiased nature of its product in this period of growing pressures of vested interests. The Czech Statistical Office is relatively well prepared for this continuing challenge by legislation, professionality of its staff and leadership, its relatively elevated position in the government hierarchy, its past history and particularly by the trust of the people it serves. Whatever the future brings, this asset is the most important of all.

CONCLUSION

At its centenary Czech official statistics are well prepared for the times ahead of us, as far as we can see today. For future challenges awaiting it in the times ahead, it can build on its tradition and history and on the professional staff it has collected and trained. We may expect it to be as successful in its second century as it was in the first.

While maintaining the optimism this conclusion gives us reasons for, it is in my opinion important for official statistics to be aware of the growing danger to its essential and cherished independence caused by the practice of using statistics as controls in complex rules-based structures in the modern world. This is the more so, because the plethora of professional and technical challenges that need to be addressed, important in their own right, may obscure this greater jeopardy in the minds of our statistical community.

References

GOODHART, C. A. E. *Monetary Theory and Practice. The U.K. Experience.* London: Palgrave, 1984.

GOODHART, CH. Problems of Monetary Management. The U.K. Experience. In: COURAKIS, A. S. eds. *Inflation, Depression, and Economic Policy in the West*, 1981, pp. 111–146.

OUTRATA, E. Influence of Governance Issues on the Quality of Official Statistics. *Proceedings 59th ISI World Statistics Congress 25–30 August 2013*, Hong Kong, 2015 (Session IPS071), pp. 669–674.

STRATHERN, M. *Improving Ratings: Audit in the British University System.* European Review 5, 1997, pp. 305–321.

International Conference on *Economic Measurement 2019*

Václav Rybáček[1] | *Czech Statistical Office, Prague, Czech Republic*

In partnership with the UK Office for National Statistics (ONS), the Economic Statistics Centre of Excellence (ESCoE)[2] held its annual conference on Economic measurement at King´s College London from 8th to 10th May 2019.[3] The conference provided an exceptional opportunity to bring together data producers, users and scholars, to discuss thoroughly the topical issues. The conference was attended by nearly three hundred of scholars, researchers, students and compilers presenting and discussing new findings, experimental calculations and bringing new ideas.

After the welcome speech delivered by John Pullinger (ONS), the conference carried on with the contribution of Vasco Carvalho (University of Cambridge) on the linkage between micro- and macrodata within the framework of input-output tables stressing a potential of data to depict the impact of idiosyncratic shocks on aggregate volatility as well as to explain pattern of knowledge diffusion. The conference was further divided into a number of the theme-oriented parallel sessions.

However, it is worth starting with the panel discussion, which took place on the second day of the conference, and which targeted the key priorities for the new SNA-manual. At the very beginning of the session, chaired by Peter Van De Ven (OECD), three main priorities as defined by expert groups were specified – digitalization, globalisation, and sustainability of well-being. Beyond that, the discussants in the panel further highlighted the importance of the ownership definition, the location of economic activities, or valuation of human capital.

As the statistical community is well aware of the urgency of tackling these issues, the overall focus of the conference was shaped accordingly. In the section devoted to *Capital*, the current definition on asset boundary was challenged. Branding, innovative financial products, social and human capital and others forms of intangible assets constituting an important input into production function in any economy were suggested as important contributions to the stock of capital, however, currently unrecognized as they go beyond the current definition of assets. This issue will undoubtedly be further a subject of intensive debates in the statistical community.

Two sessions were devoted to GDP (*"GDP and beyond"*). Two key issues addressed were the relation of GDP to well-being of citizens and an extension of production boundary especially in reaction to the ongoing technological changes. One of the key contributions pointed to the technological progress bringing statistics to a situation in which *"free digital goods are everywhere, except the national accounts."* So-called digital age gave rise to completely new forms of intangible assets from which both consumers and companies benefit. As put as an example, consumers can give a licence to a company to use his personal

[1] Czech Statistical Office, Macroeconomic Statistics Section, Na padesatem 81, Prague 10, Czech Republic. Also Jan Evangelista Purkyne University in Usti nad Labem, Moskevska 54, Usti nad Labem, Czech Republic. E-mail: vaclav.rybacek@czso.cz.

[2] The Economic Statistic Centre of Excellence is an independent research centre operating in the United Kingdom which was established in 2017. Its mission is to promote collaboration between academia, the research community and data compilers.

[3] The webpage: <https://www.escoe.ac.uk/escoe-conference-economic-measurement-2019>.

data in exchange for digital service provided (advertising, information, etc.). Such, in principle, barter transactions pose a challenge for compilers to capture the value-creating process in a comprehensive way.

A very innovative approach to the compilation of GDP was presented by Kevin Fox from the University of South Wales. His contribution addressed the occurrence of new products and the inclusion of free goods, such as Facebook, in GDP. Concerning the deflation issue, the authors´ alternative approach is rooted in microeconomics theory, using the reservation prices of the previous products in Hicksian terms, i.e. the price at which consumers are willing to give up completely a previously existing product. Applying this new approach and by inclusion of the value of free products, the authors arrived at an alternative "GDP-B".[4]

Concerning the relation of GDP to well-being, there is an overall awareness that GDP is a measure of output and it does not necessarily provide users with the idea on the evolution of well-being. Knowing that, discussants argued rather in the favour of use of disposable income which does not cover all the factors influencing well-being, but it may be realistically considered as better approximation of well-being. In the framework of the discussion, the representative of the Statistics Netherland presented estimations on well-being in the Netherlands. The publication *"Monitor of well-being: a broader picture"* is publicly available on the webpage of Centraal Bureau voor de Statistiek <www.cbs.nl>.

Changing technological environment was discussed in two aspects, in general: firstly, the valuation of the product of information and communications technologies; secondly, the exploitation of newly occurring data sources. Concerning the former, Dr. David Nguyen from The National Institute of Economic and Social Research delivered interesting presentation *"Cloud computing and national accounting"*. This research deals with the effects of the cloud computing on macroeconomic statistics. Here we are referring to the companies such as Dropbox, Google Cloud Platfrom, Amazon Web service and many other providers who enabled to turn investment in fixed capital (e.g. servers) into operating expenses, i.e. into payment for service. This, technically speaking, detachment of computing processes and data itself influence many statistical areas like trade statistics, productivity estimations, not least price statistics as the quality of those services is dramatically evolving.[5]

Broad attention was paid to new data sources such as web scrapping, scanner data, measuring retail trade using card transactional data or use of VAT for short-term economic indicators. This subject is logically highly topical as changing technological environment noticeably extends the range of data sources usable for the compilation of macroeconomic statistics. In the context of this issue, Alberto Cavallo, the Associate Professor at Harvard Business School, presented his research on the price dynamics across countries and market segments and discusses the pros and cons of the use of scanner data.[6]

The last of the SNA-priorities not yet mentioned, i.e. globalisation, was obviously addressed by many presenters at least implicitly as this general trend affects practically all statistical areas. Among many others, in the section *Trade*, researchers presented their findings on the effect of the growing specialization and fragmentation of production process across countries, which has been also gaining a growing political interest. We could mention, among others, the study on the involvement of UK regions in the network of global value chains and the discussion of a potential impact of Brexit on particular regions in the UK.

To conclude, the Conference on Economic Measurement organized by the ESCoE creates an exceptional and indeed a unique platform where key players in the field of statistics can hold open debates on topical or even controversial statistical issues and exchange their views and experience. The conference, the selection of topics and lively debates have just demonstrated how economic life has been dramatically changing and statistics shall strive to keep up with times to offer figures of satisfactory explanatory power.

[4] The paper was published by the NBER and it is online available at: <https://www.nber.org/papers/w25695>.

[5] Discussion paper "Cloud computing and national accounts" is available on the webpage: <www.escoe.ac.uk>.

[6] Research papers are available on the webpage "The Billion Prices Project" <http://www.thebillionpricesproject.com>.

Recent Publications and Events

New publications of the Czech Statistical Office

Czech Republic in International Comparison. Prague: CZSO, 2019.
Česko 15 let v Evropské unii (Czechia 15 years in the EU). Prague: CZSO, 2019.
Development of the Information Society in the Czech Republic and Other Countries 2018. Prague: CZSO, 2018.
Export and Import Price Indices in the Czech Republic in 2018. Prague: CZSO, 2019.
Focus on Women, on Men 2018. Prague: CZSO, 2019.
Indicators of Social and Economic Development of the Czech Republic 2000–4th quarter 2018. Prague: CZSO, 2019.

Other selected publications

Analysis of Correlated Data with SAS and R. 4th Ed. London and New York: CRC Press, 2018.
BING, L. *Sufficient Dimension Reduction. Methods and Applications with R.* London and New York: CRC Press, 2018.
ČERNÁ, I, MÜLLER, D., ŠTERBOVÁ, L. *Foreign Direct Investment and Investment Policy.* Prague: Oeconomica, 2017.
EFROMOVICH, S. *Missing and Modified Data in Nonparametric Estimation. With R Examples.* London and New York: CRC Press, 2018.
EUROSTAT. *Harmonised Index of Consumer Prices (HICP). Methodological Manual. November 2018.* Luxembourg: Publication Office of the European Union, 2018.
FAGERLAND, M. W., LYDERSEN, S., LAAKE, P. *Statistical Analysis of Contingency Tables.* London and New York: CRC Press, 2017.
HRONOVÁ, S., SIXTA, J., FISCHER, J., HINDLS, R. *Národní účetnictví – od výroby k bohatství* (National accounts – from production to wealth). 1st Ed. Prague: C. H. Beck, 2019. ISBN 978-80-7400-738-5
LEY, C. AND VERDEBOUT, T. *Modern Directional Statistics.* London and New York: CRC Press, 2017.
ŘEZANKOVÁ, H, LÖSTER, T., ŠULC, Z. *Úvod do statistiky* (Introduction to Statistics). Prague: Oeconomica, 2019.
VLČKOVÁ, J. *Global Production Networks in Central European Countries: the Case of the Visegrad Group.* Prague: Oeconomica, 2017.

Conferences

The ***62nd ISI World Statistics Congress*** will take place in **Kuala Lumpur, Malaysia, from 18th to 23rd August 2019**. More information available at: <*http://www.isi2019.org*>.

The ***22nd AMSE 2019 Conference*** will be held in **Nižná, Slovakia, from 28th August to 1st September 2019**. The conference is held under the auspices of the President of the Czech Statistical Office and of the President of the Statistical Office of the Slovak Republic and is dedicated to the 100th anniversary of statistics in Czechoslovakia (*https://www.czso.cz/csu/czso/history_of_czech_statistics_after_1918*). More information available at: <*http://www.amse-conference.eu*>.

The ***27th Interdisciplinary Information Management Talks (IDIMT 2019 Conference)*** will be held in **Kutná Hora, Czech Republic, during 4–6 September 2018**. More information available at: <*http://idimt.org*>.

The ***13th International Days of Statistics and Economics (MSED 2019 Conference)*** will take place **in the University of Economics, Prague, Czech Republic, from 5th to 7th September 2019**. The conference is jointly organized by the Department of Statistics and Probability and the Department of Microeconomics, University of Economics, Prague, Czech Republic; Faculty of Economics, the Technical University of Košice, Slovakia; and the Ton Duc Thang University, Vietnam. The aim of the conference is to present and discuss current problems of statistics, demography, economics and management and their mutual interconnection. More information available at: <*https://msed.vse.cz*>.

The ***37th International Conference on Mathematical Methods in Economics (MME 2019)*** will be held **in České Budějovice, Czech Republic, during 11–13 September 2019**. The conference is traditional meeting of professionals from universities and businesses interested in the theory and applications of operations research and econometrics. More information available at: <*https://mme2019.ef.jcu.cz*>.

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

All views expressed in the journal of Statistika are those of the authors only and do not necessarily represent the views of the Czech Statistical Office, the staff, the Executive Board, the Editorial Board, or any associates of the journal of Statistika.

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