STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **97** (1) 2017

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

EXECUTIVE BOARD

MANAGING EDITOR

CONTENTS

ANALYSES

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 journal included (since 2015) in the citation database of peer-reviewed literature **Scopus**, in the **Web of Science** *Emerging Sources Citation Index* (*ESCI* of **Thomson Reuters**, 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

Input-Output Approach to Regional Employment

Jaroslav Sixta [1] | *University of Economics, Prague, Czech Republic*

Abstract

The paper deals with statistical data on regional employment that was constructed on the basis of regional input-output tables. Both regional input-output tables and product linked regional employment were constructed within the research project. This data fits well the purposes of detailed analysis of regional economy since the data is broken down by two-digits level of product classification (CZ-CPA). Employment is presented on the level of the regions (NUTS 2) of the Czech Republic for 2011. The paper briefly describes procedures allowing construction of regional employment by products and the links to data from official statistics. Some of analytical possibilities of data on regional employment are illustrated by simple input-output analysis with three scenarios. The regions are also tested for output and employment sensitivity by estimating multipliers and elasticities. The interpretation of obtained results including hierarchical clustering is provided. The paper also presents discussion about the use of regional input-output tables and regional employment in regional analyses for policy measures.

Keywords

Regional input-output tables, employment, input-output analysis

JEL code

C67, R11

INTRODUCTION

Input-Output Tables (IOTs) are regarded as a suitable tool for sophisticated economic analyses. Besides, they can be used for environmental or social analyses, as well. They allow for a wide range of scientific studies and analytical works conducted by university researchers, analysts or official authorities. IOTs are also used by the OECD for its statistical outcomes such as increasingly important Trade in Value Added (TIVA), see OECD (2016). Traditional input-output tables contain three quadrants with monetary values and some additional indicators, mainly employment and capital stocks. Symmetric Input-Output Tables for national economy are usually officially compiled every five years by official statistical authorities,[2] see Eurostat (2013). In the Czech Republic, the most popular form of IOTs are product by product tables complemented by employment by products. These tables provide a powerful tool for construction of economic models, predictions and analyses.[3]

Input-output models built at the national level can be disaggregated or even differently constructed at the regional level. For the Czech Republic, Regional Input-Output Tables (RIOTs) were constructed

[1] Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: sixta@vse.cz. Author is also working at the Czech Statistical Office, Na Padesátém 81, 100 82 Prague 10, Czech Republic.

[2] <http://apl.czso.cz/pll/rocenka/rocenkaout.dod_uziti?mylang=EN>.

[3] This is possible on annual basis only since quarterly input-output tables are not very common, see Marek et al. (2016).

for the year 2011 by the University of Economics,[4] see Sixta and Vltavská (2016). The tables are constructed as symmetrical for 82 product groups at basic prices in line with ESA 1995 methodology. There are only few countries that officially publish RIOTs, e.g. Finland (Piispalla, 1999), United States, Italy (Benvenuti et al., 1995), Spain (INE, 2010) and therefore RIOTs belong mainly to research agenda. The construction of RIOTs and its possibilities were introduced by Kahoun and Sixta (2013).

This paper is particularly aimed at the discussion of employment data on the regional level. Such employment data are linked to products within the boundary of national accounts and should be used as auxiliary indicators to RIOTs. The data originates in academic research and therefore the construction of regional employment is presented at first. Brief description of the methods used for transformation of officially published data into product based data is included, as well. Data on regional employment is broken down by 82 products (adjusted CZ-CPA) and it comes solely from described computations. The most important advantage of this data is that they combine usability for economic modelling and protection of individual data since it is not linked directly to companies. The paper also brings a brief illustration of the possibilities of such data and provides analysis based on the elasticities and multipliers for three scenarios of external shocks. Such sensitivity analysis is used for the presentation of some of the possibilities with the help of the simple static input-output analysis. The first part of the analysis is aimed at the industry with the highest regional output, the second part at the industry of maximum regional employment, the third at construction industry. Finally, the purpose of the paper is also to promote the use of more advanced models based on freely accessible IOTs and RIOTs since they can provide interesting feedback to policymakers.

1 REVIEW OF LITERATURE

Researchers dealing with input-output tables can find plenty of more or less recent scientific literature dealing with regional input-output tables and regional input-output analysis. One of the most relevant information sources offers the publication *Input-output analysis: foundations and extensions* by Miller and Blair (2009). Basic categories and models, mainly inter-regional and multiregional models are deeply described. From the theoretical description to practical compilation issues is a far way. In some countries, regional input-output tables are from time to time available. In Europe, RIOTs are very well described mainly for Finland, Spain and Netherlands. Compilation issues of RIOTs for Finland are described in Louhela and Koutaniemi (2006), Dutch case was introduced in Eding et al. (1999) and Spanish in INE (2010). There were also some scientific works dealing with the case of the Czech Republic, e.g. applying of GRIT method on Czech data, see Semerák et al. (2010). Different approach based on location coefficients is described in Flegg and Tohmo (2013) for the case of Finland.

Input-output researches and fans usually contribute to the journal Economic Systems Research specifically aimed at this area. From this field, recent and closely related paper dealing with regional input-output tables was written by Többen and Konenberg (2015). Besides scientific works dealing with sub-national regional input-output tables, specialised research agenda aimed at the group of regions or countries can be found. There are at least three big databases, EORA (see Lenzen et al., 2012), WIOD[5] or EXIOBASE[6] that are covering specific input-output related information or extensions. Besides regional product flows, distribution effects should not be omitted as well since they can be linked with RIOTs.[7]

[4] <http://kest.vse.cz/veda-a-vyzkum/vysledky-vedecke-cinnosti/regionalizace-odhadu-hrubeho-domaciho-produktu-vydajovou-metodou>.
[5] <http://www.wiod.org/new_site/home.htm>.
[6] <http://www.exiobase.eu/index.php>.
[7] Some basic of linking of product flows with distribution of income can be found in Šimková and Langhamrová (2015).

Regional input-output tables or regional input-output analysis is also tackled for price level measurement since regional structures (Kramulová and Musil, 2013) are often used for these computations, see Čadil et al. (2014). Analytical potential of input-output analysis on both regional and national level is considerable. From this field, discussion about regional economic development and ways of planning can be found in Stimson et al (2006). Regional input-output analysis is also useful for assessing the economic side of organising of cultural events (Raabova, 2010).[8]

In the context of the level of European Union, the case of regional accounts is not dramatically emphasised and regional input-output tables are out of the focus. That is a pity because academic environment can provide methodology or research papers and studies but can hardly provide necessary financial resources. Regular compilation of RIOT done by official statistics is possible but it would have to be supported by European regulation. Otherwise statistical office will not allocate enough resources for this agenda.

2 METHODOLOGY

Data on regional employment expressed in persons, hours worked and full time equivalent areregularly published by the Czech Statistical Office[9] on annual basis. These figures are published on the level of sections of industries (CZ-NACE). Data useful for IO models has to be in the same classification and dimension as regional input-output tables. Estimated regional employment as an additional indicator for input-output analysis is broken down by NUTS 3 level.[10] Therefore, the transition from industries (CZ-NACE) to products (CZ-CPA) has to be done. It is a similar procedure as transformation of use table into symmetric input-output tables. Nevertheless at first, the data has to be prepared. For the purposes of this paper, data on employment in persons was selected. The reason is the simplicity of data and expected higher quality data on regional level. Of course, full time equivalent data offers more appropriate picture of economy (see Fischer and Sixta, 2009) but for the regional comparison such data would be sufficient. Data on hours worked and full time equivalent may be transformed in the same way.

The first step includes the split of published data on regional employment by sections of CZ-NACE into 82 industries in line with national accounts figures. The key assumption is that the gross value added (a) per worker (l) is identical within each section of CZ-NACE across all regions. It means that the first estimate of employment in region r is obtained as:

$$l^r_{i,s} = \frac{a_{i,s}}{l_{i,s}} l^r_s , \tag{1}$$

where:

- s section of CZ-NACE,
- i industry within the section s,
- a gross value added,
- l employment,
- r region.

The next step consists of computation of the difference between the sum of regional employment by sections and the sum of employment by industries (82). The following conditions should be set. The sum of regional figures across all industries must correspond to national totals (2) and the sum of figures by industries within each section of CZ-NACE must correspond to regional figures (3):

[8] More recent research papers can be found at: <http://www.idu.cz/media/document/multiplikacni-efekty-2010.pdf>.
[9] <http://apl.czso.cz/pll/rocenka/rocenka.indexnu_reg>.
[10] The Czech official name for this regional level is "Kraj".

$$l_i = \sum_r l_i^r . \tag{2}$$

$$l_s^r = \sum_s l_{s,i} . \tag{3}$$

Solving formulas (2) a (3) was done by iterative method RAS (see Vavrla and Rojíček, 2006) in two rounds with original RAS method for two constraints (rows and columns of the table). Resulting matrix $\mathbf{L^I}$ where rows correspond to the regions and columns to industries was used as source data for transformation. In each region, product technology was used to transformed industry based data for product based data (method A described in IO Manual, see Eurostat, 2008):

$$l_p^r = l_i^r (V')^{-1} \hat{q} , \tag{4}$$

where:

l vector of employment by products in region r by products (p) of industries (i),
V output matrix (product, industries),
$\hat{q}$ diagonal matrix of output.

Table 1 Regional employment by products, 2011, CZK mil.

Region	Total	Agri-culture	Mining Manu-facturing	Construc-tion	Trade Transport Hotels	Informa-tion activities	Banking Insurance	Real estate	Services for compa-nies	Public adminis-tration Education Health	Other services
	T	A	B+C+D+E	F	G+H+I	J	K	L	M+N	O+P+Q	R+S+T+U
	5 043 438	159 221	1 358 862	508 623	1 275 415	126 966	91 070	32 235	444 488	866 364	180 194
Pha	899 746	3 634	51 853	81 455	284 954	67 047	43 586	23 424	156 794	144 762	42 237
Stc	551 394	25 547	164 943	52 732	157 213	4 829	4 389	637	37 365	87 853	15 886
Jhc	299 676	17 847	85 902	33 265	75 814	2 367	3 822	733	17 749	50 558	11 619
Plz	277 778	12 241	93 446	24 097	62 062	3 680	3 311	518	22 729	46 402	9 292
Kar	141 032	2 810	40 574	13 140	40 612	435	1 334	75	8 461	28 617	4 974
Ust	352 404	8 478	93 774	46 734	83 792	2 402	3 703	904	27 889	71 659	13 069
Lib	193 887	3 614	74 176	17 747	43 660	1 883	2 625	511	10 144	32 763	6 764
Krh	251 777	11 112	81 903	22 178	59 587	2 496	3 169	170	14 288	46 895	9 979
Par	235 667	12 820	85 271	21 676	53 668	2 268	3 503	539	12 877	36 734	6 311
Vys	224 014	18 677	89 013	23 450	39 328	2 029	1 646	598	7 190	36 994	5 089
Jhm	553 654	16 621	139 861	59 796	135 499	21 161	8 766	1 436	53 847	93 554	23 113
Olm	270 893	11 592	93 849	26 191	57 338	3 510	2 876	630	15 653	52 124	7 130
Zln	263 414	6 154	108 128	28 688	57 800	3 023	2 054	638	12 305	39 156	5 468
Mrs	528 102	8 074	156 169	57 474	124 088	9 836	6 286	1 422	47 197	98 293	19 263

Note: Names of CZ-CPA codes were shortened, official names can be found at: <https://www.czso.cz/csu/czso/klasifikace-produkce-cz-cpa>. Full names of regions can be found in the Annex.
Source: Own computation

Employment broken down by product classification is an important part of input-output analysis. The figures correspond to the number of persons needed for production of a particular product. Fourteen resulting vectors obtained by formula (4) are arranged in a matrix $\mathbf{L^p}$ with dimension (14 regions, 82 products). Aggregated figures are described in Table (1). Obtained technical coefficients, defined as number of employed persons divided by output can be downloaded from: <kest.vse.cz>.[11] The elements of matrix **E** (regions x products) were obtained as:

$$e_j^r = l_j^r / x_j^r , \tag{5}$$

where:

- e technical coefficient of employment,
- x output,
- j index pro product.
- r region.

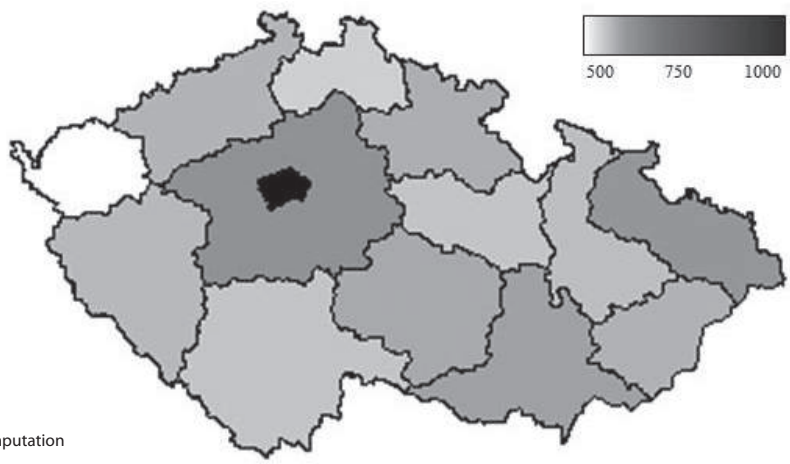
3 ANALYSIS OF REGIONAL EMPLOYMENT

Regional employment corresponds to the size of the region and these figures provide informative value only in connection with production or value added. The following Figure 1 presents regional map of gross value added per worker employed in the region; productivity of employment at current prices (h). For computational purposes, the figures were rescaled (standardized) to be presented in the form of map, see formula (6):

$$h_{st}^r = \left[h^r - \min(h) \right] / \left[\max(h) - \min(h) \right] . \tag{6}$$

The highest productivity is in the capital city of Prague, over CZK 0.9 mil. In three regions (Středočeský, Jihomoravský and Moravskoslezský kraj), the productivity exceeds CZK 640 thousand. In five regions productivity still exceeds CZK 600 thousand and in four regions CZK 560 thousands. The weakest region, Karlovarský kraj, has the productivity of employment on the level of 500 thousands CZK. In comparison with Prague, it is just about a half.

Figure 1 Regional productivity of employment, 2011, thousand CZK per person

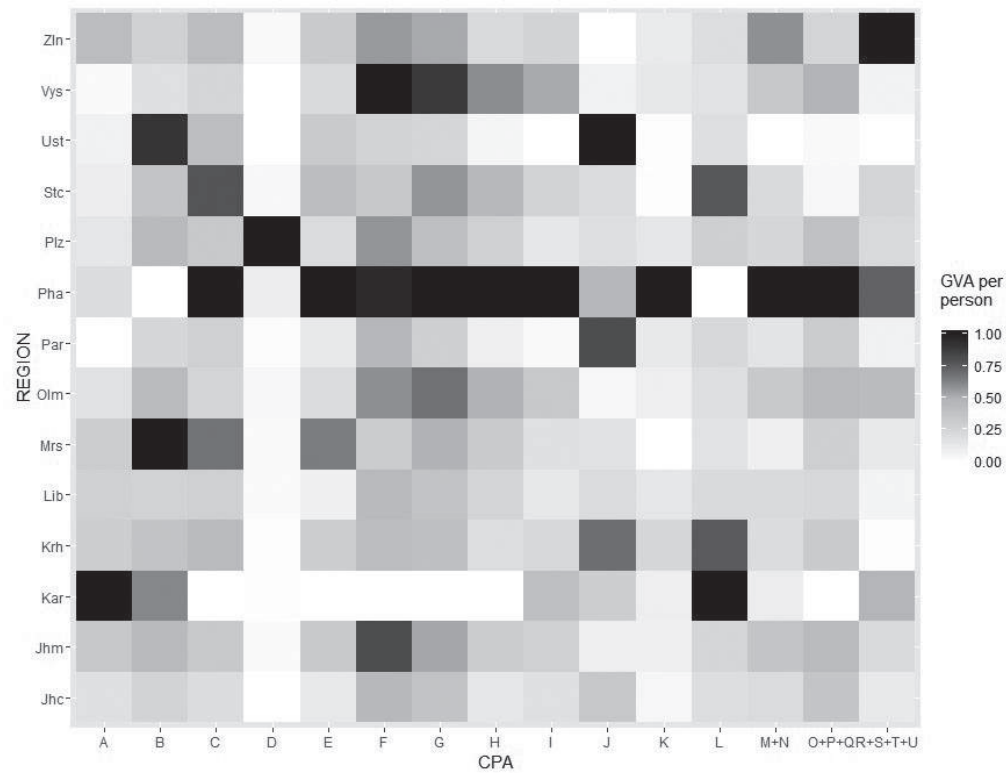


Source: Own computation

[11] The link is active from 1.12.2016.

Regional industrial specifics can be clearly presented on the heat map for emphasizing the differences. Again, heat map was constructed from standardized data (see formula 6). Prague as a capital city has the highest employment productivity for most of products. On the contrary, some regions are very specific. Středočeský kraj is very much linked to manufacturing, trade and real estate products. The region with the lowest employment productivity, Karlovarský kraj, has the highest productivity in agriculture and forestry and real estate products. For all other products, the productivity of this region is very weak. There are two regions without specifically high productivity, Jihočeský and Liberecký kraj. Although Ústecký and Moravskoslezský kraj are connected mainly with mining products, automotive industry in Moravskoslezský kraj creates high value added per worker. Interesting figures can be found also for Zlínský kraj, where other services (recreation, social and other) creates relatively high values of gross value added with relatively low employment and so high productivity of employment (Figure 2).

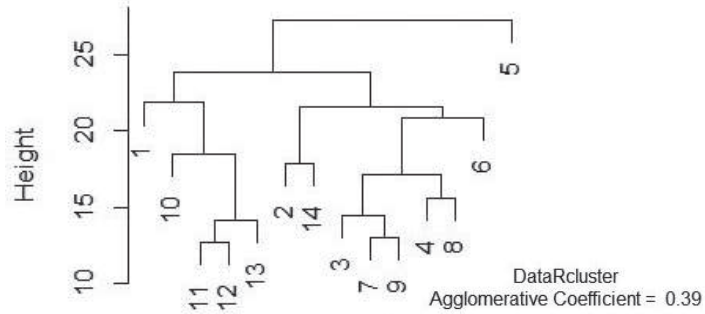
Figure 2 Heat map of regional employment productivity, 2011



Note: The scale for a heat map is based on standardized data, where 0 is given by the minimum and 1 by the maximum value.
Source: Own computation

Similar information is found when using simple cluster analysis based on Euclidean distance. There are some regions that are very different from others, mainly Karovarský kraj and the capital city of Prague. Unsurprisingly, similarity is found between Středočeský and Moravskoslezský kraj (mainly due to automotive industry) and between Jihomoravský and Olomoucký kraj. Close similarity was also identified between Liberecký and Pardubický kraj. Finally, there are found two big clusters and one very distant region, Karlovarský kraj. Clustering of employment productivity was not very successful since agglomeration coefficient is very low (0.39), see Figure 3.

Figure 3 Dendrogram of the clusters of Czech regions, 2011



Note: Regions codes refer the order used in the Table 1 or see the Annex.
Source: Own computation

4 TESTING OF REGIONAL EMPLOYMENT SENSITIVITY

Since regional employment and regional specialization is often discussed, I prepared a simple test related to regional employment. Such discussions take place on both national and regional level, regarding mainly the sensitivity of the Czech Republic on changes in automotive industry and final demand for cars. With respect to the specialization of regions mentioned above, it can be expected that each region has its own sensitivity on external shocks. Regional input-output analysis on preliminary results and selected regions was also provided in Sixta and Fischer (2015).

With respect to the specific output of the region, regional economy would react on external incentives (shocks) differently. From this perspective, it crucially depends on regional employment productivity, the mixture of imported and produced products and regional capacities. Testing regional sensitivity of employment is based on computation of elasticities of employment from simple static input-output analysis with all the limitations and assumptions (e.g. free capacities), see Leontief (1986). Three scenarios were selected to illustrate sensitivity of regional employment. The first scenario consists of external shock in the industry with the highest regional output (SC 1). The second scenario consists of external shock in the industry of maximum regional employment (SC 2). The third scenario represents external shock

Table 2 Selected products for scenarios

Nb.	Region	Scenario 1	Scenario 2	Scenario 3
0	CZ	29	46+47	41_42_43
1	Pha	46+47	46+47	41_42_43
2	Stc	29	46+47	41_42_43
3	Jhc	35	46+47	41_42_43
4	Plz	26	46+47	41_42_43
5	Kar	86	46+47	41_42_43
6	Ust	19	46+47	41_42_43
7	Lib	29	46+47	41_42_43
8	Krh	29	46+47	41_42_43
9	Par	26	46+47	41_42_43
10	Vys	35	01	41_42_43
11	Jhm	46+47	46+47	41_42_43
12	Olm	46+47	46+47	41_42_43
13	Zln	22	22	41_42_43
14	Mrs	29	46+47	41_42_43

Source: Own computation

in construction industry (SC 3). The first two scenarios are specifics for each region. On the contrary, the third scenario was selected to explain sensitivity on the same product. Following Table 2 shows selected products (on the level of two digits CZ-CPA) for all three scenarios in the regions including the Czech Republic.

The external shock is modelled simply in line with traditional statistic input-output analysis, see formula (7). It is used for illustrative purposes only. The change of output vector (x) is derived from the change of vector of final use (y). The final impact on employment is measured by the share of employment in output, see formula (7) and scalar multiplication (8).

$$\Delta x = (I - A)^{-1} \Delta y \,, \tag{7}$$

$$\Delta l = e^{r} \cdot \Delta x \,, \tag{8}$$

where:

l vector of employment by products,
x output matrix (product, industries),
y vector of final use,
e^{r} vector of technical coefficients of employment for region r.

For computation purposes, the change of final use counted 20 CZK bn. or 10 CZK bn. depending on the size of industry in a particular region. The comparison is based on relative figures, simple elasticizes of employment were calculated as:

$$\varepsilon^{r} = \frac{\sum_{i} \Delta l_{i,t}}{\sum_{i} l_{i,t}} \Bigg/ \frac{\sum_{i} \Delta y_{i,t}}{\sum_{i} y_{i,t}} \,. \tag{9}$$

Besides employment elasticities, output elasticities were computed as well. Computed elasticities cover both direct and indirect effects. It means that direct effects can be observed in the affected products (industry) by external shock (e.g. construction in case of the decrease of government investment into public infrastructure). Indirect effects can be identified in transport, trade, construction materials, etc. The overview of all three scenarios is shown in Table 3, data are recalculated for the change of final use by one CZK mil.

Table 3 Impact of the change in final use on output and employment

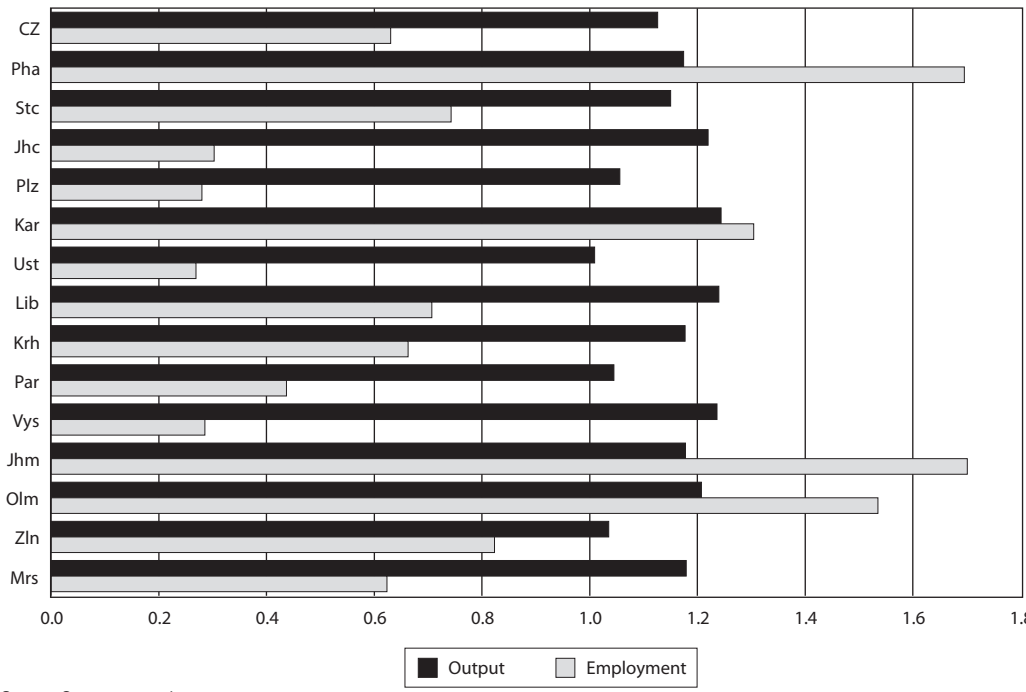
	SC1		SC2		SC3	
	Δx	Δl	Δx	Δl	Δx	Δl
CR	1.51	0.44	1.60	1.25	2.04	1.13
Pha	1.84	1.00	1.84	1.00	2.19	0.77
Stc	1.46	0.43	1.48	1.25	1.93	1.18
Jhc	1.63	0.26	1.47	1.43	1.96	1.31
Plz	1.37	0.22	1.51	1.38	2.02	1.26
Kar	1.32	1.13	1.44	1.58	1.80	1.46
Ust	1.28	0.18	1.48	1.42	1.87	1.39
Lib	1.52	0.56	1.42	1.43	1.99	1.29
Krh	1.51	0.50	1.44	1.40	1.97	1.30
Par	1.29	0.28	1.42	1.43	1.97	1.30
Vys	1.57	0.21	1.59	1.37	1.92	0.98
Jhm	1.64	1.39	1.64	1.39	2.23	1.16
Olm	1.49	1.28	1.49	1.28	2.00	1.27
Zln	1.26	0.59	1.26	059	1.95	1.24

Source: Own computation

Scenario 1

When comparing resulting elasticities for output and employment, very different results are obtained. It is caused by the difference between output in monetary values and number ofworkers necessary for the production, see Figure 4. Only in two regions the difference between output and employment elasticity is low (Karlovarský and Zlínský). On the level of the Czech Republic, the highest output is observed in automotive industry (29) and the elasticity of output is about 1.13 and elasticity of employment only 0.63. It means that the change of final use by one percent leads to the increase of overall output by 1.13% and the increase of employment by 0.63%. In nominal terms it means that additional CZK one million spent in final demand for products of automotive industry leads to CZK 1.5 million of output and 0.4 workers needed, see Table 3. The lowest employment elasticities are observed for regions oriented for energy products, Vysočina, Jihočeský, Ústecký kraj (0.29, 0.3, 0.27) and Plzeňský kraj (0.28) with high share of computers production. The highest employment elasticities are observed for Praha and Jihomoravský kraj, both connected with trade activities (46+47), reaching 1.7.

Figure 4 Elasticities for Scenario 1

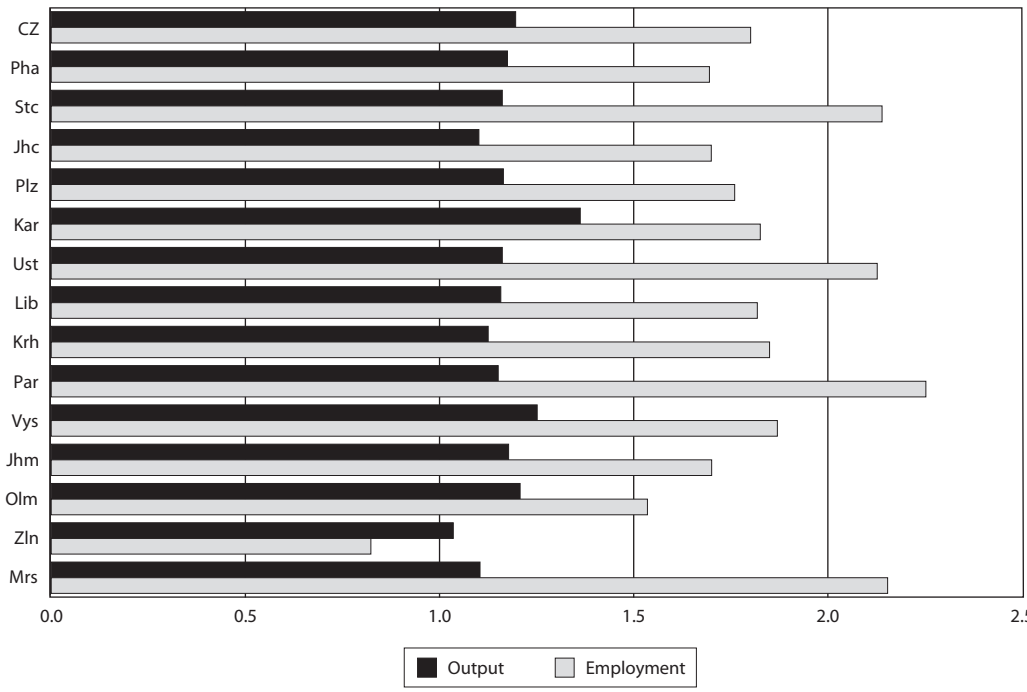


Source: Own computation

Scenario 2

When focusing on products (industries) with the highest employment, stable differences between output and employment elasticities are observed, see Figure 5. Only Zlínský kraj has the lowest employment elasticity, only about 0.823. In nominal terms it means that additional one million CZK creates 0.6 jobs. In all other regions, the elasticity of employment is higher than 1 with maximum about 2.25 in Pardubický kraj. In 12 regions the highest employment is connected with trade products (industries) (46+47), only in Vysočina and Zlínský kraj the most important products (industries) are different. Vysočina is oriented to agriculture and forestry (01) and Zlínský kraj to rubber and plastic products (22).

Figure 5 Elasticities for Scenario 2



Source: Own computation

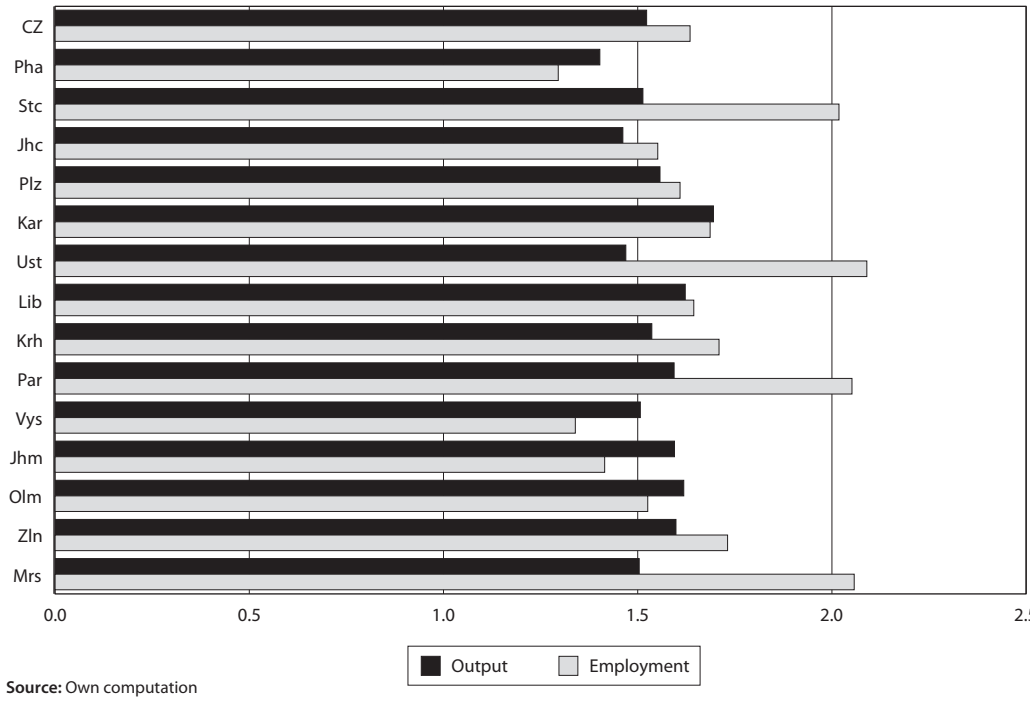
Scenario 3

The third scenario describes the situation when final demand for construction services is increased. This can be reached either by private or public expenditures. It also refers to the frequently discussed role of construction of public infrastructure for the reduction of regional structural unemployment. In this case, the output elasticities do not provide adequate information since construction industry is very interdependent. There are many elements of production chain before completed product (e.g. building or structure) finds its final customer. The analysis of employment elasticities shows that in some regions the possibility of influencing the employment by induced construction works is very limited, see Figure 6. For example, in the capital city of Prague, employment elasticity is very low, additional CZK 1 mil. creates only 0.8 jobs. The elasticity reaches only 1.3 that is the minimum of all regions while the average elasticity for the Czech Republic is 1.63. On the contrary, the same amount invested in Ústecký kraj leads to the increase of jobs by 1.4 with maximum regional elasticity of 2.09. The second lowest values of elasticities are found for Vysočina (1.34) even though the productivity of employment of construction services is relatively high but there is a big share of goods and services imported to construction industry. In all cases it means that incentives aimed at regional employment do not provide adequate response if the regional economy is not equipped by suitable free capacities.

CONCLUSION

The paper provided the information about data on regional employment linked to products within national accounts' framework. The methodology of construction of regional employment data is described including the links to officially published data. Besides, analytical possibilities of this data are presented. Regional data was constructed for 14 regions of the Czech Republic and broken down by product

Figure 6 Elasticities for Scenario 3



Source: Own computation

classification (CZ-CPA) counting 82 categories. Employment by products is an important analytical indicator to be used mainly for input-output based economic models. Such data is not officially published in the Czech Republic and therefore they are based on academic research that follows previous work in this area, see (Sixta and Vltavská, 2016).

Even though presented figures are prepared for the year 2011 and based on ESA 1995 (SNA 1993) methodology, their usage is not significantly affected. The foundations of input-output analysis lie in the structures and relation of costs and production, i.e. output in national accounts methodology. The methodology of construction was developed for the condition of Czech statistics where some principles relating to the breakdown of statistical units in national accounts are not fully implemented. It refers mainly to the definition of local kind of activity unit and widely used principal activity approach and local units. Anyway, the methodology is transferrable to other cases and can be used for other countries.

Analysis of regional employment allows for identifying regional specifics and their links to other regions. Presented heat map shows the significance of employment in different groups of products (homogenous industries). Regions were also compared from the perspective of their distance with simple cluster analysis. The specifics of the capital city of Prague were illustrated. Such information was used for simple static input-output analysis and computation of regional employment elasticities (multipliers) describing the sensitivity of employment. The analysis was presented in three scenarios for selected final demand shocks. It should provide information about the possibility of affecting regional employment.

Regional input-output tables including auxiliary indicators such as employment should serve for regional economic modelling. Optimal regional economic policy should tackle specific regional problems like structural unemployment. Statistical data serve as basis for such analyses and their availability and quality significantly set the limits to their users. Despite high demand for detailed regional data, European official statistics do not provide adequate amount of information for construction of advanced

regional models. Our research data covering regional input-output tables and regional employment can be downloaded from kest.vse.cz/ and they are intended to be exploited in users' analyses.

References

BENVENUTI, S. C., MARTELLATO, D., RAFFAELLI, C. INTEREG: A Twenty-Region Input-Output Model for Italy. *Economic Systems Research*, 1995, 7, pp. 101–116.

ČADIL, J., MAZOUCH. P., MUSIL. P., KRAMULOVÁ, J. True regional purchasing power: evidence from the Czech Republic. *Post-Communist Economies*, 2014, 26, pp. 241–256.

EDING, G., OOSTERHAVEN. J., VET DE, B., NIJMEIJER, H. *Constructing Regional Supply and Use Tables: Dutch Experiences.* Understanding and Interpreting Economic Structure, Berlin: Springer Verlag, 1999, pp. 237–263.

EUROSTAT. *European System of Accounts – ESA 1995*. Luxembourg: Office for Official Publications of the European Communities, 1996.

EUROSTAT. *Eurostat Manual of Supply. Use and Input-Output Tables*. Luxembourg: Office for Official Publications of the European Communities, 2008.

EUROSTAT. *European System of Accounts – ESA 2010*. Luxembourg: Office for Official Publications of the European Communities, 2013.

FISCHER, J. AND SIXTA, J. K propočtu souhrnné produktivity faktorů. *Politická ekonomie,* 2009, Vol. 57, Iss. 4, pp. 544–554.

FLEGG, A. T. AND TOHMO, T. Regional Input-Output Tables and the FLQ Formula: A Case Study of Finland. *Regional Studies*, 2013, 47, pp. 703–721.

INE. S*panish Regional Accounts. Base 2010.* Madrid: Instituto Nacional de Estadística, 2010.

KAHOUN, J. AND SIXTA, J. Regional GDP Compilation: Production, Income and Expenditure Approach [online]. *Statistika: Statistics and Economy Journal*, 2013, 4, pp. 24–36.

KRAMULOVÁ, J. AND MUSIL, P. Experimentální odhad výdajové metody regionálního HDP. *Politická ekonomie*, 2013, 61, pp. 814–833.

LENZEN, M., KANEMOTO, K., MORAN, D., GESCHKE, A. Mapping the Structure of the World Economy. *Environmental Science & Technology,* 2012, 46(15), pp. 8374–8381.

LEONTIEF, W. *Input-Output Economics*. Oxford University Press, 1986.

LOUHELA, T., KOUTANIEMI, M. *Construction of regional input-output tables in Finland 2002* [online]. In: 46th Congress of the European Regional Science Association (ERSA), Greece, 30.8.–3.9.2006. [cit. 20.8.2016]. <http://www-sre.wu-wien.ac.at/ersa/ersaconfs/ersa06/papers/110.pdf>.

MAREK, L., HRONOVÁ, S., HINDLS, R. Příspěvek k časnějším odhadům hodnot čtvrtletních národních účtů. *Politická ekonomie,* 2016, 64, pp. 633–650.

MILLER, R. E. AND BLAIR, P. D. *Input-output analysis: foundations and extensions*. Cambridge University Press, 2009.

OECD. *Trade in Value-Added: Concepts, Methodologies and Challenges* [online]. Paris: Organisation for Economic Cooperation and Development, 2015. [cit. 4.9.2016]. <http://www.oecd.org/sti/ind/49894138.pdf>.

PIISPALA, J. Constructing Regional Supply and Use Tables in Finland. In: *European Regional Science Association (ERSA)*, 39th European Congress, Ireland, 23.–27.9.1999.

RAABOVÁ, T. *Possible Methodsf for Measuring Economic Impacts of Cultural Tourism* [online]. Prague: Arts and Theatre Institute, 2010. [cit. 15.9.2016]. <http://www.idu.cz/media/document/tereza-raabova_possible-methods-for-measuring-economic-impacts-of-cultural-tourism.pdf>.

SEMERÁK, V., ZIGIC, K., LOIZOU, E., GOLEMANOVA-KUHAROVA, A. *Regional Input-Output Analysis: Application on Rural Regions in Germany, the Czech Republic and Greece*. In: 118th seminar of the EAAE (European Association of Agricultural Economists), 'Rural development: governance policy design and delivery' Ljubljana, Slovenia, 25.–27.8.2010.

SIXTA, J. AND VLTAVSKÁ, K. Regional Input-output Tables: Practical Aspects of its Compilation for the Regions of the Czech Republic. *Ekonomický časopis,* 2016, 64, pp. 56–69.

SIXTA, J. AND FISCHER, J. Regional Input-Output Models: Assessment of the Impact of Investment in Infrastructure on the Regional Economy. In: *Mathematical Methods in Economics 2015* [online]. Cheb, 9.–11.9.2015, pp. 719–724. [cit. 20.8.2016]. <http://mme2015.zcu.cz/downloads/MME_2015_procee dings.pdf.>.

STIMSON, R. J., STOUGH, R., ROBERTS, B. H. *Regional economic development: analysis and planning strategy*. Berlin: Springer, 2006.

ŠIMKOVÁ, M. AND LANGHAMROVÁ, J. Remittances and their Impact for the Czech Economy. *Prague economic papers*, 2015, Vol. 24, Iss. 5, pp. 562–580.

TÖBBEN, J. AND KONENBERG, T. H. Construction of Multi-Regional Input-Output Tables Using the Charm Method. *Economic Systems Research*, 2015, 27, pp. 487–507.

UNITED NATIONS. *System of National Accounts 1993 – SNA 1993*. New York: United Nations, 1993.

VAVRLA, L. AND ROJÍČEK, M. Sestavování symetrických input-output tabulek a jejich aplikace. *Statistika,* 2006, 1, pp. 28–43.

ANNEX

List of the regions

Number	Short Name	Name	CZ-NUTS
0	CZ	Česká republika	CZ0
1	Pha	Hlavní město Praha	CZ010
2	Stc	Středočeský kraj	CZ020
3	Jhc	Jihočeský kraj	CZ031
4	Plz	Plzeňský kraj	CZ032
5	Kar	Karlovarský kraj	CZ041
6	Ust	Ústecký kraj	CZ042
7	Lib	Liberecký kraj	CZ051
8	Krh	Královehradecký kraj	CZ052
9	Par	Pardubický kraj	CZ053
10	Vys	Vysočina	CZ063
11	Jhm	Jihomoravský kraj	CZ064
12	Olm	Olomoucký kraj	CZ071
13	Zln	Zlínský kraj	CZ072
14	Mrs	Moravskoslezský kraj	CZ080

Illustration of Single-Regional and Inter-Regional Approach in Regional Input-Output Analysis

Karel Šafr[1] | *University of Economics, Prague, Czech Republic*
Kristýna Vltavská[2] | *University of Economics, Prague, Czech Republic*

Abstract

Analytical works usually use single-regional approach which does not demand so much data. However, this approach disregards flows of output among regions. This leads to a misrepresentation of results which can be eliminated by using Inter-regional input-output model that requires more data to be employed. This paper illustrates the differences between the two different approaches of regional input-output model construction and their results. We construct inter-regional and single-regional models for all 14 regions of the Czech Republic and with 82 products according to the Classification of Products CZ-CPA. The results are compared on the level of Leontief's matrix and multipliers. We use graphical illustrations to depict the systematicness of differences. The single-regional approach proves a systematic undervaluation of specific products and regions contrary to other regions. The graphical analysis shows the significance of the connection among regions. This illustrates the disadvantage of the single regional approach. Finally, the results confirm the idea of a significant analytical misrepresentation of impacts modelled by this approach in the case of data for the Czech Republic.

Keywords

Regional Input-Output Tables, Input-Output analysis, Leontief`s multipliers, IRIO

JEL code

C67, R13, E21

INTRODUCTION

Regional input-output analysis represents a detailed tool of economic analysis on the sub-national level. Contrary to input-output analysis (IOA), on the national level the regional IOA offers detailed information on the exact structure of impacts. An advantage of the regional IOA lies in an accurate evaluation of effects in individual regions and products. The regional analysis of national policies in context of environment (Miller and Blair, 2009) represents the most common analysis. The detailed output of IOA actually enables a connection to the environmental matrix (Suttinon et al., 2013).

[1] Dept. of Economic Statistics, Faculty of Informatics and Statistics, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. Corresponding author: email: karel.safr@vse.cz. Author is also working at the Czech Statistical Office, Na Padesátém 81, 100 82 Prague 10, Czech Republic.
[2] Dept. of Economic Statistics, Faculty of Informatics and Statistics, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic.

The regional IOA divides according to two categories of models. One category represents models based on one region with no connection to other regions while the other category comprises inter-regional analysis where researchers simultaneously consider export and import to other regions (Miller and Blair, 2009). If we disregard the connection to other regions, we can use the single-regional input-output model (SRIO). For a comparison, we use the inter-regional input-output model (IRIO). The inconsistency of SRIO and IRIO ties to a so-called problem of aggregation of regional input-output tables. When using SRIO, the aggregation of RIOTs does not lead to national IOT (Crown, 1990).

This paper aims at comparing the SRIO and IRIO approaches using the results of Leontief's matrix and multipliers. The calculations prepared according to the CZ-CPA 2 digit are demostrated at the aggregate level for individual products and regions to give a true picture of the main differences between the approaches. We expect systematic structural differences caused by disregarding relations among regions. These differences are illustrated for the visualization of their systematicness and homogeneity (heterogeneity) across individual regions and products. Moreover, we illustrate the results using figures for the Czech Republic. They clearly show the strength of the connection among geographically close regions.

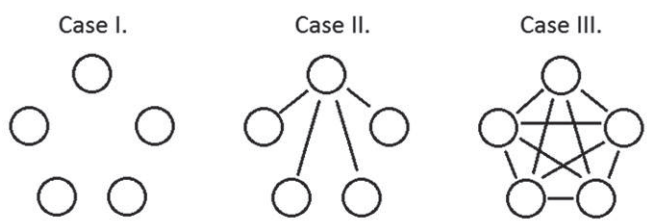
1 LITERATURE REVIEW

The growing number of methods used for regionalization of national input-output tables and the increasing amount of individual analyses (Daniels et al., 2011; Okadera et al., 2014; or Kim et al., 2004) both confirm the rising importance of regional input-output analysis. This phenomenon justifies the existence of multi-regional input-output tables for individual states (Timmer et al., 2015). The main role of regional input-output tables lies in the clarification of the decomposition process of the national impact on regional bases where individual impacts may act unequally even if the weighted sum of effects corresponds to the national analysis.

The following case illustrates this situation: a country which produces a product (Q) mostly in one region (Rq). It affects the intermediate consumption of the region as well. Thus, in the case of an exogenous impulse in another region and another product the demand for the product Q intermediate consumption could increase even if this product does not fall within the intermediate consumption of the region. This is caused by different regional structure where the product Q enters intermediate consumption. These differences bring about an inconsistency of estimations at the national level compared to the regional level.

Many regional input-output models exist which analyse multi-regional effects triggered by regional export and import (Miller and Blair, 2009). According to Lenzen et al. (2004) these models divide into three cases (Figure 1).

Figure 1 Three approaches of modelling regional IOA



Source: Authors' own elaboration based on Lenzen et al. (2004)

Case I represents the situation where individual regions create individual units with no export and import among each other (SRIO approach). Any export and import constitutes only exogenous variables. This is an analogy to the national model (Miller and Blair, 2009; EUROSTAT, 2008). Case II illustrates the model where an increase of output in one region causes an increase of output in another region.

However, this other region has no influence on the remaining regions. In this case, we are unable to discuss the so-called 'backward-linked multipliers' (Steinback, 2004). Case III works with relations among all regions. This model allows us to observe backward regional effects. The inconsistence among all cases is called information bias. It can be proved that the consistence of IRIO and national input-output tables creates a neutral bias (Crown, 1990).

Taking no account of inter-regional flows within the IRIO model comprises the root of the inconsistency between the SRIO and IRIO approaches. There is also a certain synthesis of the approaches in question, i.e. Leontief's international model (Leontief, 1953; Leontief and Strout, 1963). This model finds a way between by dividing an economy to an individual region and the rest of the given economy (e.g. Miller and Blair, 1985). This model does not allow researchers to evaluate backward effects or distinguish the target regions to which the production of the examined region multiplies. However, the construction of such model requires a calculation of the flow between the region and the rest of the economy.

Two basic regional input-output models exist for Case III, i.e. Isard's IRIO model (Isard et al., 1960) and Chenery's Multi-regional input-output model, abbreviated as MRIO (Chenery, 1953). The main difference between these models lies in the detail of calculation. While MRIO does not consider a detailed allocation of flows among regions, IRIO requires such data. With respect to the detail of IRIO, we decided to use this approach. IRIO allows us to investigate detailed differences among regions and effects of the flows among regions.

Even though several inputs (e.g. Lahr, 1993) indicate that data sources constructed without survey could produce biased results, we decided to base the flows among regions on minimization of distance (Šafr, 2016). Several input lead us to this choice: firstly, the homogeneity of methods used; secondly, we assume this bias as insignificant in the case of flows among regions (Sargento, Ramos and Hewings, 2012); finally, indirect estimates are accepted disregarding other available data sources.

2 METHODOLOGY[3]

2.1 National Input-Output Methodology (NIO)

The core of IOA consists in the matrix of intermediate consumption $\boldsymbol{X}$. Components of this matrix represent the flow of output from industry i to industry j. If we summarize everything that industry i supplies to other industries and add total final consumption (y) and export (e) in this industry, we get the total output of this industry. The following formula represents the basic equation of IOA (EUROSTAT, 2008):

$$\sum_{j}^{n} x_{ij} + y_i = x_i \,, \;\; j = 1,2,3\ldots\ldots, n. \tag{1}$$

x_{ij} represents the flow of intermediate consumption from industry i to industry j; y_i comprises the final use of product i (final consumption together with export). The proportions of intermediate consumption flows from industry i to industry j on total production of industry j represent technical coefficients:

$$\text{Matrix}: \mathbf{A} = (a_{ij})_{nn}, \text{where: } a_{ij} = \frac{x_{ij}}{x_j}, \;\; j,i = 1,2,3\ldots\ldots, n. \tag{2}$$

Technical coefficients represent production functions of individual industries which remain stable over a long time period. Moreover, they show how many inputs of intermediate consumption one unit

[3] This part of the paper was published in Šafr and Vltavská (2016): *The evaluation of economic impact using the regional input-output model: the case study of Czech regions in context of national input-output tables* (14th International Scientific Conference 'Economic Policy in the European Union Member Countries'). As we think it necessary for the clarification of the method used, we publish this part in this paper as well.

of output of the industry *i* requires. We constructed the fundamental input-output model from equation (2) which describes the value of total output necessary for fulfilling final use:

$$\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{y}, \tag{3}$$

where **I** represents the identity matrix and $(\mathbf{I}–\mathbf{A})^{-1}$ Leontief inversion.

2.2 Single Region Input-Output Analysis (SRIO)

Regional relations are similar to national ones. Miller and Blair (2009) describe the fundamental production function:

$$\sum_{j}^{n} x_{ij}^{R} + y_{ij}^{R} = x_{i}^{R}, \; j = 1,2,3\ldots\ldots,n, \tag{4}$$

where x_{ij}^{R} represents the flow of intermediate consumption from industry *i* to industry *j* in region *R*; y_{ij}^{R} represents final use of product *i* in region *R*. The difference between the national and regional model lies in export as part of final use. The regional model includes not only export outside the country ($e_i^{N_R}$) but export to other regions within the country (e_i^{R}) as well.

Following description characterises technical coefficients:

$$\text{Matrix: } \mathrm{A}^{R} = (a_{ij}^{R}), \text{ where: } a_{ij}^{R} = \frac{x_{ij}^{R}}{x_{i}^{R}}, \; j,i = 1,2,3\ldots\ldots,n. \tag{5}$$

Regional technical coefficients differ from national technical coefficients. Šafr (2016) described their relation as follows:

$$\left[\sum_{R=1}^{m} \mathbf{A}^{R} diag(\mathbf{x}^{R})\right] diag(\mathbf{x})^{-1} = \mathbf{A}. \tag{6}$$

Finally, formula (3) is adjusted for the regional model:

$$\mathbf{x}^{R} = (\mathbf{I} - \mathbf{A}^{R})^{-1}\mathbf{y}^{R}. \tag{7}$$

2.3 Inter-regional Input-Output Analysis (IRIO) – Isard's approach

IRIO is based on decomposition of matrix A (Miller and Blair, 2009). Our goal is the construction of a matrix of intermediate consumption X^T that simultaneously differentiates individual products and individual industries. This matrix consists of *n* products and *m* regions (this matrix has $m \times n$ columns and rows in total). The diagonal of X^T represents the regional matrix of intermediate consumption (X^T). Matrices outside the diagonal represent the allocation of import from region *i* to region *j*:

$$\mathrm{X}^{\mathrm{T}} = \begin{bmatrix} \mathrm{X}^{1} & \mathrm{F}^{1,2} & \cdots & & \cdots & \mathrm{F}^{1,m-1} & \mathrm{F}^{1,m} \\ \mathrm{F}^{2,1} & \mathrm{X}^{2} & \cdots & & \cdots & \mathrm{F}^{2,m-1} & \mathrm{F}^{2,m} \\ \vdots & \vdots & \ddots & & & \vdots & \vdots \\ & & & \mathrm{X}^{\mathrm{R}} & & & \\ \vdots & \vdots & & & \ddots & \vdots & \vdots \\ \mathrm{F}^{m-1,1} & \mathrm{F}^{m-1,2} & \cdots & & \cdots & \mathrm{X}^{m-1} & \mathrm{F}^{m-1,m} \\ \mathrm{F}^{m,1} & \mathrm{F}^{m,2} & \cdots & & \cdots & \mathrm{F}^{m,m-1} & \mathrm{X}^{m} \end{bmatrix} \tag{8}$$

Beside others, the condition of the transition from X^T to national matrix X applies:

$$\sum_{R=1}^{m}\sum_{\beta=1}^{m*n} x^{T}_{[(R-1)*82+i],\beta} = \sum_{j}^{n} x_{i,j.} \tag{9}$$

Columns of X^T has to follow the same condition:

$$\sum_{R=1}^{m}\sum_{\alpha=1}^{m*n} x^{T}_{\alpha,[(R-1)*82+j]} = \sum_{i}^{n} x_{i,j.} \tag{10}$$

Formulas (9) and (10) ensure comparability of inter-regional impacts and national impacts. Thus, this represents inter-regional decomposition of national matrix with respect to inter-regional particularity.

The final part presents the construction of matrix **F**. An unsolved problem lies in the construction of regional matrices $\mathbf{F}^{R,P}$where rows represent export of output from industries in region R to region P. This matrix has the same number of columns and rows as the matrix of intermediate consumption and has to respect the volume of inter-regional flows (Šafr, 2016). Using the matrix of intermediate consumption of import ($\mathbf{X}^{R,imp}$) we approximate the structure of $\mathbf{F}^{R,P}$:

$$\sum_{R=1}^{m} \mathbf{F}^{R,P} = \mathbf{X}^{P,imp}. \tag{11}$$

Moreover we gain information about $\mathbf{F}^{R,P}$:

$$\begin{bmatrix} \mathrm{F}^{R,P}_{1,1} & \cdots & \mathrm{F}^{R,P}_{1,j} & \cdots & \mathrm{F}^{R,P}_{1,n} \\ \vdots & \ddots & & & \vdots \\ \mathrm{F}^{R,P}_{i,1} & & \mathrm{F}^{R,P}_{i,j} & & \mathrm{F}^{R,P}_{i,n} \\ \vdots & & & \ddots & \vdots \\ \mathrm{F}^{R,P}_{n,1} & \cdots & \mathrm{F}^{R,P}_{n,j} & \cdots & \mathrm{F}^{R,P}_{n,n} \\ & & \sum = & & \\ i^{R,P}_{1} & \cdots & i^{R,P}_{j} & \cdots & i^{R,P}_{n} \end{bmatrix} \quad \sum = \quad \begin{matrix} e^{R,P}_{1} \\ \vdots \\ e^{R,P}_{i} \\ \vdots \\ e^{R,P}_{n} \end{matrix} \tag{11}$$

As we know sums of both rows and columns of $\mathbf{F}^{R,P}$, we can use the RAS method (Sargento et al., 2012) for the approximation of the structure of matrix $\mathbf{F}^{R,P}$ with the condition that the structure of $\mathbf{F}^{R,P}$ and $\mathbf{X}^{R,imp}$ are similar. This concept ensures the consistency between regional and national Leontief's coefficients.

Such approach ensures the consistency between regional (IRIO) and national input-output model where we disregard regions. Thus, multiregional Leontief's coefficients a represent weighted decomposition of national Leontief's multipliers. Using the IRIO approach:

$$\sum (\mathrm{I}^{T} - \mathrm{A}^{T})^{-1} \mathrm{y}^{T} = (\mathrm{I} - \mathrm{A})^{-1}, \tag{12}$$

because in general regional import and export do not exist:

$$\sum (\mathrm{I}^{R} - \mathrm{A}^{R})^{-1} \mathrm{y}^{R} = (\mathrm{I}^{T} - \mathrm{A}^{T})^{-1} \mathrm{y}^{T}. \tag{13}$$

This is the reason why the effects calculated by means of the SRIO approach (the left side of the formula 13) do not correspond with the effect calculated by means of IRIO (the right side of the formula 13). The sum of IRIO equals the national matrix (formula 12). In other words, the sum of effects through individual regions in all products in IRIO has to correspond with the national effects calculated at the national level. Due to the regional export and import this does not apply in SRIO.

3 DATA

For the analysis, we use regional input-output tables (RIOTs) constructed for the reference year 2011 by the Department of Economic Statistics from the University of Economics, Prague (Sixta and Vltavská, 2016; Sixta et al., 2014; Department of Economic Statistics, 2016). RIOTs describe the structure of output in individual regions (NUTS 3 level) corresponding to national input-output tables (IOTs) published by the Czech Statistical Office. Moreover, each region has its own IOT of imported goods. We need both these tables for the analysis and information of regional flows of import and export. However, this data source is not available. RIOTs provide us only with the total amount of import and export for individual industries without any information which region represents the resource side and which region features as the recipient. However, for the construction of IRIO we need more detailed information about the trade, such as which region imports and exports to another region. Therefore, we calculated this information using Karush-Kuhn-Tucker theorem (Šafr, 2016). We proportionally adjusted export into FOB[4] prices for the flows between regions. This ensures the consistency of intermediate consumption matrix.

Table 1 Regional import and export, share on regional output, mil CZK, %

Region	Import	%	Export	%
Jhc	49 914	7.68	43 754	6.73
Jhm	64 432	5.13	49 979	3.98
Kar	44 841	17.01	18 903	7.17
Krh	44 318	7.46	22 010	3.7
Lib	42 311	9.74	15 116	3.48
Mrs	107 683	7.26	63 420	4.28
Olm	47 814	8.48	29 996	5.32
Par	54 351	7.93	37 592	5.48
Pha	222 911	7.37	507 833	16.78
Plz	45 949	7.19	30 567	4.78
Stc	166 479	9.46	89 103	5.06
Ust	79 864	8.17	72 787	7.44
Vys	50 770	9.26	41 331	7.54
Zln	58 437	9.06	57 683	8.94
CZE	1 080 074	7.99	1 080 074	7.99

Note: CZE – the Czech Republic, Pha –Prague, Stc – Central Bohemia Region, Jhc – South Bohemia Region, Plz – the Plzen Region, Kar – the Karlovy Vary Region, Ust – the Usti Region, Lib – the Liberec Region, Krh – the Hradec Kralove Region, Par – the Pardubice Region, Vys – the Vysocina Region, Jhm – the South Moravian Region, Olm – the Olomouc Region, Zln – the Zlin Region, Mrs – the Moravian-Silesian Region.
Source: Authors' calculation

Table 1 shows that the highest absolute value of import and export reaches Prague. However, the results differ if one uses the relative share on the region's output. From such perspective, Prague comprises the most important exporter (16.78%) and an average importer (7.37%). On the contrary, Karlovy Vary Region records the most important relative import with 17.01%.

If we provisionally assume that the regional structure of output and intermediate consumption keep the same level in all regions, we conclude that the Hradec Králové Region is the most undervalued and

[4] FOB prices – Free On Board pricing.

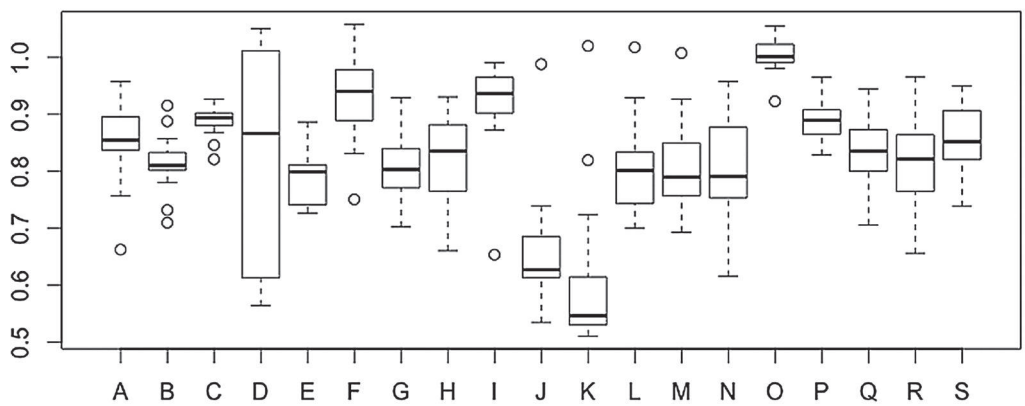
the South Moravian Region the least undervalued area. The highest share of multiplication flows to Prague. However, these results depend on the precise structure of RIOTs which differ from region to region to a certain extent and on the structure of suppliers and consumers of regional output. Moreover, it depends on the right links among individual regions.

4 RESULTS

Leontief's multipliers represent one of the most common tools of IOA. Using IRIO the part of multiplication comprises transfer to import among regions with no influence by other multiplication and it is considered a final quantity. On the other hand, the interregional approach assumes export and import among regions as endogenous variables. This causes an increase of the share of import on output. It further influences export, which constitutes a part of final use. The increase of final use leads to another multiplication. When using IRIO, the export and import among regions ensure the consistency of impacts calculated at regional level with the national level.

For an illustration of bias between SRIO and IRIO, we used the share of SRIO multipliers on IRIO multipliers (Figure 2). Products K (Financial and insurance activities), J (Information and Communication), E (Water supply; sewerage, waste management and remediation activities), A (Agriculture, forestry and fishing) and B (Mining and quarrying) show the most significant differences. On average across all products, SRIO multipliers are undervalued by 14% compared to IRIO. The offer of product K mostly concentrates in Prague, which also causes the underestimation of the product. This induced the fact that product K has notably a role of regional import. Similar situation applies for product J.

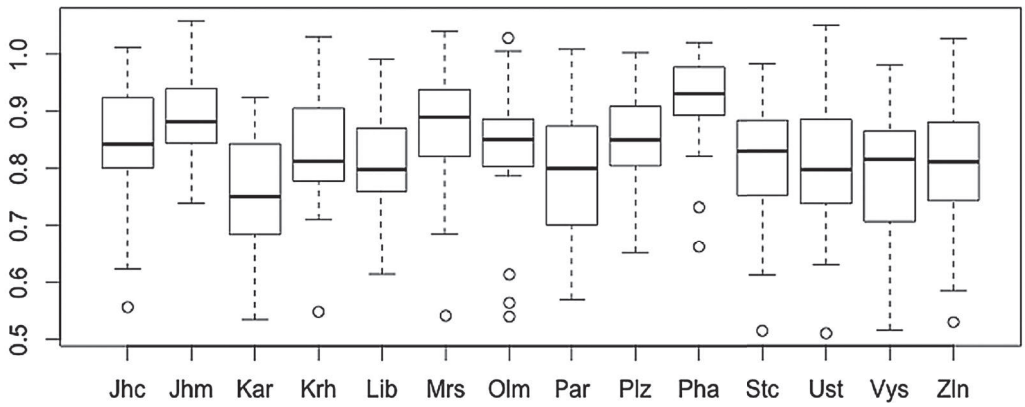
Figure 2 The share of SRIO and IRIO multipliers on products, %



Note: A – Agriculture, forestry and fishing, B – Mining and quarrying, C – Manufacturing, D – Electricity, gas, steam and air conditioning supply, E – Water supply; sewerage, waste management and remediation activities, F – Construction, Services: G – Wholesale and retail trade; repair of motor vehicles and motorcycles, H – Transportation and storage, I – Accommodation and food service activities, J – Information and communication, K – Financial and insurance activities, L – Real estate activities, M – Professional, scientific and technical activities, N – Administrative and support service activities, O – Public administration and defence; compulsory social security, P – Education, Q – Human health and social work activities, R - Arts, entertainment and recreation, S – Other service activities.
Source: Authors' calculation

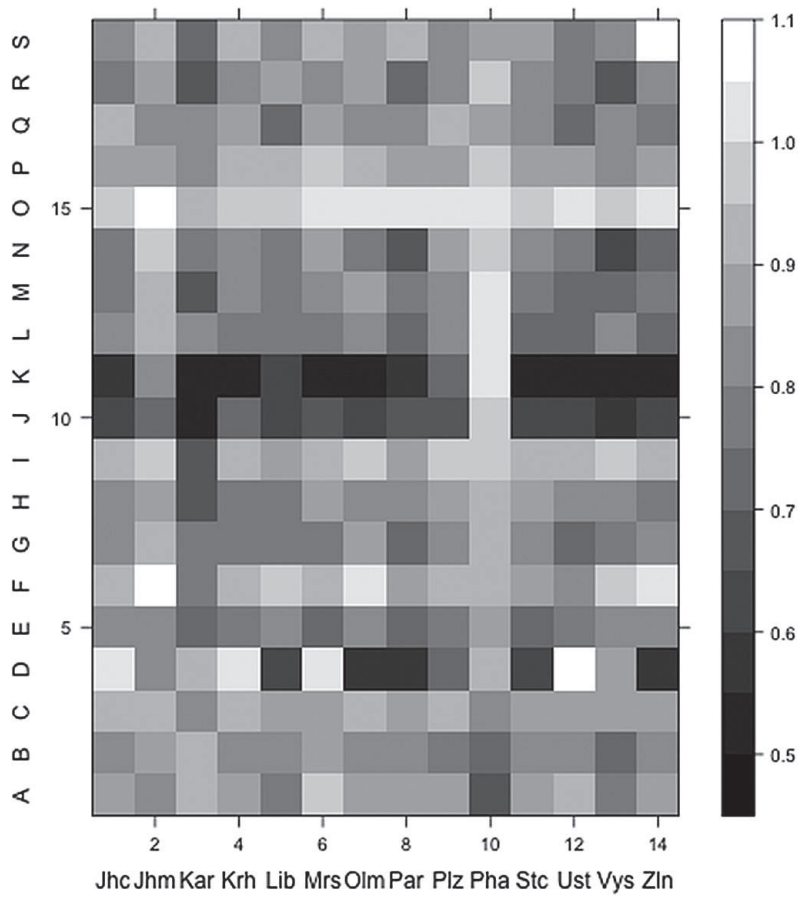
This analysis finds use not only for products but for regions as well. Figure 3 shows that the least undervalued region is Prague, due to a share of export and low share of import on output. This allows us to expect that a lot of output of other regions is multiplied in Prague. This represents an opposite situation than for example in the South Moravian Region. This region demonstrates a low share of import on the output but a low share of export as well.

Figure 3 The share of SRIO and IRIO multipliers in regions, %



Source: Authors' calculation

Figure 4 Level-map of fraction of SRIO and IRIO multipliers at matrix products by regions

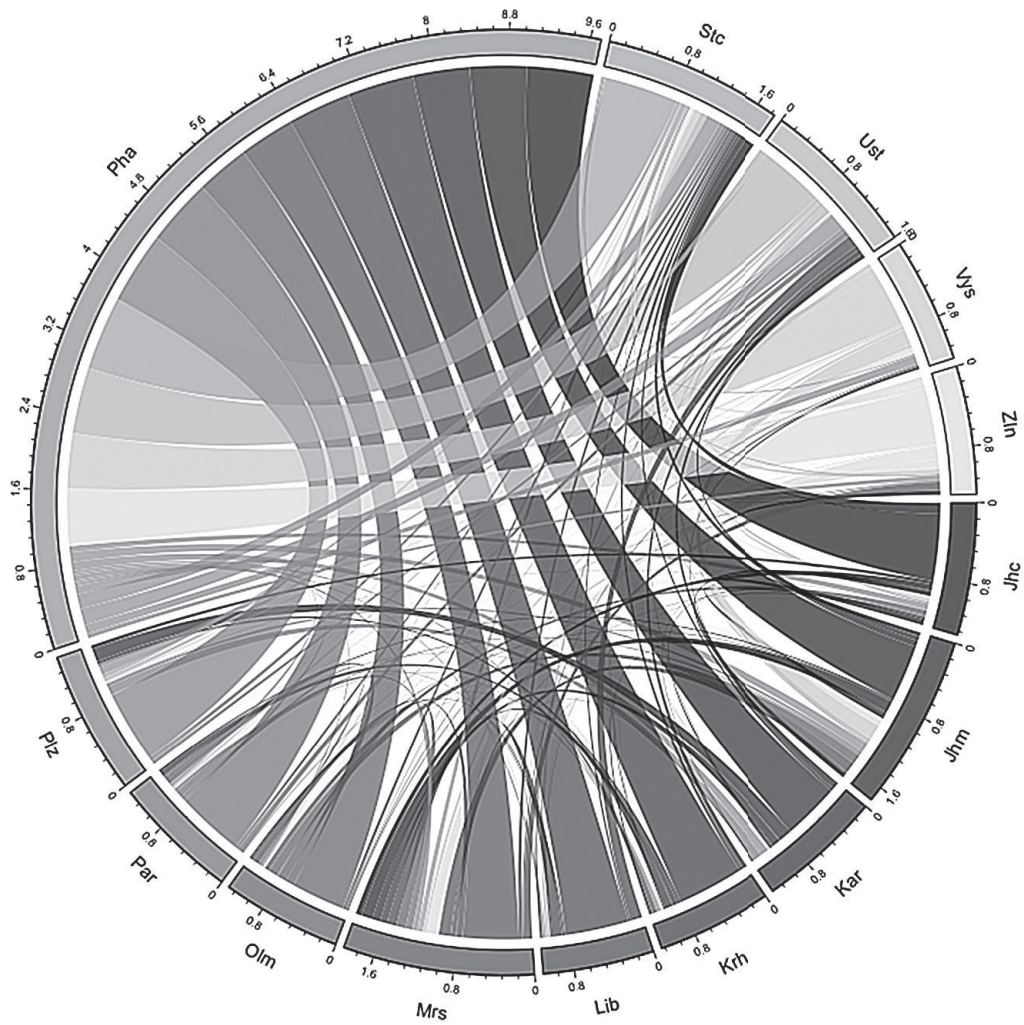


Source: Authors' calculation

We use a level map for an illustration of simultaneous analysis (Figure 4). One can see that mainly products J and K show systematic undervaluing of multipliers. On the other hand, product O (Public administration and defence) has the least undervalued multiplier. Therefore, it distinctly corresponds with the multiregional approach. This figure confirms characteristics of Prague and the South Moravian Region. Their SRIO multipliers correspond more to IRIO than multipliers in other regions.

All these characteristics prove the unique status of Prague among all regions. Figure 5 confirms the idea that if export reaches high and import stays as low as the share on the regional output, the output of other regions multiplies in Prague.

Figure 5 Chord diagram/network circle of multiplication out of regions



Source: Authors' calculation

Figure 5 illustrates multiplication outside regions. The boldness of the link describes the flow from the given region to other regions. If the colour of the flow is different from the one used for the sector, it designates the multiplication into the region. This figure effectively illustrates to which region the output is multiplied due to an increase of uses in the region. The whole diagram corroborates the idea that a significant value of output from other regions multiplies in Prague.

Table 2 describes the strength of IRIO multipliers, i.e. how much of the average multiplier flows outside the region on average.

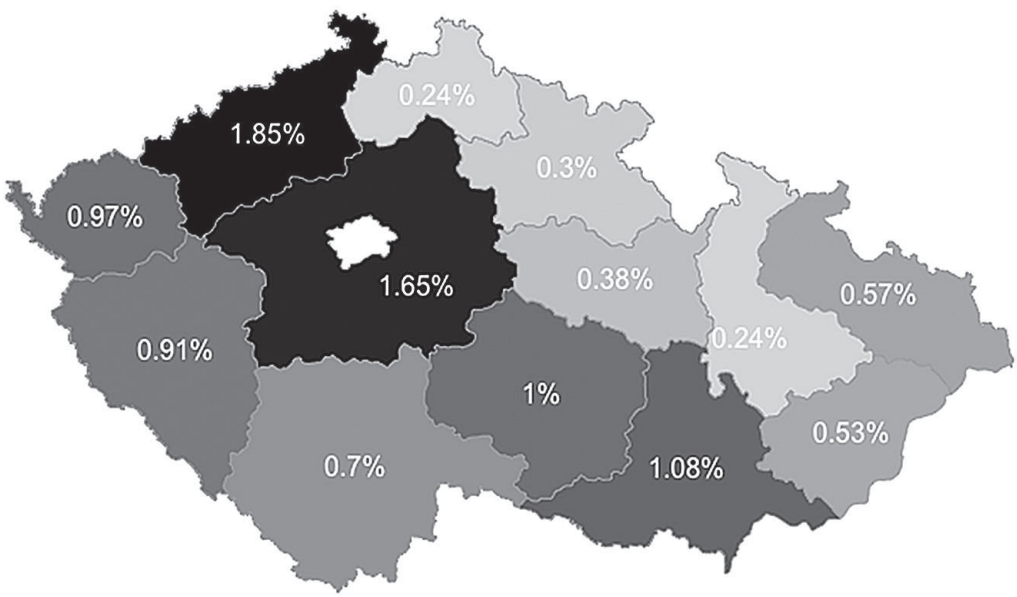
Table 2 Summary statistics of IRIO multipliers, %

Region	Average IRIO multipliers	Part of m. outside region	%	Region	Average IRIO multipliers	Part of m. outside region	%
Jhc	1.61	0.18	0.11	**Par**	1.62	0.20	0.13
Jhm	1.59	0.10	0.06	**Plz**	1.57	0.14	0.09
Kar	1.62	0.25	0.15	**Pha**	1.67	0.10	0.06
Krh	1.60	0.18	0.11	**Stc**	1.62	0.18	0.11
Lib	1.61	0.20	0.12	**Ust**	1.64	0.21	0.13
Mrs	1.57	0.15	0.09	**Vys**	1.66	0.24	0.14
Olm	1.57	0.17	0.11	**Zln**	1.62	0.20	0.12

Source: Authors' calculation

Figures 2 to 5 and Tables 1 and 2 confirm the important status of Prague among the regions. Aiming directly at Prague allows us to find the source regions with the highest average of multiplied output (Figure 6) with the total multiplication in Prague serving as a baseline.

Figure 6 Source regions of the highest multipliers of output in Prague (% part of Prague multiplier)

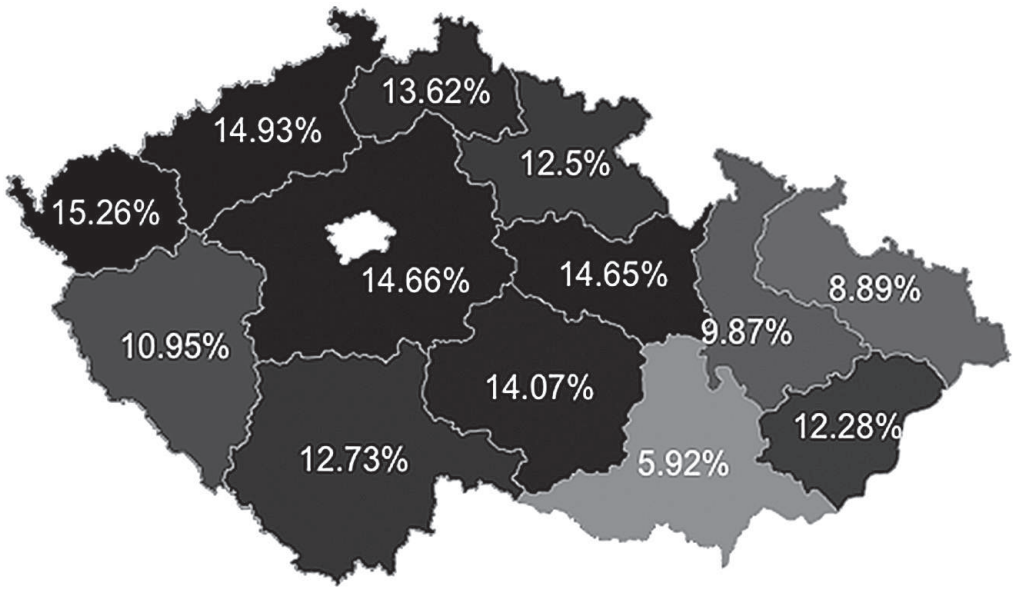


Source: Authors' calculation

Figure 6 proves the strongest connection of Prague and regions geographically close to Prague. The output multiplied in Prague (throughout all categories of products) comes mainly from the Usti Region (1.85%) and the Central Bohemia Region (1.65%).

Figure 7 depicts target regions to which the output of Prague is mostly multiplied. We can even see the decomposition of average multiplication of Prague's output outside the region. A weak connection of the South Moravian Region reflects a generally weak links of this region to the other regions. Table 1 demonstrates it rather clearly.

Figure 7 Regions according to targeting of the multiplied output from Prague (% part of their multipliers)

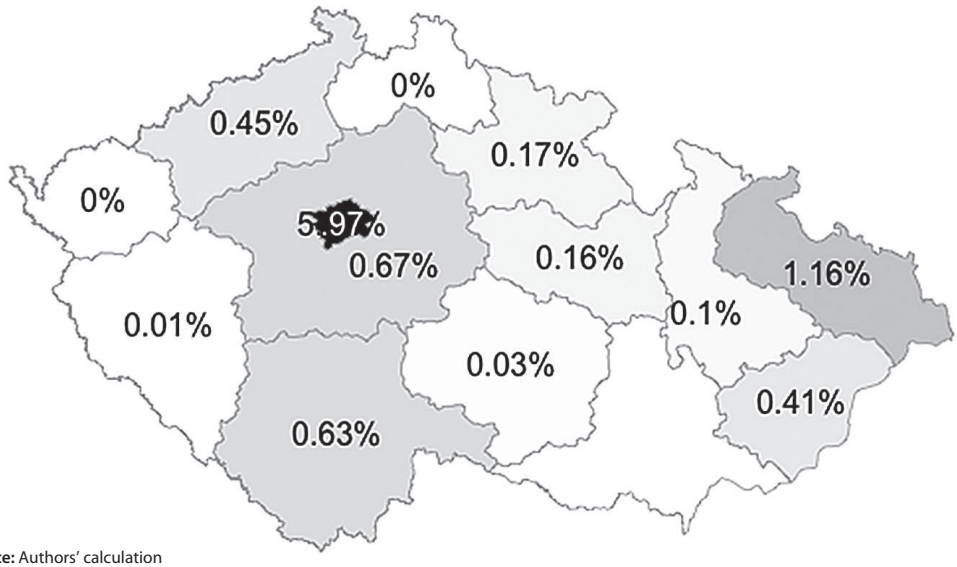


Source: Authors' calculation

The Karlovy Vary Region, the Usti Region and the Central Bohemia Region record the highest multiplication from Prague. Moreover, Figure 7 depicts the influence of the distance of the average connection on the multiplication of the output from Prague. Regions geographically closer to Prague have a stronger link it than a distant region. The costs on import may be the principal cause here. It goes along with economic assumptions about consumers' and producers' behaviours. Minimized costs on import lead to minimizing the import distance which results in strong multiplication mostly to the surrounding regions.

The second example focuses on the region with trends opposite of Prague, i.e. the South Moravian region with the weakest connection to Prague (see Figures 6 and 7) among Czech regions. Figures 8 and 9 summarize the connection of the South Moravian region to all other regions.

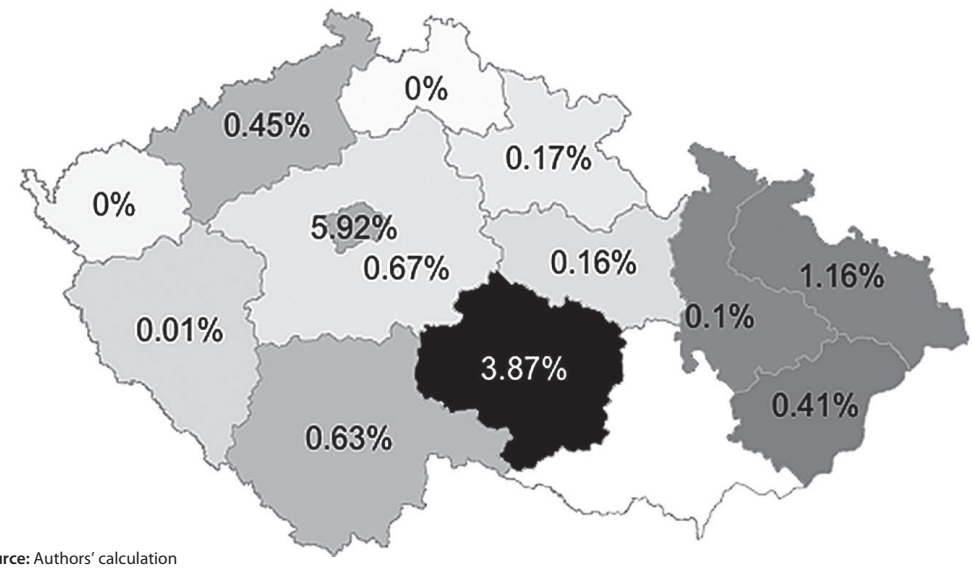
Figure 8 Source regions of the highest multipliers of output to the South Moravian region (% part of the South Moravian region multiplier)



Source: Authors' calculation

Multiplication of the output of the South Moravian region is mainly allocated in Prague, which correspond with the rest of the regions. Prague maintains the strongest position among all regions. Figure 9 shows the ratio of multiplication for the South Moravian region from others regions. This figure proves the minimization of costs on the imported products. Thus, the highest share of the multiplication transfers to the neighbouring regions.

Figure 9 Target regions for the multiplied output of the South Moravian region (% part of their multipliers)



Source: Authors' calculation

DISCUSSION AND CONCLUSION

This paper aimed to illustrate the differences between SRIO and IRIO approaches in IOA. These approaches differ in their construction as well as in the required data. SRIO analyses only one region with no relations to other regions. Thus, this approach proves less data demanding. IRIO approach analysed each region in context of all regions. The different construction and data sources give rise to the following two hypotheses. Firstly, regions analysed by means of SRIO are systematically undervalued. Secondly, Prague has a unique position among the regions.

We employed Leontief matrix for an analytical illustration of the differences between these two approaches. Using the matrix, we illustrate a significant undervalue of single-regional Leontief multipliers' in comparison to IRIO. This undervalue characterizes both average Leontief multipliers for whole regions and partial multipliers for individual products (Figures 2 and 3). Detailed estimations on the level of individual products show wide variability. On the other hand, one can see identical undervalue calculated in individual regions.

Figure 4 proves these expectations for aggregated Leontiefs' products in individual regions. A similar structure of the undervaluation applies to virtually all regions. While the undervaluation which we evaluate at the level of individual regions (in all regions) varies a lot, mainly in the structure of products. These results prove the highest undervaluation of products J and K. Figures 2 to 4 confirm the idea about the undervalue which we expected when using SRIO. The undervaluation is clearly systematic (mainly seen in Figure 4). The undervaluation is traceable in narrow quantiles of individual products using boxplots (Figures 2 and 3).

Figure 5 confirms the hypothesis about a specific position of Prague to which all regions maintain a strong link. Thus, all regions are unequally strongly connected to Prague opposed to the other regions. Figures 6 to 9 illustrate this phenomenon for Prague and South Moravian region. The most significant undervaluation takes place in regions and products which are import demanding on their own output.

The main goal of the paper was achieved mainly by using graphical analyses of Leontiefs matrix. Graphical analysis was necessary due to the impossibility of testing of the data by means of common statistical methods (e.g. t-tests). This analysis was supplemented with an estimation of the undervaluation in individual regions and in the Czech Republic as a whole. Two minor hypotheses (the systematicness of the undervaluation and a unique position of Prague) were illustrated using graphical analysis as well. The results and illustrations confirm both these minor hypotheses. The undervaluation gives ground for questioning of the results by means of SRIO.

The data and the model demonstrate regional heterogeneity of Leontiefs coefficients calculated by means of SRIO and IRIO, as we expected. The fact that various researchers use different data confirms this heterogeneity (e.g. Freeman, Alperovich and Weksler; 1985). The results (undervaluation, systematicness etc.) follow economic and statistical theory behind these models (Lezen et al., 2004).

Recommendation for future work lies in the idea that disregarding the measurement of the connection of a region to other regions produces systematically undervalued results. This bias is noticeable in all regions. Using a more sophisticated regional model, such as MRIO (or used IRIO), although more data demanding, can eliminate such bias. Therefore, a further research lies mainly in more precise and robust methods of the extrapolation of the data sources needed, for example methods for the estimation of the flows of output among regions, or ensuring the consistency of regional estimations with national figures (e.g. 3D RAS etc.).

ACKNOWLEDGEMENT

This paper is prepared under the support of the "Regional estimates of gross domestic product based on the expenditure approach" project of the Czech Science Foundation, project No. 13-15771S and by Institutional Support for Long Period and Conceptual Development of Research and Science at Faculty of Informatics and Statistics, University of Economics, Prague.

References

CHENERY, H. B. Regional Analysis. In: CHENERY, H. B., CLARK, P. G. and PINNA, V. C., eds. *The Structure and Growth of the Italian Economy,* 1953, pp. 91–129.

CROWN, W. H. An Interregional Perspective on Aggregation Bias and Information Loss in Input-Output Analysis. G*rowth And Change*, Vol. 21, No. 1, 1990, pp. 11–29, EconLit with Full Text, EBSCO*host* (viewed 7th September 2016).

DANIELS, P. L., LENZEN, M., KENWAY, S. J. The ins and Outs of Water use – A Review of Multi-Region Input-Output Analysis and Water FootPrints for Regional Sustainability Analysis and Policy. *Economic Systems Research*, Vol. 23, No. 4, 2011, pp. 353–370.

DEPARTMENT OF ECONOMIC STATISTICS. *Regionalization of gross domestic product employing expenditure approach* (data) [online]. Department of Economic Statistics, University of Economics, Prague, 2016. [cit. 1.6.2016]. Available from: <*http://kest.vse.cz/veda-a-vyzkum/vysledky-vedecke-cinnosti/regionalizace-odhadu-hrubeho-domaciho-produktu-vydajovou-metodou*>.

EUROSTAT. *Manual of Supply, Use and Input-Output Tables* [online]. Eurostat, 2016. [cit. 1.6.2016]. Available from: <*http://apl.czso.cz/nufile/sut/IO_manual.pdf*>.

FREEMAN, D., GERSHON A., ITZHAK, W. Inter-regional input-output model – the Israeli case [online]. *Applied Economics*, Vol. 17, No. 3, 1985, pp. 381–393. [cit. 22.9.2016].

TIMMER, M., DIETZENBACHER, E., LOS, B., STEHRER, R., DE VRIES, G.. An Illustrated User Guide to the World Input–Output Database: the Case of Global Automotive Production. *Review of International Economics*, Vol. 23, No. 3, 2015, pp. 575–605.

ISARD, W. et al. *Methods of Regional Analysis: An Introduction to Regional Science.* MIT Press, 1960.

KIM, E., HEWINGS, D. J., HONG, CH. An Application of an Intergrated Transport Network-Multiregional CGE Model: a Framework for the Economic Analysis of Highway Projects. *Economic Systems Research,* Vol. 16, No. 3, 2004, pp. 235–258.

LAHR, M. L. A Review of the Literature Supporting the Hybrid Approach to Constructing Regional Input-Output Models [online]. *Economic Systems Research,* Vol. 5, No. 3, 1993, pp. 277–293. [cit. 22.9.2016].

LENZEN, M., PADE, L., MUNKSGAARD, J. CO2 Multipliers in Multi-region Input-Output Models. *Economic Systems Research,* Vol. 16, No. 4, 2004, pp. 391–412.

LEONTIEF, W. Interregional Theory. In: LEONTIEF, W. W., CHENERY, H. B., CLARK, P. G. *Studies in the structure of the American economy*, Oxford University Press, New York, 1953, pp. 93–115.

LEONTIEF, W., STROUT, A. Multi-regional Input-Output Analysis. In: BARNA, T., eds. *Structural Interdependence and Economic Development,* London: St. Martin's Press, 1963.

MILLER, R. E., BLAIR, P. D. *Input-Output Analysis: foundations and ex-tensions.* Cambridge University Press, 2009.

MILLER, R. E., BLAIR, P. *Input-Output Analysis – Foundations and Extensions.* Prentice-Hall, 1985.

OKADERA, T., OKAMOTO, N., WATANABE, M., CHONTANAWAT, J. Regional water footprints of the Yangtze river: an interregional Input-Output Approach. *Economic Systems Research,* Vol. 26, No. 4, 2014, pp. 444–462.

ŠAFR, K. Allocation of commodity flows in the regional Input-Output tables for the Czech Republic. In: *Proceedings of 19th International Scientific Conference Application of Mathematics and Statistics in Economics,* Banská Štiavnica, 2016 (in print).

ŠAFR, K., VLTAVSKÁ, K. The evalutation of economic impact using regional Input-output model: The case study of Czech regions in context of national Input-Output tables. In: P*roceedings of 14th International Scientific Conference "Economic Policy in the European Union Member Countries",* Petrovice u Karviné, 2016 (in print).

SARGENTO, A. L. M., RAMOS, P. N., HEWINGS, G. J. D. Inter-regional Trade Flow Estimation through Non-survey Models: An Empirical Assessment. *Economic Systems Research,* Vol. 24, No. 2, 2012, pp. 173–193.

STEINBACK, S. R. Using Ready-Made Regional Input-Output Models to Estimate Backward-Linkage Effects of Exogenous Output Shocks. The Review of Regional Studies, Vol. 34, No. 1, 2004, pp. 57–71.

SIXTA, J., FISCHER, J., ZBRANEK, J. Regional Input-Output Tables. In: *Proceedings of 32nd International Conference on Mathematical Methods in Economics,* Olomouc, 2014, pp. 896–901.

SIXTA, J., VLTAVSKÁ, K. Regional Input-output Tables: Practical Aspects of its Compilation for the Regions of the Czech Republic. *Ekonomický časopis,* Vol. 64, No. 1, 2016, pp. 56–69.

SUTTINON, P., NASU, S., IHARA, T., BONGOCHGETSAKUL, N., UEMOTO, K. Water Resources Management in Shikoku Region by Inter-regional Input-Output Table. *Review Of Urban And Regional Development Studies,* Vol. 25, No. 2, 2013, pp. 107–127, EconLit with Full Text, EBSCO*host* (viewed 7th September 2016).

Detection of Breaks in a Capital Structure: a Case Study

Jaromír Antoch [1] | *Charles University, Prague, Czech Republic*
Daniela Jarušková [2] | *Czech Technical University, Prague, Czech Republic*

Abstract

The main goal of this paper is to present an analysis of financial quarterly time series describing the level of book leverage of U.S. companies selected from different industries in the period 1991–2014. The basic question is whether the sub-prime crisis 2007–2008 caused a change in the behavior of the respective companies. More generally, we are interested whether the time series may be considered stationary. Statistical methods suitable for the detection of breaks (changes) for individual and panel data are presented together with their pros and cons. Against our expectations, the analysis did not reveal a significant change due to the sub-prime crisis. On the other hand, all series contain at least one change, most of the changes occurring around the year 2000, thus offering room for an economic explanation.[3]

Keywords

Change point problem; abrupt, gradual and multiple changes; stationarity in the mean; sum and maximal test statistics; panel data; book leverage

JEL code

C10, C23

INTRODUCTION

Capital structure determines the relative ownership of the firm by creditors and equity holders, as represented by the relative weights of debt and equity in the company. Therefore, how a firm chooses its capital structure is one of the fundamental questions in corporate finance, and financial economic research focuses on variables that help explain capital structure decisions. For details see, e.g., seminal paper by Lemmon et al. (2008).

The key variable in capital structure is leverage, so that one of the basic research questions is whether the leverage, or any other key characteristic describing the capital structure, is time invariant or whether it contains a breaking point(s). If it does contain a breaking point(s), then the question is how to estimate them and how to decide which phenomenon is behind them.

In this paper we concentrate on selected issues from the change-point methodology and illustrate advantages and pitfalls of the selected approach on the analysis of real financial data. More specifically,

[1] Faculty of Mathematics and Physics, Department of Probability and Mathematical Statistics, Sokolovská 83, 186 75 Prague, Czech Republic. Corresponding author: e-mail: antoch@karlin.mff.cuni.cz.
[2] Faculty of Civil Engineering, Department of Mathematics, Thákurova 7, 166 29 Prague 6, Czech Republic. E-mail: daniela.jaruskova@cvut.cz.
[3] This article is based on contribution from the conference *Robust 2016*.

we are interested in breaks in the model describing the level of leverage of selected U.S. companies from different industries around the sub-prime crisis. Recall that the sub-prime crisis is generally defined between the fourth quarter of 2007 and the end of 2008, see Santos João (2011) and Dick-Nielsen et al. (2012) for details. Time span of the considered data covers the period from 1Q 1991 to 4Q 2014.

This paper is organized as follows. Section 1 describes statistical methods suitable for detection of changes in the underlying model. Section 2 describes the data and its analysis. Finally, Section 3 summarizes selected conclusions.

1 DESCRIPTION OF STATISTICAL METHODS USED

In the scope of mathematical statistics, whether an observed series has remained stationary or whether a change of a specific kind has occurred, the outcome is usually based on hypotheses testing. The null hypothesis claims that the process is stationary while the alternative hypothesis claims that the process is nonstationary and the stationarity was violated in a specific way. In our case we will mainly be interested in stationarity in the mean of the observed series.

We usually start the statistical inference process by analyzing information on one series describing the behavior of one company. We assume that the data $Y_1,..., Y_n$ was collected at time moments $t_1 < ... < t_n$, so that they form a time series. In our case the time moments can be, without a loss of generality, replaced by their indices 1,..., n. When studying the data for one company, our goal is to decide whether the sequence $Y_1,..., Y_n$, is stationary or whether its mean has changed. We assume that a potential change of the analyzed series' mean occurred in a short, with respect to n, time period. Thus we can make a slight simplification dealing with time series models that contain a sudden shift in the mean at an unknown time point.

In our case the null hypothesis claims that the characteristic has not changed while the alternative claims that the analyzed characteristic has changed in an assumed manner. For testing which of these hypotheses is true we use statistics developed in the field of change-point detection. For more details and different approaches to the problem see, e.g., Csörgö et Horváth (1997), Bai et Perron (1998), Antoch et al. (2002, 2004, 2007, 2008), Antoch et Jarušková (2013) or Horváth et Rice (2014).

Recall that analogous methodology has been developed for gradual changes as well. Nevertheless, it is well known that procedures developed for detection of a sudden change also respond in the case of gradual changes, and vice versa. However, one must keep in mind that in such a case they lose a power. For details see, e.g., Antoch et al. (2002) or Jarušková (1998).

First let us explain how the test statistics applied in our paper are constructed. Suppose for a while that we know the position of a potential change point (break). In other words, if we know that a change occurred, then it certainly occurred at the time k. In such a case, for deciding whether or not the analyzed series has changed, one may use a classical two-sample test statistic for testing the equality of the mean of the first part $Y_1, ..., Y_k$ to the mean of the second part $Y_{k+1}, ..., Y_n$, of the original series. A natural estimate of the first mean is the average $\overline{Y}_k = \sum_{j=1}^{k} Y_j/k$ and, similarly, the estimate of the second mean is the average $\overline{Y}_k^0 = \sum_{j=k+1}^{n} Y_j/(n-k)$.

Supposing moreover that the variance σ^2 of the series remains the same over the entire time span $j = 1, ..., n$, then the test statistic T_k has the form:

$$T_k = \sqrt{\frac{k(n-k)}{n}}\,\frac{\overline{Y}_k - \overline{Y}_k^0}{\hat{\sigma}} = \sqrt{\frac{n}{k(n-k)}}\,\frac{\sum_{j=1}^{k}(Y_j - \overline{Y}_n)}{\hat{\sigma}}. \tag{1}$$

Notice that test statistic T_k may be obtained as the maximum likelihood estimator under the assumption that observations $\{Y_i\}$ are independent normally distributed random variables. For a detailed derivation see Section 3 in Antoch et al. (2002).

In Formula (1) the standard deviation of the analyzed time series has been replaced by its estimate, that can be calculated either as:

$$\hat{\sigma}_1 = \sqrt{\left[\sum_{j=1}^{k}\left(Y_j - \bar{Y}_k\right)^2 + \sum_{j=k+1}^{n}(Y_j - \bar{Y}_k^0)^2\right]/(n-2)},$$

or as:

$$\hat{\sigma}_2 = \sqrt{\sum_{j=1}^{n}(Y_j - \bar{Y}_n)^2/n}.$$

In the situation where data is dependent and forms a linear process, σ must be estimated more carefully. A Bartlett type estimator adjusted to a possible change is usually recommended in the literature as the first choice. A detailed description can be found in Antoch et al. (1997).

In a case where the means of the first and second parts of the series coincide, the $\hat{\sigma}_1$ and $\hat{\sigma}_2$ do not differ substantially. If the means differ, then $\hat{\sigma}_1$ attains a smaller value than $\hat{\sigma}_2$ with a large probability, so that the test statistic using this value has a larger power for change point detection. It is well known that when k and $n - k$ are large, then the statistic T_k has approximately a standard normal distribution, and the hypothesis claiming that the means of the first and second parts are the same is rejected if $|T_k| > u_{1-\alpha/2}$ with $u_{1-\alpha/2}$ being the $(1 - \alpha/2)100\%$ quantile of $N(0,1)$.

If we do not know the position of a potential change point (break), then we calculate the value of the statistic T_k for all possible $k = 1, \ldots, n - 1$, and plot the sequence $\{|T_k|\}$ against time points $\{k; k = 1, \ldots, n - 1\}$. The plot provides us with important visual information about eventual change point(s). As the sequence $\{T_k\}$ is a standardized CUSUM sequence of residuals $\{\bar{Y_j} - Y_n\}$, which starts ($k = 0$) and ends ($k = n$) at zero. If a sudden shift occurs at a time k, the sequence $\{|T_k|\}$ attains a large value for such a k. A magnitude of this value is given by a difference between the means of the first and second parts of the series, i.e., by the size of a shift in the mean. If there are several sudden changes that are well separated, then the sequence $\{|T_k|\}$ has more peaks.

In addition to the sequence $\{T_k\}$ we may also compute a weighted sequence $\{w_k T_k\}$. The most frequently applied weights are:

$$w_k = \sqrt{k(n-k)/n^2}, \quad k = 1, \ldots, n-1, \tag{2}$$

leading to the statistic:

$$T_k^* = \sqrt{\frac{1}{n}}\frac{\sum_{j=1}^{k}(Y_j - \bar{Y}_n)}{\hat{\sigma}}. \tag{3}$$

As shown in James et al. (1987), the statistics T_k^* may be obtained using the modified likelihood principle.

It is not surprising that a decision on existence of a change point is based on the maximum of the statistics $\{|T_k|\}$, i.e.,

$$max_{1 \le k \le n-1} |T_k|, \tag{4}$$

respectively on the maximum of the statistics $\{w_k|T_k|\}$, i.e.,

$$max_{1 \le k \le n-1} |T_k^*|, \tag{5}$$

sometimes called a weighted maximum type test statistic. Notice that some authors use the term "penalized maximum type test statistic" here. However, the terminology is not uniform and we will use the term "weighted" throughout this paper.

Recall that besides test statistics (4) and (5) we might also use test statistics that are sums of $\{T_k^2\}$ or $\{(w_k T_k)^2\}$. Because we do not apply them in this paper, we refer to Antoch et al. (2002) and MacNeill (1974) for more details.

As an estimator of the time of change one usually takes that index $\hat{k}_0$, for which the sequence of the statistics $\{T_k\}$ attains its maximum, i.e.,

$$\hat{k}_0 = argmax_{1\le k\le n-1} |T_k|, \tag{6}$$

respectively, where the sequence of statistics $\{w_k|T_k|\}$ attains its maximum, i.e.,

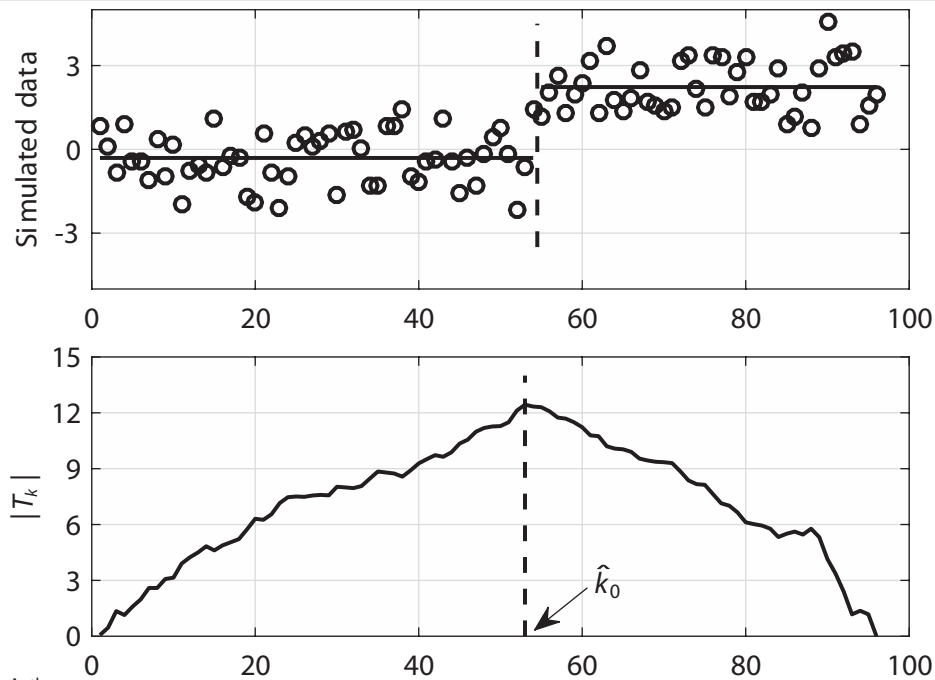
$$\hat{k}_0^* = argmax_{1\le k\le n-1} |T_k^*|. \tag{7}$$

If the maximum is not unique, so that a maximum is attained for a set of indices, we usually take as the estimator the smallest index of this set. Nevertheless, such an issue usually indicate that more than one change can be detected in the data, and one should deal with this issue. For details see, e.g., paper Antoch et Hušková (1998).

It is worth noticing that statistics (6) and (7) do not necessarily correspond to the location of a change provided the series exhibits a gradual instead of sudden change. In such a case, other estimators have to be used; for details see, e.g., Antoch et Hušková (1998) and Antoch et al. (2002). See also discussion in Section 2.

Clearly, the test statistic (5) detects more easily a change in the middle of the sequence while the statistic (4) detects more easily a change at the beginning or at the end of the series. As an illustration, Figure 1 shows a simulated time series with a shift in its mean and a corresponding behavior of the sequence $\{|T_k|\}$. Behavior of both sequences $\{|T_k|\}$ and $\{|T_k^*|\}$ when applied to the real data describing the level of the book leverage of U.S. companies selected from different industries, can be seen in Figure 2.

Figure 1 Simulated data and behavior of statistics $\{|T_k|\}$



Source: Authors

The exact distribution of statistics (4) and (5) is too complex; hence approximate critical values have to be applied. The approximate critical values may be obtained by simulations, where the observations $\{Y_i\}$ are taken from a standard normal distribution. These critical values may be applied to a broad class of distributions thanks to the invariance principle. For $n \approx 100$ the 5% critical value of statistic (4) is 3.17 and the 1% critical value is 3.70. For $n \approx 100$ the 5% approximate critical value of statistic (5) is 1.29, while the 1% approximate critical value is 1.55. As argued above, it is always useful to plot either statistics $\{|T_k|\}$ or $\{|T_k^*|\}$ against time points $\{k; k = 1, ..., n - 1\}$.

To get approximations to the distribution of the considered test statistics, different versions of the bootstrap were suggested in the literature. Because this issue goes far beyond the scope of this paper, we refer the reader to Antoch et al. (1995) and Horváth et Rice (2014) for details and additional references.

If the series contains more than one change, and the changes are well separated, the statistics (4) and (5) are able to reject the null hypothesis of stationarity in the means well. For estimating multiple change points, a sequential procedure proposed in Vostrikova (1981), and later modified by many other authors, may be applied. The basic idea may be described as follows. If a change is detected, the series is split into two parts, i.e., the part before the detected change point and the part after it. Then the same procedure is applied to both subseries recursively. Another possibility is to use the MOSUM approach discussed, e.g., in Antoch et al. (2002), or to employ a test statistic proposed for detecting several changes developed, e.g., in Antoch et Hušková (1994) and Antoch et Jarušková (2013).

The critical values for statistics (4) and (5) presented above were obtained under an assumption that $\{Y_i\}$ form a sequence of independent variables. When $\{Y_i\}$ form an ARMA sequence or, more generally, a linear process, the same test statistics may be applied, but σ^2 must be estimated more carefully and the critical values must be adapted. For more details see Antoch et al. (1997).

Finally, consider a situation when the data comes from I companies and are obtained during the same time moments $t_1 < ... < t_n$. We say that they form a so-called "panel". Suppose that $Y_j(i)$ denotes a value of variable of interest, e.g. book leverage, at time t_j for a company i. Then we can organize the data into a matrix with n rows and I columns. Moreover, we assume that if there is a change point k_0, then any series $\{Y_j(i), j = 1, ..., n\}$ either changes at time k_0 or does not change at all. For the i^{th} company we compute $\bar{Y}(i) = n^{-1} \sum_{j=1}^{n} Y_j(i)$ and $s(i) = \sqrt{n^{-1} \sum_{j=1}^{n} \left(Y_j(i) - \bar{Y}(i)\right)^2}$. Then for the i^{th} company and for $k = 1, ..., n$ we compute either

$$t_k(i) = \frac{n}{k(n-k)} \left(\sum_{j=1}^{k} \frac{Y_j(i) - \bar{Y}(i)}{s(i)} \right)^2,$$

or

$$v_k(i) = \frac{1}{n} \left(\sum_{j=1}^{k} \frac{Y_j(i) - \bar{Y}(i)}{s(i)} \right)^2.$$

Notice that for the i^{th} company $t_k(i) = T_k^2$ and $v_k(i) = (w_k T_k)^2$, where w_k are defined in (2). Further, for any time point $k = 1, ..., n$ we compute statistics:

$$U_k = I^{-1} \sum_{i=1}^{I} (t_k(i) - 1),$$

or a test statistic:

$$Z_k = I^{-1} \sum_{i=1}^{I} \left(v_k(i) - \frac{k(n-k)}{n} \right).$$

Similar to the one-dimensional case, the resulting panel test statistic can be either the maximum or sum of statistics $\{U_k\}$, respectively $\{Z_k\}$. We will not discuss here the details of either the appropriate normalization or finding the corresponding critical values, because such considerations go beyond the scope of this paper, being technically too complicated. The interested reader can find a detailed description and more about the analysis of panel data in Hušková et Horváth (2012) or, e.g., Baltagi (2013), Antoch (submitted).

Analogous to the case of statistics $\{T_k\}$ and $\{T_k^*\}$, the plot of $\{U_k\}$ and/or $\{Z_k\}$ provides us with important visual information about an eventual change point for panel data. Values of statistics of $\{U_k\}$ and $\{Z_k\}$, when applied to our book leverage data, can be seen in Figure 7.

2 DESCRIPTION OF DATA AND ITS ANALYSIS

To illustrate our approach, quarterly accounting data describing behavior of more than 300 U.S. companies from different industries was selected from the well-known FAMA/FRENCH database. At the beginning of our analysis we had at our disposal financial quarterly time series describing, among others, the level of book leverage collected during the period 1Q 1983 to 4Q 2014. Note that all financial indicators are in USD. After careful inspection, however, only 46 companies remained for subsequent change-point analysis. Both rough and detailed industry classification according to the SIC Code of the respective companies can be found in Tables 2 and 3. The main reasons why we could not include data about remaining companies into our analysis were the following:

1. A company disappeared from the market before the end of the year 2014.
2. The data series was too short for the purposes of our analysis.
3. There were too many values missing from the data series for a given company.

If only a few observations were missing, we replaced them by their estimates obtained by combining neighboring observations. In practice we used linear interpolation. In this way we obtained a panel of 46 companies observed during the last 24 years, more precisely 96 quarters of the period 1Q 1991 through 4Q 2014. The variable of interest was the book leverage, i.e., the size of the debt with respect to debt plus shareholders' equity. These data will be used to illustrate our approach. Complete data we worked with is available upon request from the authors of this paper.

Table 1 Identifiers of analyzed companies

1004	1078	1104	1161	1166	1173	1230	1300	1327	1356
1380	1408	1468	1585	1602	1613	1618	1678	1686	1704
1728	1773	1783	1823	1864	1913	1920	1926	1968	1988
2044	2049	2055	2061	2086	2136	2154	2184	2220	2269
2282	2282	2285	2290	2312	2403	2411			

Source: Authors

Table 2 Rough categories of analyzed companies according to the SIC Code

SIC Code	Standard Industrial Classification	#
10–14	Mining	2
20–39	Manufacturing	32
40–49	Transportation & Public Utilities	3
50–51	Wholesale Trade	3
52–59	Retail Trade	3
70–89	Services	4

Source: Authors

Table 3 Detailed categories of analyzed companies according to the SIC Code

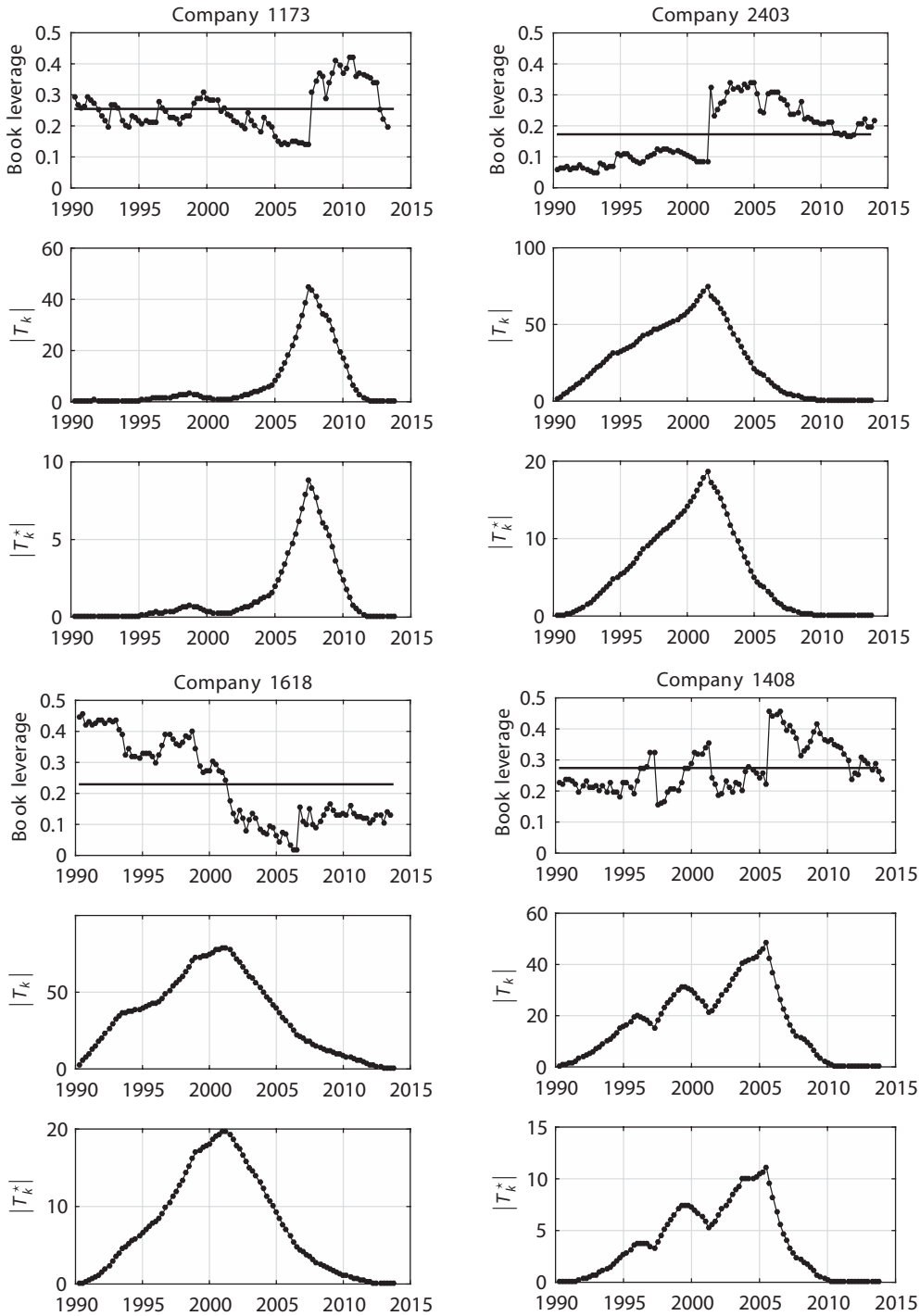
SIC Code	Standard Industrial Classification	#
10	Metal Mining	1
13	Oil and Gas Extraction	1
20	Food and Kindred Products	1
26	Paper and Allied Products	2
27	Printing, Publishing, and Allied Industries	1
28	Chemicals and Allied Products	7
29	Petroleum Refining and Related Industries	1
31	Leather and Leather Product	1
32	Stone, Clay, Glass, and Concrete Products	1
33	Primary Metal Industries	1
34	Fabricated Metal Products, except Machinery and Transportation Equipment	4
35	Industrial and Commercial Machinery and Computer Equipment	3
36	Electronic and other Electrical Equipment and Components, except Computer Equipment	4
37	Transportation Equipment	1
38	Measuring, Analyzing, and Controlling Instruments; Photographic, Medical and Optical Goods; Watches and Clocks	5
45	Transportation by Air	1
48	Communications	2
50	Wholesale Trade-Durable Goods	2
51	Wholesale Trade-Nondurable Goods	1
54	Food Stores	1
57	Home Furniture, Furnishings, and Equipment Stores	1
58	Eating and Drinking Places	1
72	Personal Services	1
73	Business Services	2
80	Health Services	1

Source: Authors

Typical representative behavior of the analyzed data can be seen in Figure 2. As an illustration, we included here companies exhibiting different financial strategies. While some studied time series exhibit a sudden shift in book leverage as in the case of companies 1173 and 2403, some others exhibit gradual change, such as in the data of company 1618. A typical example of several sudden changes is given by the data of company 1408. Finally, different levels of the average book leverage are illustrated by companies 1988 and 2184.

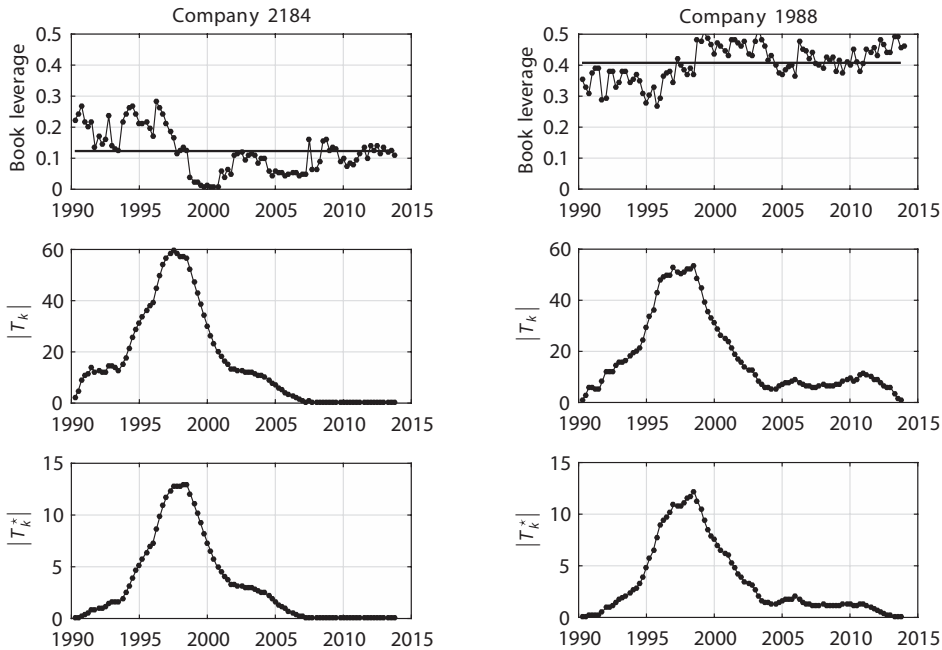
Notice that the scale for the book leverage is the same in all subfigures. On the other hand, this is not true for the scale of the values of the test statistics $\{|T_k|\}$ and $\{|T_k^*|\}$. The reason is a very high variability of the values of respective test statistics; if the same scale were used, then some figures would become unreadable.

Figure 2 Typical representatives of the analyzed data



Source: Authors

Figure 2 Typical representatives of the analyzed data – continuation

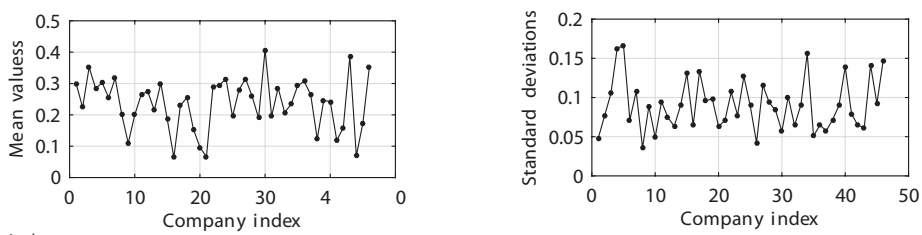


Source: Authors

First, the data passed a "visual inspection", which gave us an initial idea about the "behavior patterns" of individual companies. It is worth noticing that statistics $\{T_k\}$ and $\{T_k^*\}$ constructed for detection of sudden changes indicate a break also when the data exhibits a gradual change, as is the case of company 1618, see Figure 2. However, in such a case one must be careful when interpreting a course of $\{T_k\}$ and/or $\{T_k^*\}$, because the locations of corresponding maxima, i.e., statistics (6) and (7), do not necessarily correspond to the location of a change in behavior of the studied time series. For more details show to proceed in such a case see, e.g., papers Antoch et Hušková (1998) and Antoch et al. (2002).

Second, the mean value and standard deviation of the book leverage of each individual company has been calculated. The results can be seen in Figure 3. It appears that the mean values do not contain any outliers and follow our expectations. The same holds for the standard deviation values.

Figure 3 Mean values and standard deviations of the book leverage for individual companies

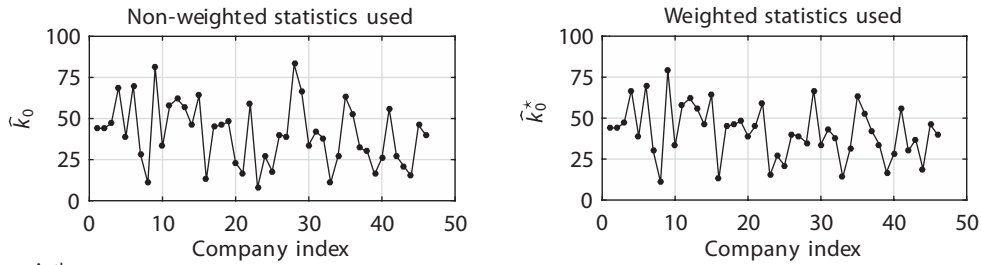


Source: Authors

Third, test statistics (4) and (5) suggested for the detection of a sudden change in the behavior of individual time series have been calculated for each company. It was a bit surprising that all test statistics for individual series are statistically significant on the 5% level for non-weighted statistic (4), and all but

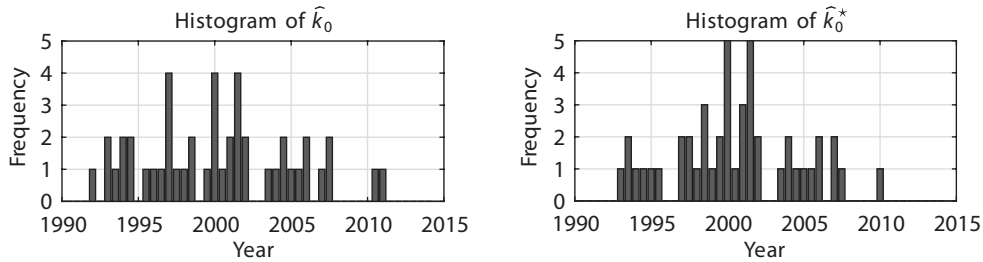
two test statistics for individual series are statistically significant on the 5% level when weighted statistic (5) has been used. Several companies exhibited two to three detectable changes. Therefore, we estimated the times of the change for each individual data time series using both statistic (6) and (7). Estimated change points are presented in Figure 4, and summarized using histograms in Figure 5. From Figures 4 and 5 we can see that the most change points have been detected around the year 2000, followed by the years 1997 and 2005. Against our expectations, this analysis has not shown any breaks around the time of the sub-prime crisis.

Figure 4 Estimated individual change points using Formulas (6) and (7) ($\widehat{k}_0 = 1$ corresponds to 1Q 1983, while $\widehat{k}_0 = 100$ corresponds to 4Q 2015)



Source: Authors

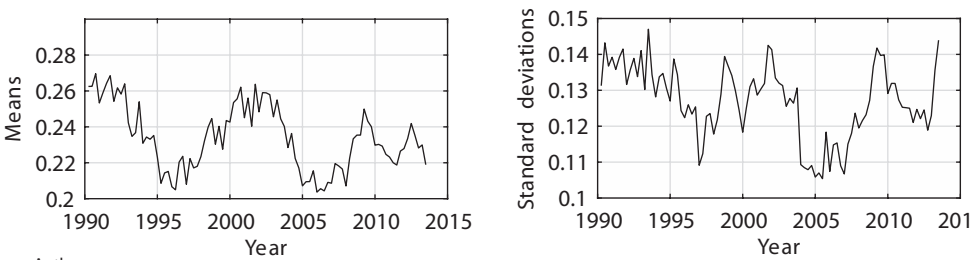
Figure 5 Histograms of estimated change points for estimators based on both non-weighted and weighted test statistics



Source: Authors

We also calculated averages and standard deviations of the set of analyzed companies during the period 1Q 1991 through 4Q 2014. The results are plotted in Figure 6. It is very interesting that the character of both means and standard deviations changes practically at the same time when many individual companies exhibited a sudden change in their book leverage levels.

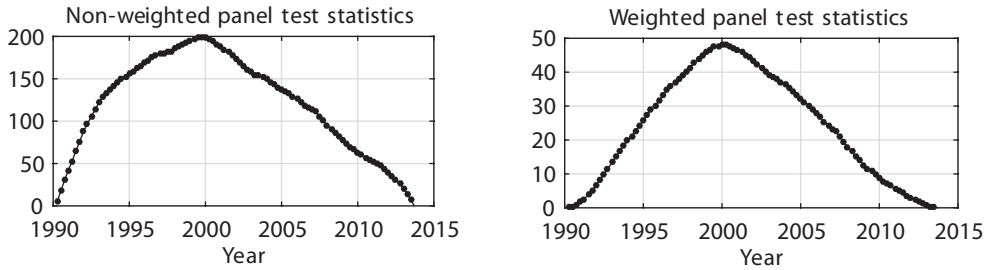
Figure 6 Average book leverage of actual portfolios during the time and the corresponding standard deviations



Source: Authors

Finally, we calculated panel test statistics (8) and (9). The results are presented in Figure 7. Both statistics indicate the change around the year 2000. The non-weighted panel test statistic also reflects the changes in the individual behavior of each company around the year 1997, compare Figures 4 and 5. The courses of both considered test statistics correspond to the analysis of individual companies.

Figure 7 Panel test statistics



Source: Authors

CONCLUSIONS

In our paper we describe analysis of stationarity (in the mean) of the book leverage data of 46 U.S. companies selected from different industries, see Tables 2 and 3. First, we analyzed each company separately, and then we analyzed all companies together using methods suggested for analysis of panel data. Moreover, we assumed that if there is a change at time k_0, then any series either changes at time k_0 or does not change at all.

Change point methods for panel data are proposed for deciding which of two hypotheses holds true. One hypothesis claims that the series do not change while the second claims that the series change at a certain unknown time. In the case of our data, none of these hypotheses seems to be true because each series exhibits a change, but at a different time. Most series changed near the year 2000, some also around 1997 and 2005, and as a consequence the hypothesis that all series are stationary was rejected.

The changes during the period 1997–2000 may reflect the Asian financial crisis that gripped much of East Asia, beginning in July 1997 and raised fears of a worldwide economic meltdown due to financial contagion. The Asian "flu" had also put pressure on the United States and Japan. Their markets did not collapse, but they were severely hit. On 27 October 1997, the Dow Jones industrial plunged 554 points or 7.2%, amid ongoing worries about the Asian economies. The New York Stock Exchange briefly suspended trading. The crisis led to a drop in consumer and spending confidence. Indirect effects included the dot-com bubble, and years later the housing bubble and the sub-prime mortgage crisis. Recall that many economists believe that the Asian crisis was created not by market psychology or technology, but by policies that distorted incentives within the lender-borrower relationship. For more details see, e.g., Goldstein (1998) or Muchhala (2007).

Another important goal of our analysis was to decide whether a sub-prime crisis in the period 2007–2008 caused a significant change in companies' behavior. Even though few analyzed series changed within this time period, see, e.g., company 1173, most series exhibit significant changes at some other times. Therefore, the panel statistics do not show a change in the period 2007–2008. We conclude that we did not discover a change of analyzed book leverage data of the selected U.S. companies due to the sub-prime crisis.

The fact that all 46 companies included in our study existed at least during 24 consecutive years, i.e., 1Q 1991 through 4Q 2014, and were able to report regularly, indicates that they represent rather powerful companies. Therefore, one should be careful about making sweeping generalizations on the whole U.S. economy, because in addition to these strong companies, many weaker ones were also present on the market.

ACKNOWLEDGEMENT

The work was supported by the Czech Science Foundation under Grant No. P403/15/09663S. The authors would like to thank Professor Jan Hanousek from CERGE and two unknown referees for their valuable comments that helped considerably improve the contents of the paper.

References

ANTOCH, J., GREGOIRE, G., HUŠKOVÁ, M. Tests for continuity of regression functions. *J. of Statistical Planning and Inference*, Vol. 137, 2007, pp. 753–777.

ANTOCH, J., GREGOIRE, G., JARUŠKOVÁ, D. Detection of structural changes in generalized linear models. *Statistics & Probability Letters*, Vol. 69, 2004, pp. 315–332.

ANTOCH, J., HANOUSEK, J., HORVÁTH, L., HUŠKOVÁ, M., WANG, S. Structural breaks in panel data: Large number of panels and short length time series. *Econometric Reviews* (submitted).

ANTOCH, J. AND HUŠKOVÁ, M. Procedures for detection of multiple changes in series of independent observations. In: HUŠKOVÁ, M., MANDL, P. eds. *5th Prague Symp. on Asymptotic Statistics,* Physica-Verlag, Heidelberg, 1994, pp. 3–20.

ANTOCH, J., HUŠKOVÁ, M., VERAVERBEKE, N. Change-point problem and bootstrap. *J. of Nonparametric Statistics*, Vol. 5, 1995, pp. 123–144.

ANTOCH, J. AND HUŠKOVÁ, M. Estimators of changes. In: GHOSH, S. AND DEKKER, M. *Nonparametrics, Asymptotics and Time Series*, New York: M. Dekker, 1998, pp. 533–578.

ANTOCH, J., HUŠKOVÁ, M., JANIC, A., LEDWINA, T. Data driven rank test for the change point problem. *Metrika,* Vol. 68, 2008, pp. 1–15.

ANTOCH, J., HUŠKOVÁ, M., JARUŠKOVÁ, D. Off-line statistical process control. In: LAURO et al. eds. *Multivariate Total Quality Control,* Springer/Physica-Verlag, Heidelberg, 2002, pp. 1–86. ISBN 3-7908-1383-4.

ANTOCH, J., HUŠKOVÁ, M., PRÁŠKOVÁ, Z. Effect of dependence on statistics for determination of change. *J. of Statistical Planning and Inference,* Vol. 60, 1997, pp. 291–310.

ANTOCH, J. AND JARUŠKOVÁ, D. Testing for multiple change points. *Computational Statistics*, Vol. 30, 2013, pp. 2161–2183.

BAI, J. AND PERRON, P. Estimating and testing linear models with multiple structural changes. *Econometrica*, Vol. 66, 1998, pp. 47–78.

BALTAGI, B. H. *Econometric Analysis of Panel Data.* 5th Ed. New York: John Wiley, 2013.

CSÖRGŐ, M. AND HORVÁTH, L. *Limit Theorems in Change-Point Analysis*. New York: J. Wiley, 1997.

DICK-NIELSEN, J., FELDHŰTTER, P., LANDO, D. Corporate bond liquidity before and after the onset of the subprime crisis. *J. of Financial Economics*, Vol. 103, 2012, pp. 471–492.

FAMA, E. F. AND FRENCH, K. Common risk factors in the returns on stocks and bonds. *J. of Financial Economics,* Vol. 33, 1993, pp. 3–56.

FAMA-FRENCH DATABASE [online]. <http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/index.htm>.

GOLDSTEIN, M. *The Asian Financial Crisis: Causes, Cures, and Systemic Implications.* Institute for International Economics, 1998. ISBN 0-88132-261-X.

HORVÁTH, L. AND RICE, G. Extensions of some classical methods in change point analysis. With discussion. *Test*, Vol. 23, 2014, pp. 219–290.

HUŠKOVÁ, M. AND HORVÁTH, L. Change-point detection in panel data. *J. of Time Series Analysis*, Vol. 33, 2012, pp. 631–648.

JAMES, B., JAMES, K. L., SIEGMUND, D. Tests for a change-point. *Biometrika*, Vol. 74, 1987, pp. 71–83.

JARUŠKOVÁ, D., Testing appearance of linear trend. *J. of Statistical Planning and Inference*, Vol. 70, 1998, pp. 263–276.

LEMMON, M. L., ROBERTS, M. R., ZANDER, J. F. Persistence and the cross-section of corporate capital structure. *J. of Finance*, Vol. 63, 2008, pp. 1575–1608.

MAC NEILL, I. B. Tests for change of parameter at unknown time and distribution of some related functionals of Brownian motion. *Annals of Statistics,* Vol. 2, 1974, pp. 950–962.

MUCHHALA, B. *Ten Years After: Revisiting the Asian Financial Crisis*. Woodrow Wilson International Center for Scholars, Asia Program, 2007. ISBN 1-933549-24-6.

SANTOS JOÃO, A. C. Bank corporate loan pricing following the subprime crisis. *Review of Finance Studies,* Vol. 24, 2011, pp. 1916–1943.

SIC CODE [online]. <http://siccode.com/en/siccode/list/directory>.

VOSTRIKOVA, L. Detecting disorder in multidimensional random processes. *Soviet Mathematical Doklady*, Vol. 24, 1981, pp. 55–59.

Statistical Inference Based on L-Moments

Tereza Šimková [1] | *Technical University of Liberec, Liberec, Czech Republic*

Abstract

To overcome drawbacks of central moments and comoment matrices usually used to characterize univariate and multivariate distributions, respectively, their generalization, termed L-moments, has been proposed. L-moments of all orders are defined for any random variable or vector with finite mean. L-moments have been widely employed in the past 20 years in statistical inference. The aim of the paper is to present the review of the theory of L-moments and to illustrate their application in parameter estimating and hypothesis testing. The problem of estimating the three-parameter generalized Pareto distribution's (GPD) parameters that is generally used in modelling extreme events is considered. A small simulation study is performed to show the superiority of the L-moment method in some cases. Because nowadays L-moments are often employed in estimating extreme events by regional approaches, the focus is on the key assumption of index-flood based regional frequency analysis (RFA), that is homogeneity testing. The benefits of the nonparametric L-moment homogeneity test are implemented on extreme meteorological events observed in the Czech Republic.[2]

Keywords

L-moment, parameter estimation, generalized Pareto distribution, homogeneity testing, precipitation extreme events, Czech Republic

JEL code

C02, C12, C13, C14, C15

INTRODUCTION

Moments, such as mean, variance, skewness and kurtosis, are traditionally used to describe features of a univariate distribution. Hosking (1990) introduced an alternative approach using L-moments, which are defined as certain linear combinations of order statistics. The main L-moments' advantage, in comparison to conventional moments, is their existence of all orders under only a finite mean assumption. When describing a multivariate distribution, the situation is very similar. The mean vector and covariance, coskewness and cokurtosis matrices with elements the covariance, coskewness and cokurtosis are the characteristics usually used to summarize features of a multivariate distribution. However, central comoments (covariance, coskewness, cokurtosis, etc.) are defined under finiteness of central moments of lower orders. To avoid this drawback, Serfling and Xiao (2007) proposed multivariate L-moments with elements the L-comoments as analogues to central comoments, without giving assumptions to finiteness of second and higher central moments.

L-moments, being measures of shape of a probability distribution, may be used for summarizing data drawn from both univariate and multivariate probability distributions. Besides description statistics,

[1] Technical University of Liberec, Faculty of Science, Humanities and Education, Department of Applied Mathematics, Studentská 1402/2, 461 17 Liberec 1, Czech Republic. E-mail: tereza.simkova@tul.cz, phone: (+420)485352973.

[2] This article is based on contribution at the conference *Robust 2016*.

L-moments play an important role also in inferential statistics. In the past 25 years the method of L-moments has been used as a convenient alternative to the traditional estimation method of moments and maximum likelihood method, mainly in hydrology, climatology and meteorology (e.g., Kyselý and Picek, 2007), but also in economics and socioeconomics (e.g., Bílková, 2014). The L-moments based estimates are obtained in a similar way as in the moment method, which means the population L-moments are equated to their corresponding sample quantities. Hosking (1990) gives parameter estimators of some common univariate distributions and highlights L-moments, because they sometime provide better estimates than the maximum likelihood method (particularly for small samples and heavy-tailed distributions). Several other studies have shown that the L-moment method in some cases outperforms also other estimation methods, including the well-known method of moments and relatively new methods of TL- and LQ-moments when estimating their parameters or high quantiles (Hosking, Wallis and Wood, 1985; Hosking and Wallis, 1987; Martins and Stedinger, 2000; Delicado and Goria, 2008; Šimková and Picek, 2016), as well. Moreover, L-moments based estimates are more tractable than maximum likelihood estimates. Besides parameter estimating, L-moments are also employed in hypothesis testing, particularly in RFA which yields reliable estimates of high quantiles of extreme events using data from sites, which have similar probability distributions. A univariate approach based on L-moments introduced by Hosking and Wallis (1997) has been routinely used in areas such as hydrology, climatology and meteorology, among others (Chen et al., 2006; Kyselý, Picek and Huth, 2006; Kyselý and Picek, 2007; Viglione, Laio and Claps, 2007; Noto and La Loggia, 2009; Kyselý, Gaál and Picek, 2011). Attention to multivariate RFA has been devoted recently in works of Chebana and Ouarda (2007), and Chebana and Ouarda (2009), in which the main steps of univariate index-flood based RFA of Hosking and Wallis (1997) were generalized using multivariate L-moments, copulas and quantile curves. Now multivariate RFA based on L-moments becomes popular in practice, because it improves analysis of the studied phenomenon by considering more available information (Chebana et al., 2009; Ben Aissia et al., 2015; Requena, Chebana and Mediero, 2016).

The paper gives a brief review of the theory of L-moments and their selected applications and uses already known methods to illustrate their use in specific examples in practice. First, the usefulness of L-moments is shown in the problem of estimating the GPD parameters often used in modelling extreme events. Although various techniques, such as the moments, L-moments or maximum likelihood methods, have been proposed in the literature for estimating parameters of a probability distribution, some of them are more accurate for data of certain properties as it has been already shown in several comparison studies (Hosking, Wallis and Wood, 1985; Hosking and Wallis, 1987; Martins and Stedinger, 2000; Delicado and Goria, 2008). Hence, a small simulation study is performed to compare several estimation methods and to show the superiority of the method based on L-moments for estimating GPD parameters in some cases. However, nowadays L-moments are mainly used in RFA to reliably estimate high quantiles of extreme events. The second illustration uses L-moments in hypothesis testing. Several papers have already dealt with both univariate and multivariate RFA based on L-moments of extreme precipitation events in the Czech Republic (Kyselý, Picek and Huth, 2006; Kyselý and Picek, 2007; Kyselý, Gaál and Picek, 2011; Šimková, Picek and Kyselý, in preparation). All these studies employed for homogeneity checking the parametric Hosking and Wallis (1997) or generalized Chebana and Ouarda (2007) L-moment homogeneity tests, which preceded model's parameters estimation and also relatively labouring selection of the best copula in the bivariate case. The nonparametric procedure is more powerful and easier to implement than the parametric one, because it does not require estimation of model's parameters nor specification of the copula in the multivariate case, and, hence, the homogeneity testing becomes simpler and quicker. Here, the benefits of the nonparametric homogeneity testing based on L-moments are implemented on bivariate extreme meteorological events observed in the Czech Republic. We will investigate whether the regions' homogeneity will be confirmed also by the nonparametric test, but in much easier way.

Because the nonparametric test of Masselot, Chebana and Ouarda (2016) has been introduced very recently, this is one of the first attempts of implementing the nonparametric procedure on the real-world data.

The paper is organized as follows: In the first section, the theory of both univariate and multivariate L-moments and their use in statistical inference are briefly reviewed. The use of L-moments in specific tasks, particularly in estimating the GPD parameters and nonparametric checking of regions' homogeneity formed in the area of the Czech Republic, and results obtained are presented in Section 2. The paper closes with summary section.

1 METHODOLOGY

1.1 Univariate L-moments

1.1.1 Population univariate L-moments

A population L-moment is defined to be a certain linear combination of order statistics (the letter L just emphasizes that the L-moment is a linear combination) which exists for any random variable with finite mean. Hosking (1990) defined the population L-moment of the *r*th order as a linear combination of the expectations of the order statistics $X_{1:n} \leq X_{2:n} \leq \ldots \leq X_{n:n}$ of a random sample of size *n* drawn from a univariate distribution of a random variable *X* with cumulative distribution function *F:*

$$\lambda_r = \frac{1}{r}\sum (-1)^k \binom{r-1}{k} EX_{r-k:r}, r = 1,2,\ldots \tag{1}$$

When comparing L-moments to conventional moments, L-moments have some merits, including their existence, uniqueness and robustness (because they are linear combinations). The Formula (1) may be rewritten to the form, which is useful particularly for computation of L-moments of a given probability distribution,

$$\lambda_r = \int_0^1 x(F) P^*_{r-1}(F(x))dF, r = 1,2,\ldots, \tag{2}$$

where:

$$P^*_r(t) = \sum_{i=0}^{r} (-1)^{r-i} \binom{r}{i} \binom{r+i}{i} t^i$$

is the *r*th shifted Legendre polynomial and *x(F)* is quantile function of a variable *X*. The first L-moment is just the mean of a random variable *X* and the second L-moment is equal to one-half of the Gini's mean difference statistic. Serfling and Xiao (2007) also present the expression of the second and higher order L-moments in the covariance representation as:

$$\lambda_r = \mathrm{cov}(X, P^*_{r-1}(F(X))), r \geq 2. \tag{3}$$

It is desirable to define dimensionless versions of higher L-moments, termed L-moment ratios, as:

$$\tau_r = \frac{\lambda_r}{\lambda_2}, r \geq 3.$$

Analogy of the coefficient of variation may be also defined in terms of L-moments as the ratio of the second L-moment λ_2 to the first L-moment λ_1

$$\tau = \frac{\lambda_2}{\lambda_1}.$$

The first two L-moments λ_1 and λ_2, termed L-location and L-scale, being measures of location and scale, and the third and fourth L-moment ratios τ_3 and τ_4, termed L-skewness and L-kurtosis, being measures of skewness and kurtosis, may be used for summarizing a univariate distribution. See Table 1 for the first four L-moments of some selected common univariate distributions which may be simply derived using the Formula (2).

Table 1 L-moments of several selected univariate distributions

Distribution	Quantile function	L-moments
Uniform	$x(F) = \alpha + (\beta - \alpha)F$	$\lambda_1 = \frac{1}{2}(\alpha + \beta), \lambda_2 = \frac{1}{6}(\beta - \alpha),$ $\lambda_3 = 0, \lambda_4 = 0$
Exponential	$x(F) = \xi - \alpha \log(1 - F)$	$\lambda_1 = \xi + \alpha, \lambda_2 = \frac{1}{2}\alpha, \lambda_3 = \frac{1}{6}\alpha,$ $\lambda_4 = \frac{1}{12}\alpha$
Normal	no explicit form, approximation used $x(F) \doteq \mu + 5.063\sigma[F^{0.135} - (1 - F)^{0.135}]$	$\lambda_1 = \mu, \lambda_2 = \pi^{-\frac{1}{2}}\sigma, \lambda_3 = 0,$ $\lambda_4 = 0.0702\sigma$
Logistic	$x(F) = \xi + \alpha \log\left(\frac{1 - F}{F}\right)$	$\lambda_1 = \xi, \lambda_2 = \alpha, \lambda_3 = 0, \lambda_4 = \frac{1}{6}\alpha$
Generalized Pareto	$x(F) = \xi + \frac{\alpha}{k}[1 - (1 - F)^k]$	$\lambda_1 = \xi + \frac{\alpha}{k+1},$ $\lambda_2 = \frac{\alpha}{(k+1)(k+2)},$ $\lambda_3 = \frac{\alpha(1-k)}{(k+1)(k+2)(k+3)},$ $\lambda_4 = \frac{\alpha(k-1)(k-2)}{(k+1)(k+2)(k+3)(k+4)}$
Generalized extreme-value	$x(F) = \xi + \frac{\alpha}{k}[1 - (-\log F)^k]$	$\lambda_1 = \xi + \frac{\alpha}{k} - \alpha\Gamma(k),$ $\lambda_2 = \alpha\Gamma(k)(1 - 2^{-k}),$ $\lambda_3 = \alpha\Gamma(k)(2 \cdot 3^{-k} + 3 \cdot 2^{-k} - 1)$ $\lambda_4 = \alpha\Gamma(k)(-5 \cdot 4^{-k} + 10 \cdot 3^{-k} - 6 \cdot 2^{-k} + 1)$

Source: Hosking (1990)

Although L-moments do not exist for distributions which have no finite mean (e.g., this happens for a Cauchy distribution, or generalized Pareto and generalized extreme-value distributions for certain values of the shape parameter), generalizations of L-moments, termed trimmed L-moments (abbreviated TL-moments) and LQ-moments, have been proposed. They always exist. See Elamir and Seheult (2003) and Mudholkar and Hutson (1998) for their definitions and properties.

1.1.2 Sample univariate L-moments

Because L-moments are defined for a probability distribution, they must be in practice estimated from an observed random sample drawn from an unknown probability distribution. The *r*th sample L-moment, being an unbiased estimator of the population L-moment λ_r, defined Hosking (1990) as a linear combination of the order sample $x_{1:n} \leq x_{2:n} \leq ... \leq x_{n:n}$ of size n

$$l_r = \binom{n}{r}^{-1} \sum_{1 \leq i_1} \sum_{< i_2 < \ldots} \cdots \sum_{< i_r \leq n} \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} x_{i_{r-k}:n}, r = 1,2,\ldots,n.$$

The first sample L-moment termed sample L-location, is equal to the sample mean, while the second sample L-moment is called sample L-scale.

Naturally, the L-moment coefficient of variation τ and L-moment ratios τ_r are estimated by the sample L-moment coefficient of variation and sample L-moment ratios given by:

$$t = \frac{l_2}{l_1}, t_r = \frac{l_r}{l_2}, r \geq 3. \tag{4}$$

Observed data may be alternatively summarized and described by the sample L-location l_1, L-scale l_2, L-skewness t_3 and L-kurtosis t_4.

1.1.3 Method of L-moments

Usually, the method of maximum likelihood and method of moments are used for estimation of parameters of a probability distribution. Following the same idea as in the case of method of moments, L-moments provide parameter estimates. Let *X* be a random variable with a probability density function $f(x; \theta_1, ..., \theta_k)$, where $\theta_1, ..., \theta_k$ are k unknown parameters. The unknown parameters are estimated by solving the system of equation which arise from matching the first k population L-moments with corresponding sample counterparts, i.e.,

$$\lambda_i = l_i, i = 1,\ldots,k. \tag{5}$$

Hosking (1990), and Hosking and Wallis (1997) give parameter estimates of selected common univariate probability distributions derived by the L-moment method. Parameter estimates of some univariate probability distributions are shown in Table 2.

Table 2 Parameter estimation of several selected univariate distributions

Distribution	Parameter estimation
Uniform	$\hat{\alpha} = l_1 - 3l_2, \hat{\beta} = 2l_1 - \hat{\alpha}$
Exponential	$\hat{\alpha} = 2l_2, \hat{\xi} = l_1 - \hat{\alpha}$
Normal	$\hat{\mu} = l_1, \hat{\sigma} = \sqrt{\pi} l_2$
Logistic	$\hat{\xi} = l_1, \hat{\alpha} = l_2$
Generalized Pareto	$\hat{k} = \frac{1 - 3t_3}{1 + t_3}, \hat{\sigma} = l_2(\hat{k} + 1)(\hat{k} + 2), \hat{\xi} = l_1 - \frac{\hat{\sigma}}{\hat{k} + 1}$
Generalized extreme-value	$\hat{k} \approx 7.859z + 2.9554z^2$, where $z = \frac{2}{3+t_3} - \log_3 2, \hat{\sigma} = \frac{l_2}{\Gamma(\hat{k})(1-2^{-\hat{k}})}, \hat{\xi} = l_1 - \frac{\hat{\sigma}}{\hat{k}} + \hat{\sigma}\Gamma(\hat{k})$

Source: Hosking (1990)

1.2 Multivariate L-moments

1.2.1 Population L-comoments

Serfling and Xiao (2007) defined L-comoments, which describe a multivariate distribution only under finite mean assumptions, analogously to the forms of central comoments and univariate L-moments in the covariance representation given by the Formula (3). Hence, it is worth remembering central comoments.

Let's have a bivariate random vector (X_1, X_2) with cumulative distribution function F, marginal distribution functions F_1, F_2, finite means μ_1, μ_2 and central moments $\mu_k^{(1)}, \mu_k^{(2)}, k \geq 2$. The rth central comoment of variable X_1 with respect to variable X_2 is defined as:

$$\xi_{r[12]} = \operatorname{cov}(X_1, (X_2 - \mu_1^{(2)})^{r-1}), r \geq 2.$$

The second, third and fourth central comoments $\xi_{2[12]}, \xi_{3[12]}, \xi_{4[12]}$ are covariance, coskewness and cokurtosis, respectively. Dimensionless versions of central comoments are given by:

$$\psi_{r[12]} = \frac{\xi_{r[12]}}{\sqrt{\mu_2^{(1)}(\mu_2^{(2)})^{r-1}}}, r \geq 2.$$

The second, third and fourth central rescaled comoments $\psi_{2[12]}, \psi_{3[12]}, \psi_{4[12]}$, are called correlation, coskewness and cokurtosis coefficients, respectively.

Let's have a bivariate random vector (X_1, X_2) with cumulative distribution function F, marginal distribution functions F_1, F_2 and finite means μ_1, μ_2. The rth L-comoment of variable X_1 with respect to variable X_2 (in this order) is defined as:

$$\lambda_{r[12]} = \operatorname{cov}(X_1, P_{r-1}^*(F_2(X_2))), r \geq 2,$$

(the version $\lambda_{r[21]}$ is defined similarly). Generally, $\lambda_{r[12]}$ and $\lambda_{r[21]}$ are not equal. Having $X_1 = X_2$, L-comoments reduce to univariate L-moments. The second to the fourth L-comoments may be regarded as alternatives to central comoments $\xi_{2[12]}, \xi_{3[12]}, \xi_{4[12]}$.

Scale-free versions of L-comoments, so-called L-comoment coefficients, are defined in similar way as L-moment coefficient of variation and L-moment ratios given by Formula (4):

$$\tau_{2[12]} = \frac{\lambda_{2[12]}}{\lambda_1^{(1)}}, \tau_{r[12]} = \frac{\lambda_{k[12]}}{\lambda_2^{(1)}}, r \geq 3.$$

Computation of population L-comoments may be simplified when variables X_1, X_2 meet certain conditions, particularly when variables are jointly distributed with affinely equivalent marginal distributions and one variable has linear regression on the other (for a detailed formulation see Proposition 3 in Serfling and Xiao (2007)).

1.2.2 Estimation of L-comoments

As it is in the case of univariate L-moments, L-comoments must be in practice estimated from an observed random sample drawn from an unknown multivariate distribution. This is made in terms of concomitants. Consider a sample $\{(x_i^1, x_i^2), 1 \leq i \leq n\}$ drawn from an unknown bivariate distribution. When the sample $\{x_1^2, ..., x_n^2\}$is sorted to a non-decreasing sequence, then the element of the sample $\{x_1^1, ..., x_n^1\}$ which is paired to the rth order statistic $x_{r:n}^2$ is called the concomitant of $x_{r:n}^2$ and denoted by $x_{[r:n]}^{12}$. The unbiased estimator of the rth L-comoment $\lambda_{r[12]}$ is defined as a linear combination of concomitants:

$$\hat{\lambda}_{r[12]} = \frac{1}{n}\sum_{r=1}^{n} w_{k:n}^{(r)} x_{[r:n]}^{12}, r \geq 2, \tag{6}$$

where:

$$w_{k:n}^{(r)} = \frac{1}{n}\sum_{j=0}^{\min\{r-1,k-1\}} (-1)^{r-j-1} \binom{r-1}{j}\binom{r-1+j}{j}\binom{n-1}{j}^{-1}\binom{k-1}{j}$$

are the weights.

1.2.3 Multivariate L-moments as L-comoment matrices

Consider a d-variate random vector $\boldsymbol{X} = (X_1, ..., X_d)$. The multivariate L-moment of the first order is the vector mean:

$$\Lambda_1 = E(X_1, \ldots, X_d),$$

while the second and higher orders multivariate L-moments are defined in a matrix form with elements the rth L-comoments of variables $X_i, X_j, 1 \le i,j \le d$,

$$\Lambda_r = \left(\lambda_{r[ij]}\right)_{d\times d}. \tag{7}$$

The second, third and fourth multivariate L-moments $\Lambda_2, \Lambda_3, \Lambda_4$ are termed L-covariance, L-coskewness and L-cokurtosis matrices, respectively. Scale-free versions of L-comoment matrices Λ_r, $r \ge 2$, labelled as the L-comoment coefficient matrices Λ_r^* consist of L-comoment coefficients of variables $X_i, X_j, 1 \le i,j \le d$,

$$\Lambda_r^* = \left(\tau_{r[ij]}\right)_{d\times d}.$$

The diagonal elements of matrices Λ_r and Λ_r^* are obviously the univariate L-moments and L-moment ratios.

Multivariate L-moments of three selected bivariate distributions are presented in Table 3.

Table 3 L-moments of several selected bivariate distributions

Distribution	Joint density function and L-moments
Normal	$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\},$ $x, y, \mu_1, \mu_2 \in R, \sigma_1, \sigma_2 > 0, \rho \in (-1, 1)$ $\Lambda_1 = (\mu_1, \mu_2), \Lambda_2 = \frac{1}{\sqrt{\pi}} \cdot M, \Lambda_3 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \Lambda_4 = 0.0702 \cdot$ M, where: $M = \begin{pmatrix} \sigma_1 & \rho\sigma_1 \\ \rho\sigma_2 & \sigma_2 \end{pmatrix}$
Pareto Type I	$f(x,y) = \frac{\alpha(\alpha+1)}{\sigma_1\sigma_2}\left(\frac{x}{\sigma_1} + \frac{y}{\sigma_2} - 1\right)^{-\alpha-2}, x \ge \sigma_1 > 0, y \ge \sigma_2 > 0, \alpha > 0$ $\Lambda_1 = \left(\frac{\alpha\sigma_1}{\alpha-1}, \frac{\alpha\sigma_2}{\alpha-1}\right), \Lambda_2 = \frac{1}{(\alpha-1)(2\alpha-1)} \cdot M, \Lambda_3 = \frac{\alpha+1}{(\alpha-1)(2\alpha-1)(3\alpha-1)} \cdot M, \Lambda_4 = \frac{(\alpha+1)(2\alpha+1)}{(\alpha-1)(2\alpha-1)(3\alpha-1)(4\alpha-1)} \cdot M$, where: $M = \begin{pmatrix} \alpha\sigma_1 & \sigma_1 \\ \sigma_2 & \alpha\sigma_2 \end{pmatrix}$
Pareto Type II	$f(x,y) = \frac{\alpha(\alpha+1)}{\sigma_1\sigma_2}\left(\frac{x-\mu_1}{\sigma_1} + \frac{y-\mu_2}{\sigma_2} + 1\right)^{-\alpha-2}, x > \mu_1, \mu_1 \in R, \sigma_1 > 0, y > \mu_2, \mu_2 \in R, \sigma_2 > 0, \alpha > 0$ $\Lambda_1 = \left(\mu_1 + \frac{\sigma_1}{\alpha-1}, \mu_2 + \frac{\sigma_2}{\alpha-1}\right), \Lambda_2 = \frac{1}{(\alpha-1)(2\alpha-1)} \cdot M, \Lambda_3 = \frac{\alpha+1}{(\alpha-1)(2\alpha-1)(3\alpha-1)} \cdot M, \Lambda_4 = \frac{(\alpha+1)(2\alpha+1)}{(\alpha-1)(2\alpha-1)(3\alpha-1)(4\alpha-1)} \cdot M$, where: $M = \begin{pmatrix} \alpha\sigma_1 & \sigma_1 \\ \sigma_2 & \alpha\sigma_2 \end{pmatrix}$

Source: Serfling and Xiao (2007), own construction

Multivariate L-moments are estimated by considering estimates of L-comoments, defined by Formula (6), in the matrix given in Formula (7). In the similar way, the L-comoment coefficient matrices are estimated.

1.3 Regional frequency analysis

Occurrence of extreme events, e.g., in hydrology, meteorology and climatology, among others, observed nowadays in many parts of the world may impact negatively on human society. Therefore, to reduce their impact, it is important to best estimate hiqh quantiles of a given return period. Although, in many practical applications the number of measurements is not sufficient to reliably estimate high quantiles (e.g., when annual maxima are measured), the same variable is often measured in other sites. In theses cases, RFA then provides more accurate estimates of high quantile in comparison to local approaches by taking into account data from different sites which have probability distributions similar to that site of interest. In index-flood based RFA introduced by Dalrymple (1960), a set of sites must meet a homogeneity condition, which means that all sites within a region have identical probability distributions apart from a site-specific scale factor (regions that meet this condition are termed homogeneous, otherwise they are termed heterogeneous). The multivariate quantile $\boldsymbol{Q}_i(F)$, $0 < F < 1$, at site is estimated as:

$$\widehat{\boldsymbol{Q}}_i(F) = \widehat{\boldsymbol{\mu}}_i\widehat{\boldsymbol{q}}(F),$$

where $\widehat{\boldsymbol{\mu}}_i$ corresponds to an estimate of the index-flood scale factor at site (usually estimated by sample mean or median) and $\widehat{\boldsymbol{q}}(\cdot)$ is an estimate of the regional growth curve which is a dimensionless quantile function of the probability distribution that is common to all sites in the region.

Generally, RFA consists of two main parts: 1) identification of homogeneous regions, and 2) quantile estimation. Here, the focus is on identification of homogeneous regions, i.e., groups of sites having probability distributions identical apart from a site-specific scale factor, because it is a key task in index-flood based RFA.

At first, a region must be proposed. Generally, it is recommended to form sites into groups on the basis of the site characteristics, such as the geographical location and elevation. They should not be based on at-site characteristics, because they are used for homogeneity testing as will be shown later. Several procedures have been proposed in the literature to form groups of similar sites, however, cluster analysis is the most practical method (Gordon, 1981; Everitt, 1993). When the region has been already proposed, it is desirable to decide whether it may be regarded as homogeneous, and, hence, the data from other sites may be utilized to obtain accurate estimates of high quantiles. Before executing the homogeneity test, the discordancy test should be applied to detect discordant sites.

1.3.1 L-moment discordancy test

The first step in any data analysis is to check that the data are suitable for the analysis. Two kinds of errors may occur: 1) data values are incorrect, and 2) the circumstances under which the data are ollected change over time. Sample L-moments may be used to reveal these errors. The aim of the L-moment discordancy test is to detect sites which are discordant with the group of sites as a whole.

Let's have a group of N sites, with site i having the record length n_i and sample L-comoment coefficient matrices $\Lambda_2^{*(i)},\Lambda_3^{*(i)},\Lambda_4^{*(i)}$. The discordancy measure is in the form:

$$D_i = \frac{1}{3}(U_i - \overline{U})^T S^{-1}(U_i - \overline{U}),$$

where:

$$U_i^T = [\Lambda_2^{*(i)} \Lambda_3^{*(i)} \Lambda_4^{*(i)}], S = \frac{1}{N-1} \sum_{i=1}^{N} (U_i - \overline{U})(U_i - \overline{U})^T, \overline{U} = \frac{1}{N} \sum_{i=1}^{N} U_i$$

(A^T denotes a transposed matrix A). A site i is regarded to be discordant if $||D_i|| > 2.6$. The sites flagged as discordant should be further checked.

1.3.2 L-moment homogeneity testing

First in the literature, the parametric multivariate L-moment homogeneity test was introduced as a generalization of the univariate Hosking and Wallis (1997) test to the multivariate case. However, this test is parametric which means that the multivariate probability distribution common to all sites must be specified. Moreover, the threshold for decision about homogeneity comes from simulations. To avoid drawbacks of the parametric test above-mentioned, Masselot, Chebana and Ouarda (2016) have introduced three alternatives, which differ in generating synthetic homogeneous regions and in the way of decision about homogeneity in comparison to the Chebana and Ouarda (2007) procedure. From all three alternatives proposed, here, the focus is only on the permutation nonparametric test which has the best performance according to the simulation study performed.

Parametric L-moment Homogeneity Test

1) Compute the statistic

$$V_{||\cdot||} = \left(\frac{\sum_{i=1}^{N} n_i \, || \, \Lambda_2^{*(i)} - \overline{\Lambda}_2^{*} \, ||^2}{\sum_{i=1}^{N} n_i} \right)^{1/2}. \tag{8}$$

where $\overline{\Lambda}_2^{*} = (\sum_{i=1}^{N} n_i \Lambda_2^{*(i)}) / \sum_{i=1}^{n} n_i$ is the regional L-covariance coefficient matrix and $|| \cdot ||$ an arbitrary matrix norm (Chebana and Ouarda (2007) recommend the spectral matrix norm).

2) Generate a large number N_{sim} of homogeneous regions (500 regions is enough according to Chebana and Ouarda, 2007) with N sites, each having the same record length as its real-world counterpart. To get a sample with univariate margins use copulas, and to get the desired sample use the quantile function of a four-parameter kappa distribution. A copula, being very flexible in modelling the dependence structure between variables, is a multivariate distribution function whose one-dimensional margins are uniform on the interval (0, 1). Sklar's theorem (Sklar, 1959) provides the relationship between a copula C, joint distribution function H and univariate margins F_1, ..., F_d:

$$H(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)) \forall x_1, \ldots, x_d \in R.$$

Copula theory has been well developed in the literature, see e.g., Joe (1997) and Nelsen (2006) for detailed copula foundations. The regional weighted parameters of the kappa distribution are estimated using the L-moment method proposed by Hosking (1990) by fitting the kappa distribution to the regional L-moment ratios (1, t_2^R, t_3^R, t_4^R), where t_k^R is a weighted mean of the at-site L-moment ratios for k = 2,3,4, while the regional copula parameter is obtained as a weighted mean of the at-site estimates using the at-site record lengths as weights.

3) Compute the statistic $V_{||\cdot||}^{(j)}$ defined by the Formula (8) on each of the simulated homogeneous regions, j = 1, ..., N_{sim}. Standardize $V_{||\cdot||}$ computed on the observed data in the first step by the mean μ and standard deviation σ of the computed values of $V_{||\cdot||}^{(j)}$ for a large number of simulated regions, i.e.,

$$H_{||\cdot||} = \frac{V_{||\cdot||} - \mu}{\sigma}.$$

4) Categorize the region: the region is declared to be homogeneous if $H_{||\cdot||} < 1$, acceptably homogeneous if $1 \leq H_{||\cdot||} < 2$, and definitely heterogeneous if $H_{||\cdot||} \geq 2$. Naturally, other measures used by Hosking and Wallis (1997) in the univariate L-moment homogeneity testing may be considered in the multivariate case to detect heterogeneity.

Nonparametric Permutation L-moment Homogeneity Test

1) Choose a significance level $\alpha \in (0,1)$.
2) Calculate $V_{||\cdot||}$ defined by the Formula (8) on the observed data as in the first step of the parametric test.
3) Generate a large number N_{sim} of homogeneous regions, which means to reassign randomly the pooled data between N sites while preserving the real-world at-site record lengths.
4) Compute the statistic $V^{(j)}_{||\cdot||}$ defined by the Formula (8) on each of the simulated regions, $j = 1, ..., N_{sim}$.
5) Compute the *p-value* given by:

$$p - value = \frac{1}{N_{sim}} \#\{V^{(j)}_{||\cdot||} > V_{||\cdot||}\}. \tag{9}$$

The null hypothesis of homogeneity is rejected if $p - value < \alpha$.

Although RFA has been traditionally used for analysis of extreme natural phenomena, it may be also employed in other fields in which extremes appear. In particular, it seems that modelling and estimation in finance, in which the interest in multivariate heavy-tailed distributions has increased, could be improved by using RFA.

2 RESULTS

In this section, results of specific applications of L-moments in two main fields of statistical inference are presented.

2.1 Estimation of GPD parameters

The choice of an appropriate estimation method of the GPD parameters is solved in this section. The three-parametric GPD with parameters ξ (location), σ (scale) and k (shape) has cumulative distribution function in the form:

$$F(x) = \begin{cases} 1 - [1 - \dfrac{k(x-\xi)}{\sigma}]^{1/k}, & k \neq 0. \\ 1 - e^{-\frac{x-\xi}{\sigma}}, & k = 0. \end{cases}$$

The L-moments estimates are compared to estimates obtained by the moment and maximum likelihood methods. Note that population L-moments of all orders exist for $k > -1$. Matching the first three population L-moments to their sample counterparts and solving the system of equations given in (5), L-moments based parameter estimates are obtained:

$$\hat{k} = \frac{1 - 3t_3}{1 + t_3}, \hat{\sigma} = l_2(\hat{k}+1)(\hat{k}+2), \hat{\xi} = l_1 - \frac{\hat{\sigma}}{\hat{k}+1}.$$

Analogously, moments based estimates are given by:

$$g = \frac{2(1-\hat{k})\sqrt{1+2\hat{k}}}{1+3\hat{k}}, \hat{\sigma} = s(\hat{k}+1)\sqrt{1+2\hat{k}}, \hat{\xi} = \bar{x} - \frac{\hat{\sigma}}{\hat{k}+\hat{\sigma}},$$

where $\bar{x}, s^2, g$ are sample mean, variance and skewness, respectively. First, the shape parameter k must be estimated by numerically solving the first equation. In the case of the maximum likelihood method, the location parameter ξ cannot be obtained, because the likelihood function is not bounded with respect to ξ, hence, the minimum value of the sample data is used as its estimate (Singh and Guo, 1995). The estimates of σ and k are achieved by solving equations:

$$\sum_{i=1}^{n} \frac{(x_i - \xi)/\sigma}{1 - k(x_i - \xi)/\sigma} = \frac{n}{1-k}, \sum_{i=1}^{n} \ln\left[1 - k(x_i - \xi)/\sigma\right] = -nk.$$

According to Šimková and Picek (2016) this study is focused only on values of the shape parameter k in the range $-0.4 \leq k \leq 0.4$ being typical for environmental applications. For each combination of the sample size n, $n \in \{20, 50, 100\}$, and the shape parameter k, $k \in \{-0.4, -0.2, 0, 0.2, 0.4\}$ 1 000 times a sample from the GPD is drawn, while the parameters of location and scale are fixed $\xi = 0$, $\sigma = 1$. The estimation methods are compared each other according to the sample mean squared error (MSE):

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2.$$

Table 4 Parameter estimates by moment (MM), L-moment (LM) and maximum likelihood (ML) methods: sample mean over 1 000 simulations (first row) and sample MSE (second row)

	$n = 20$			$n = 50$			$n = 100$		
	MM	LM	ML	MM	LM	ML	MM	LM	ML
ξ	0.661	−0.029	0.053	0.621	−0.014	0.020	0.578	−0.010	0.010
	0.894	0.015	*0.006*	0.607	0.005	*0.001*	0.444	0.003	*0.000*
σ	2.288	1.160	1.224	2.064	1.075	1.071	1.927	1.046	1.032
	3.456	*0.270*	0.347	2.091	0.089	*0.074*	1.320	0.047	*0.034*
$k = -0.4$	0.051	−0.255	−0.240	−0.082	−0.334	−0.351	−0.139	−0.360	−0.379
	0.231	*0.100*	0.182	0.112	*0.039*	0.048	0.075	*0.021*	*0.021*
ξ	0.299	−0.015	0.053	0.235	−0.006	0.020	0.189	−0.004	0.010
	0.200	0.011	*0.006*	0.108	0.004	*0.001*	0.068	0.002	*0.000*
σ	1.642	1.103	1.213	1.428	1.041	1.0671	1.322	1.021	1.030
	0.718	*0.206*	0.300	0.290	0.068	*0.065*	0.157	0.035	*0.029*
$k = -0.2$	0.121	−0.108	−0.028	−0.008	−0.166	−0.147	−0.062	−0.183	−0.176
	0.137	*0.086*	0.163	0.051	*0.031*	0.039	0.028	*0.016*	0.017
ξ	0.121	−0.013	0.051	0.061	−0.004	0.020	0.044	−0.003	0.010
	0.065	0.010	*0.005*	0.032	0.003	*0.001*	0.019	0.001	*0.000*
σ	1.357	1.085	1.225	1.175	1.025	1.061	1.114	1.016	1.035
	0.307	*0.191*	0.307	0.095	0.060	*0.058*	0.045	0.029	*0.026*
$k = 0$	0.221	0.059	0.190	0.100	0.014	0.055	0.063	0.008	0.028
	0.096	*0.087*	0.168	0.029	*0.027*	0.032	0.014	*0.013*	0.014
ξ	0.037	−0.006	0.051	0.014	−0.002	0.019	0.005	−0.001	0.010
	0.033	0.008	*0.005*	0.015	0.003	*0.001*	0.009	0.001	*0.000*
σ	1.187	1.056	1.220	1.076	1.021	1.067	1.039	1.009	1.032
	0.165	*0.152*	0.251	0.054	0.051	*0.050*	0.027	0.026	*0.023*
$k = 0.2$	0.345	0.240	0.416	0.254	0.212	0.269	0.226	0.204	0.234
	0.070	0.078	0.157	*0.021*	0.025	0.028	*0.010*	0.013	0.011
ξ	0.010	−0.004	0.051	0.003	−0.002	0.019	0.000	−0.001	0.010
	0.020	0.007	*0.005*	0.008	0.002	*0.001*	0.005	0.001	*0.000*
σ	1.116	1.051	1.209	1.043	1.020	1.072	1.021	1.009	1.036
	0.133	0.142	0.197	*0.046*	0.049	*0.046*	0.023	0.025	*0.020*
$k = 0.4$	0.500	0.431	0.634	0.434	0.410	0.482	0.415	0.403	0.442
	0.070	0.088	0.145	*0.022*	0.029	0.028	*0.011*	0.014	*0.011*

Source: Own construction

The smallest the MSE the best the estimator is. Computations are executed in the software R (R Core Team, 2014).

Table 4 compares performance of the moments, L-moments and maximum likelihood methods (the minimum MSE value is highlighted by italics). It can be concluded that the maximum likelihood method provides the best estimate of the location parameter, and also of the scale parameter for moderate sample sizes $n = 50$, 100, while the L-moment method outperforms the moment method for small sample size $n = 20$. When estimating the shape parameter, the L-moment based estimator is recommended for heavier tails ($k \leq 0$), while the moment method yields estimates with the smallest MSE for light tails ($k > 0$).

2.2 Nonparametric homogeneity testing in bivariate RFA of extreme precipitation events

Bivariate parametric homogeneity testing based on L-moments has been already applied to data observed at meteorological stations located in the area of the Czech Republic in the study of Šimková, Picek and Kyselý (in preparation). They found out that six regions formed in the area may be regarded as homogeneous with accordance to bivariate distribution function with components the 1- and 5-day maximum annual precipitation totals. Hence, this finding justifies to use data from an entire region for estimating quantiles in any target site in region. However, the procedure used required the user interventions: a bivariate copula specification, and kappa distribution and copula parameters estimation. Here, the homogeneity of regions is also checked by the nonparametric permutation test proposed recently by Masselot, Chebana and Ouarda (2016), which is easy to apply. We want to test the null hypothesis:

H_0 Region is homogeneous,

against the alternative:

H_1 Region is not homogeneous,

on the 5% significance level.

Maximum annual 1- and 5-day precipitation amounts measured mostly from 1961 to 2007 at 210 stations covering the area of the Czech Republic are used as the input dataset. The data were provided

Figure 1 Delineation of stations into six regions

Source: Šimková (accepted)

by the Czech Hydrometeorological Institute (CHMI), where they underwent basic quality checking. Kyselý (2009) also checked thoroughly the data for errors and missing readings. Delineation of the stations to regions shown in Figure 1 is that presented first in the study of Šimková (accepted). Basic information concerning the datasets for each region is summarized in Table 5. See Šimková (accepted) for more description of regions studied.

Table 5 Information on the input datasets

Region	1	2	3	4	5a	5b
Number of stations	75	79	33	16	4	4
Overall record length	3 438	3 633	1 508	719	188	141
Minimal record length (years)	33	37	33	36	47	47
Maximal record length (years)	47	47	47	47	47	47
Average record length (years)	45.8	46.0	45.7	44.9	47	47
Altitude range (m a.s.l.)	[150, 400]	[410, 1 118]	[220, 1 490]	[255, 572]	[315, 440]	[398, 778]
Average altitude (m a.s.l.)	270.1	550.3	411.1	412.6	361	523.3

Source: Šimková, Picek and Kyselý (in preparation)

The problem of determining discordant sites and their retention in regions have been already discussed by Šimková, Picek and Kyselý (in preparation). To estimate p-values given by Formula (9), 500 synthetic regions were generated by permuting bivariate data between sites, while the values of $V_{||\cdot||}$ have been already calculated on real observed data by Šimková, Picek and Kyselý (in preparation). Table 6 shows the values of $V_{||\cdot||}$, and compares the results of parametric and nonparametric homogeneity testing via the heterogeneity measures and p-values obtained. Values of the heterogeneity measure $H_{||\cdot||}$ are those presented by Šimková, Picek and Kyselý (in preparation). The parametric test gives evidence about homogeneity of all regions because $H_{||\cdot||}$ values are less than 2, while the nonparametric version rejects the null hypothesis H_0 of homogeneity for region 1 on the 5% significance level.

Table 6 Homogeneity testing results

Region	$V_{\|\cdot\|}$	$H_{\|\cdot\|}$	*p-value*
1	0.0603	1.4884	0.006
2	0.0570	1.1316	0.226
3	0.0536	–1.3857	0.994
4	0.0551	0.7111	0.090
5a	0.0181	–1.5305	0.976
5b	0.0278	–0.7635	0.770

Source: Šimková, Picek and Kyselý (in preparation), own construction

Hence, region 1 should be redefined. Because Šimková, Picek and Kyselý (in preparation) proposed region 1 as a unification of three smaller regions labelled 1a, 1b and 1c, presented in the study of Kyselý, Gaál and Picek (2011), the original delineation could be now considered. Homogeneity of these smaller regions has been checked again and they have finally met the homogeneity condition.

Table 7 Homogeneity testing results for redefined region 1

Region	$V_{\|\cdot\|}$	$H_{\|\cdot\|}$	*p-value*
1a	0.0364	–2.4479	0.996
1b	0.0590	0.8400	0.116
1c	0.0426	–0.8644	0.792

Source: Own construction

CONCLUSION

Alternatives to traditionally used moments and comoments labelled L-moments, which exist under only finite mean assumptions, have been introduced in the paper presented. L-moments, being measures of shape of a probability distribution, may be used to describe a probability distribution and to summarize sample data. Population L-moments of several selected both univariate and bivariate distributions have been also presented. The paper also shows selected already established L-moments based techniques and implements them in particular tasks of statistical inference.

The problem of estimating GPD parameters has been resolved. In a small comparison simulation study, in which three parameters of the GPD were estimated, it has been shown that the method based on L-moments outperforms other usually used estimation methods, such as the maximum likelihood and moments methods. This happens particularly for heavier tailed distributions and small to moderate samples. These results are consistent with those obtained for other probability distributions.

The benefits of the nonparametric test based on L-moments have been applied for regions formed in the area of the Czech Republic. The results obtained by the nonparametric test have confirmed those obtained by the parametric one (except one region), but in a much shorter time and without estimating parameters and selection of a suitable bivariate copula family, which is substantially more advantageous. Although RFA has been traditionally used for analysis of natural phenomena, such as floods and precipitation, nothing prevents to use it also for example in finance or economics, because extremes also appear there.

ACKNOWLEDGMENT

The study was supported by the Czech Science Foundation under project 15-00243S and by the Student Grant Competition under project 21116 at the Faculty of Science, Humanities and Education, Technical University of Liberec. The author would like to thank two anonymous reviewers for valuable comments which helped to improve the paper.

References

BEN AISSIA, M. A., CHEBANA, F., OUARDA, T. B. M. J., BRUNEAU, P., BARBET, M. Bivariate Index-flood Model for a Northern Case Study. *Hydrological Sciences Journal*, 2014. DOI: 10.1080/02626667.2013.87517.

BÍLKOVÁ, D. Alternative Means of Statistical Data Analysis: L-Moments and TL-Moments of Probability Distributions [online]. *Statistika: Statistics and Economy Journal*, 2014, 2, pp. 77–94.

CHEBANA, F. AND OUARDA, T. B. M. J. Multivariate L-moment homogeneity test. *Water Resources Research*, 2007, 43. DOI: 10.1029/2006WR005639.

CHEBANA, F. AND OUARDA, T. B. M. J. Index-flood Based Multivariate Regional Frequency Analysis. *Water Resources Research*, 2009, 45. DOI: 10.1029/2008WR007490.

CHEBANA, F., OUARDA, T. B. M. J., FAGHERAZZI, L., BRUNEAU, P., BARBET, EL ADLOUNI, S., LATRAVERSE, M. Multivariate Homogeneity Testing in a Northern Case Study in the Province of Quebec, Canada. *Hydrological Processes*, 2009, 23(12), pp. 1690–1700.

CHEN, Y. D., HUANG, G., SHAO, Q., XU, C.-Y. Regional Analysis of Low Flow using L-moments for Dongjiang Basin, South China. *Hydrological Sciences Journal*, 2006, 51(6), pp. 1051–1064.

DALRYMPLE, T. Flood Frequency Analyses. *Water Supply Paper 1543-A*, U.S. Geological Survey.

DELICADO, P. AND GORIA, M. N. A Small Sample Comparison of Maximum Likelihood, Moments and L-moments Methods for the Asymmetric Exponential Power Distribution. *Computational Statistics and Data Analysis*, 2008, 52(3), pp. 1661–1673.

ELAMIR, E. A. H. AND SEHEULT, A. H. Trimmed L-moments. *Computational Statistics and Data Analysis*, 2003, 43(3), pp. 299–314.

EVERITT, B. S. *Cluster Analysis*. London: Edward Arnold, 1993.

GORDON, A. D. *Classification: Methods for the Exploratory Analysis of Multivariate Data*. London: Chapmann & Hall, 1981.

HOSKING, J. R. M. L-moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of Royal Statistical Society (Series B)*, 1990, 52, pp. 105–124.

HOSKING, J. R. M. AND WALLIS, J. R. Parameter and Quantile Estimation for the Generalized Pareto Distribution. *Technometrics*, 1987, 29, pp. 339–349.

HOSKING, J. R. M. AND WALLIS, J. R. *Regional Frequency Analysis: An Approach Based on L-moments*. Cambridge: Cambridge University Press, 1997.

HOSKING, J. R. M., WALLIS, J. R., WOOD, E. F. Estimation of the Generalized Extreme-value Distribution by the Method of Probability-weighted Moments. *Technometrics*, 1985, 27, pp. 251–261.

JOE, H. *Multivariate Models and Dependence Concepts*. London: Chapman & Hall, 1997.

KYSELÝ, J. Trends in Heavy Precipitation in the Czech Republic over 1961–2005. *International Journal of Climatology*, 2009, 29, pp. 1745–1758.

KYSELÝ, J. AND PICEK, J. Regional Growth Curves and Improved Design Value Estimates of Extreme Precipitation Events in Czech Republic. *Climate Research*, 2007, 33, pp. 243–255.

KYSELÝ, J., GAÁL, L., PICEK, J. Comparison of Regional and At-site Approaches to Modelling Probabilities of Heavy Precipitation. *International Journal of Climatology*, 2011, 31, pp. 1457–1472.

KYSELÝ, J., PICEK, J., HUTH, R. Formation of Homogeneous Regions for Regional Frequency Analysis of Extreme Precipitation Events in the Czech Republic. *Studia Geophysica et Geodaetica*, 2006, 51, pp. 327–344.

MARTINS, E. S. AND STEDINGER, J. R. Generalized Maximum-likelihood Generalized Extreme-value Quantile Estimators for Hydrologic Data. *Water Resources Research*, 2000, 36(3), 737–744.

MASSELOT, P., CHEBANA, F., OUARDA, T. B. M. J. Fast and Direct Nonparametric Procedures in the L-moment Homogeneity Test. *Stochastic Environmental Research and Risk Assessment*, 2016. DOI: 10.1007/s00477-016-1248-0.

MUDHOLKAR, G. S., HUTSON, A. D. LQ-moments: Analogs of L-moments. *Journal of Statistical Planning and Inference*, 1998, 71, pp. 191–208.

NELSEN, R. B. *An Introduction to Copulas*. New York: Springer-Verlag New York, 2006.

NOTO, L. V., LA LOGGIA, G. *Use of L-Moments Approach for Regional Flood Frequency Analysis in Sicily, Italy*. Water Resources Management, 2009, 23, pp. 2207–2229.

R CORE TEAM. R: A Language and Environment for Statistical Computing (Version 3.3.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing, 2016.

REQUENA, A. I., MEDIERO, L., GARROTEL, L. A Complete Procedure for Multivariate Index-flood Model Application. *Journal of Hydrology*, 2016, 535, pp. 559–580.

SERFLING, R. AND XIAO, P. A Contribution to Multivariate L-comoments: L-comoment Matrices. *Journal of Multivariate Analysis*, 2007, 98, pp. 1765–1781.

SINGH, V. P. AND GUO, H. Parameter Estimation for 3-parameter Generalized Pareto Distribution by the Principle of Maximum Entropy (POME). *HydrologicaL Sciences Journal*, 1995, 40(2), pp. 165–181.

SKLAR, A. Fonctions de Rèpartition à *n* Dimensions et Leurs Marges. *Publications de l'Institut de Statistique de L'Universitè de Paris*, 1959, 8, pp. 229–231.

ŠIMKOVÁ, T. L-moment Homogeneity Test in Trivariate Regional Frequency Analysis of Extreme Precipitation Events. *Meteorological Applications* (accepted).

ŠIMKOVÁ, T. AND PICEK, J. A Comparison of L-, Lq-, Tl-Moment and Maximum Likelihood High Quantile Estimates of the Gpd and Gev Distribution. *Communications in Statistics – Simulation and Computation*, 2016. DOI: 10.1080/03610918.2016.1188206.

ŠIMKOVÁ, T., PICEK, J., KYSELÝ, J. *Bivariate Regional Frequency Analysis of Extreme Precipitation Events in the Czech Republic* (in preparation).

VIGLIONE, A., LAIO, F., CLAPS, P. A *Comparison of Homogeneity Tests for Regional Frequency Analysis*. Water Resources Research, 2007, 43(3). DOI:10.1029/2006WR005095.

Kriging Methodology and Its Development in Forecasting Econometric Time Series

Andrej Gajdoš [1] | *Pavol Jozef Šafárik University in Košice, Slovak Republic*
Martina Hančová [2] | *Pavol Jozef Šafárik University in Košice, Slovak Republic*
Jozef Hanč [3] | *Pavol Jozef Šafárik University in Košice, Slovak Republic*

Abstract

One of the approaches for forecasting future values of a time series or unknown spatial data is kriging. The main objective of the paper is to introduce a general scheme of kriging in forecasting econometric time series using a family of linear regression time series models (shortly named as FDSLRM) which apply regression not only to a trend but also to a random component of the observed time series. Simultaneously performing a Monte Carlo simulation study with a real electricity consumption dataset in the R computational language and environment, we investigate the well-known problem of "negative" estimates of variance components when kriging predictions fail. Our following theoretical analysis, including also the modern apparatus of advanced multivariate statistics, gives us the formulation and proof of a general theorem about the explicit form of moments (up to sixth order) for a Gaussian time series observation. This result provides a basis for further theoretical and computational research in the kriging methodology development.[4]

Keywords

Forecasting models, linear regression models, best linear unbiased prediction, approximation of mean squared error, moments of random vectors

JEL code

C10, C53, C59, C60, Q47

INTRODUCTION

Data of many economic, financial, insurance or business variables can be generally considered as time series datasets – sets of observations tracking the same type of information at multiple points in time. Modern time-series econometrics, representing an interconnection of mathematical, statistical and computer methods, allows us to model, forecast, interpret and describe various real phenomena dealing with these types of data (Andersen et al., 2009; Box et al., 2008; Brockwell and Davis, 2006; Cipra, 2013; Enders, 2014; Tsay, 2010). Last twenty-five years brought notable advances in the time-series econometrics (Escobari, Ngo, 2014) and moreover, its applications and tools led to several Nobel Prize Awards

[1] Faculty of Science, Department of Economic and Financial Mathematics, Jesenná 5, 04001 Košice, Slovak Republic. Corresponding author: e-mail: andrej.gajdos@student.upjs.sk, phone: (+421)904002040.
[2] Faculty of Science, Department of Economic and Financial Mathematics, Jesenná 5, 04001 Košice, Slovak Republic. E-mail: martina.hancova@upjs.sk.
[3] Faculty of Science, Institute of Physics, Park Angelinum 9, 04001 Košice, Slovak Republic.
[4] This article is based on contribution at the conference *Robust 2016.*

in Economics – namely R. S. Shiller, E. F. Fama and L. P. Larsen in 2013, Ch. A. Sims in 2011 as well as R. F. Engle III and C. W. J. Granger in 2003.

One of the most important areas of time series theory application is the forecasting which solves a task how to predict future values of a time series from its current and past values (Hyndman and Athanasopoulos, 2014). From a practical point of view, information obtained by forecasting provides a crucial knowledge for effective and efficient planning or decision making. The Box-Jenkins methodology (Box et al., 2008; Cipra, 2013; Enders, 2014; Tsay, 2010), belonging to the most popular methodologies for modeling and forecasting econometric time series data (see current real econometric applications e.g. in Pošta, Pikhart, 2012; Salamaga, 2015; Šimpach, 2015) is based fundamentally on ARMA, ARIMA models or their vector counterparts (VARMA, VARIMA). But there exist other advanced and powerful forecasting alternatives such as exponential smoothing methods (Cipra, 2013; Hyndman and Athanasopoulos, 2014), neural networks models (Andersen et al., 2009; Crone et al., 2011; Fomby and Terrell, 2006), linear regression models (Brockwell and Davis, 2006; Chatterjee and Hadi, 2012; Cipra, 2013; Enders, 2014; Štulajter, 2002) or dynamic regression models (Pankratz, 1991; Shumway and Stoffer, 2011).

The prediction theory using linear regression models called kriging (Christensen, 2001; Cressie and Wikle, 2011; Moore, 2001; Stein, 1999; Štulajter, 2002) represents a process of finding the optimal linear prediction for random processes or random fields. The process is based on modeling in an appropriate general class of linear regression models where the following analytical or numerical optimization finds out the best unbiased linear predictor (BLUP) on a set of all linear unbiased predictors. The optimization criterion is a minimization of the mean squared error (MSE) among considered predictors.

Finally it is worth to mention that although the kriging was originally developed for predictions in spatial data (geostatistics and meteorology, Cressie, 1993), the idea of the BLUP brings fruitful results in a much broader set of problems (Harville, 2008; Murphy, 2012; Rao and Molina, 2015; Robinson, 1991), e.g. small-area estimation in economics, the prediction of breeding values in genetics, the estimation of treatment contrasts (e.g. in drug development, agriculture or manufacturing), the analysis of longitudinal data, insurance credibility theory, noise removing from images or machine learning.

Our paper deals with an application of kriging in forecasting econometric time series. Since kriging is not well-known in econometric journals and literature, the first section summarizes a general framework how the kriging methodology works. To not be distracted by many technical details and to focus on main ideas, we illustrate each step of kriging using a real econometric time series dataset dealing with electricity consumption, and reducing the number of used formulas as much as possible (an interested reader will find explicit references for all data and formulas). The illustrative example brings us naturally to a problem of a kriging failure when standard computational methods dealing with considered estimates of nonnegative variance parameters give us negative values. The second section continues in this generally well-known estimation problem. Here we numerically manifest the practical commonness (non-rareness) of this situation by a simulation study numerically quantifying a relative occurrence of explored cases. In the final third section of the paper, we analyze the mentioned problem in the broader context of theoretical developments in kriging methodology using appropriate advanced methods of multivariate statistics. Our analysis results in the formulation and proof of a general theorem about the explicit form of moments for a Gaussian time series observation.

As for numerical calculations, we carried out our computational research producing final results (tables and figures) of the paper in the R statistical computing language (<https://www.r-project.org>; Chambers, 2008; R Development Core Team, 2016) in a powerful integrated development environment called RStudio (<https://www.rstudio.com>; Verzani, 2011). At present, the free and open source R computational environment rapidly improves its capabilities (now there are almost 10 000 statistical packages) which R ranks as one of the best statistical tools for the high-quality computational time series research (McLeod et al., 2012).

1 FORECASTING TIME SERIES USING KRIGING

Forecasting time series within the framework of kriging consists of the following stages (Christensen, 2001; Stein, 1999; Štulajter, 2002): (i) selecting sufficiently general and broad class of linear regression models; (ii) obtaining an empirical realization of a given time series and its modeling; (iii) choosing predicted time series values and finding the BLUP for them; (iv) estimating model parameters on which the BLUP depends and using empirical ("plug-in") BLUP; and finally (v) exploring the impact of the estimation on properties, especially mean squared error, of the BLUP. Let us briefly illustrate this scheme in the case of a real econometric dataset which also brings us naturally to our research problem.

1.1 The first stage – a general class of models

As we mentioned above, in the first stage of kriging we select some general class of linear regression models. In our research, we are concerned with the so-called finite discrete spectrum linear regression models (FDSLRM) – a class of time series models whose mean values (trend) are given by linear regression and random components (error terms) are a linear combination of uncorrelated zero-mean random variables and white noise, which together can be interpreted in terms of finite discrete spectrum (Priestley, 2004).

This parametric family of time series models, a direct extension of classical regression models with many practical applications, was introduced in 2002–2003 by Štulajter (2002, 2003). Especially the monograph from 2002 focusing on forecasting econometric time series in terms of kriging has started a mathematical and statistical research of FDSLRM dealing with its properties and applications (Hančová, 2008, 2011; Hančová et al., 2015; Harman and Štulajter, 2010; Štulajter, 2007; Štulajter and Witkovský, 2004). The exact formal definition of FDSLRM is the following:

A model of time series $X(.)$ is said to be the finite discrete spectrum linear regression model (FDSLRM) iff $X(.)$ satisfies:

$$X(t) = \sum_{i=1}^{k} \beta_i f_i(t) + \sum_{j=1}^{l} Y_j v_j(t) + w(t);\ t \in \mathcal{T}, \tag{1}$$

where:

$\mathcal{T}$ representing a time domain is a countable subset of the real line $\mathbb{E}^1$,

k a l are fixed known non-negative integers, i.e. $k, l \in \mathbb{N}_0$, ,

$\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)' \in \mathbb{E}^k$ is a vector of regression parameters,

$\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_l)'$ is a random vector with $E(\boldsymbol{Y}) = 0$ and with covariance matrix $Cov(\boldsymbol{Y}) = diag\{\sigma_j^2\}$ of size $l \times l$, where $\sigma_j^2 \geq 0, j = 1,2, \dots, l$,

$f_i(.);\ i \in \{1, 2, \dots, k\}$ and $v_j(.);\ j \in \{1, 2, \dots, l\}$ are real functions defined on $\mathbb{E}^1$,

$w(.)$ is a white noise uncorrelated with $\boldsymbol{Y}$ and with dispersion $D[w(t)]=\sigma^2 > 0$.

In FDSLRM applications (Štulajter, 2003, 2007; Štulajter and Witkovský, 2004) the most frequently considered time domain set $\mathcal{T}$ is the set of natural numbers $\mathbb{N} = \{1,2, \dots\}$.

For further considerations, we remind one of basic properties of the FDSLRM (Štulajter, 2003), which says that a finite FDSLRM observation $\boldsymbol{X} = \big(X(1), X(2), \dots, X(n)\big)'$, $n \in \mathbb{N}$ satisfies a linear mixed model (LMM) of the form:

$$\boldsymbol{X} = \mathrm{F}\boldsymbol{\beta} + \mathrm{V}\boldsymbol{Y} + \boldsymbol{w} \text{ with } E(\boldsymbol{w}) = \boldsymbol{0}, Cov(\boldsymbol{w}) = \sigma^2 \mathrm{I}_n, Cov(\boldsymbol{Y}, \boldsymbol{w}) = \boldsymbol{0}, \tag{2}$$

where matrices F (size $n \times k$) and V (size $n \times l$) are known design matrices given by values of functions $f_i(.), v_j(.)$ for times $t = 1,2,\dots,n$ and $\boldsymbol{w} = (w(1),\dots,w(n))'$ stands for a finite n-dimensional white noise observation. In the language of LMM terminology $\boldsymbol{\beta}$ would represent the k-vector of fixed effects and

the random component would depend on l-vector $\boldsymbol{Y}$ of random effects and n-vector $\boldsymbol{w}$ of random errors. This fundamental FDSLRM property allows us to apply many results and mathematical techniques of LMM methodology (e.g. Demidenko, 2013; McCulloch et al., 2008; Searle et al., 2006; Witkovský, 2012).

The last remark in our formal introduction of FDSLRM deals with the variance parameters of $\boldsymbol{Y}$ and $w(.)$. It is common to describe the parameters by one vector $\boldsymbol{\nu} = (\sigma^2, \sigma_1^2, \ldots, \sigma_l^2)'$, so $\boldsymbol{\nu}$ becomes an element of the parametric space $\Upsilon = (0, \infty) \times [0, \infty)^l$. Because of several practical or theoretical reasons (similar as in the case of LMM; see Remark 1), it is common to work only with a restricted space $\Upsilon^* = (0, \infty)^{l+1}$.

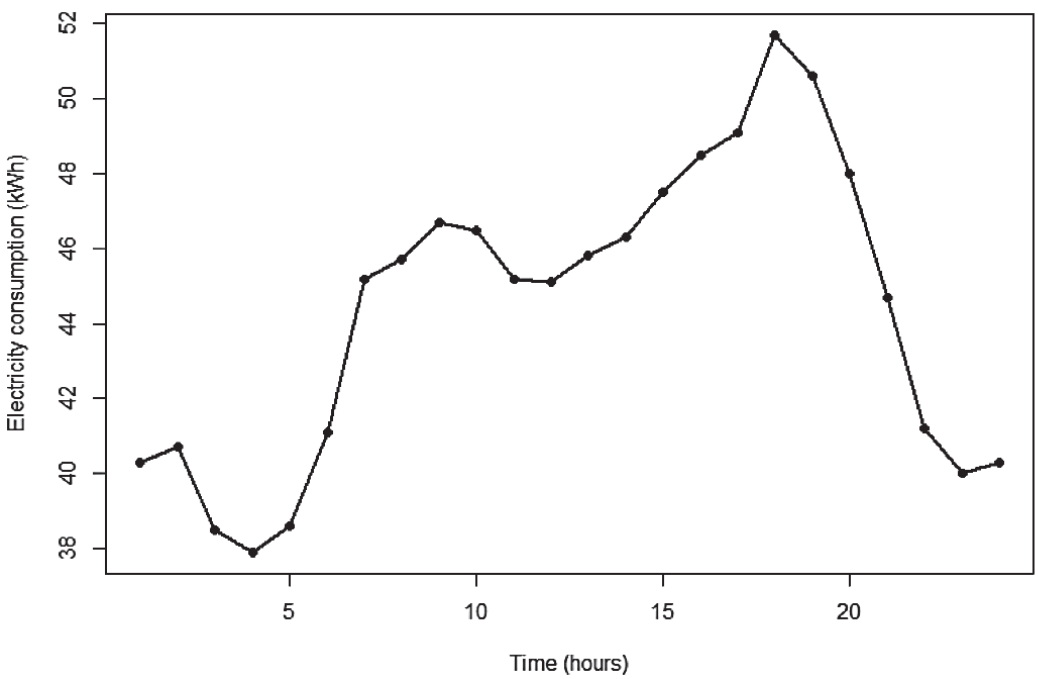
Remark 1 (Parametric space for variance parameters $\boldsymbol{\nu}$)

One practical reason for working only with the restricted space Υ^* is that any zero variance σ_j^2 implies almost sure zero random component Y_j (i.e. Y_j has a degenerate distribution with $P(Y_j = 0) = 1$), which in practice means ignoring this component in the model (1). Another research reason is to avoid technical or numerical problems dealing with zero variances in developing theory. However, there can be considerable interest not to reduce Υ, e.g. in testing (e.g. testing for overdispersion, where we would carry out testing for zero variance components and dropping them from the model) or in guaranteeing the existence of estimators and predictors (e.g. in some cases, estimates of $\boldsymbol{\nu}$ based on least-square minimization or likelihood maximization exist only in space Υ, but not in restricted space Υ^*).

1.2 The second stage – time series data

In the second stage of kriging we observe an empirical realization of finitely many values $\boldsymbol{X} \equiv (X(1), X(2), \ldots, X(n))'$ of time series $X(.)$. As a real data example, we use a microeconomic time series dataset (Figure 1) from Štulajter and Witkovský (2004).

Figure 1 Time series data of electricity consumption during 24 hours in a department store



Source: Authors' figure created in R software (R Development Core Team, 2016; real data from the table of Example 4.1, p.116, Štulajter, Witkovský, 2004; the dataset as a text file are available at: <https://goo.gl/tiJjvr>.)

This time series can be modeled by an adequate Gaussian FDSLRM, if we employ generally applicable empirical considerations commonly used in economics and business (Štulajter, 2002; Štulajter and Witkovský, 2004). First of all, economic or business data often show some periodic (seasonal) patterns as they are influenced by seasons or regularly repeating events. To identify significant frequencies describing the periodic behavior, we apply spectral time series analysis (Andersen et al., 2009; Brockwell and Davis, 2006; Priestley, 2004; Štulajter, 2002). Generally, there are more than one frequency. Lower frequencies appear in the trend, and higher frequencies are included in the random component. According to the periodogram,[5] the main tool of the spectral analysis, there are three most significant Fourier frequencies $\lambda_1 = 2\pi/24, \lambda_2 = 2\pi/8, \lambda_3 = 2\pi/6$. Considering and checking all mentioned facts in same way as in Štulajter and Witkovský (2004), we get the following FDSLRM (1) with $k = 3, l = 4$ for the explored consumption dataset:

$$X(t) = \beta_1 + \beta_2 \cos\lambda_1 t + \beta_3 \sin\lambda_1 t + Y_1 \cos\lambda_2 t + Y_2 \sin\lambda_2 t + Y_3 \cos\lambda_3 t + Y_4 \sin\lambda_3 t + w(t)\,; t \in \{1, 2, \ldots, 24\}. \quad (3)$$

1.3 The third stage – the BLUP for a chosen future value

As for the third kriging stage, finding the BLUP, in this case, is straightforward. Mathematically, model (3) represents an orthogonal version of FDSLRM (Štulajter, 2003) for which exists a closed analytic form of the BLUP (theorem 2.1 in Štulajter, 2003, p. 129) for any future value $X(n + d), d \in \mathbb{N}$. This form denoted by $X^*(n + d)$ generally depends on variance parameters $\boldsymbol{\nu}$:

$$X^*(n + d) \equiv X^*_{\boldsymbol{\nu}}(n + d).$$

1.4 The fourth stage – estimation of models parameters and use of the EBLUP

In practical situations like this one, we need to estimate regression parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)' \in \mathbb{E}^3$ and variance parameters $\boldsymbol{\nu} = (\sigma^2, \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)' \in \Upsilon^* = (0, \infty)^5$. Various standard and also nonstandard mathematical techniques for estimating $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ can be found in above mentioned references dealing with kriging and FDSLRM (Christensen, 2001; Hančová, 2008; Štulajter, 2002; Štulajter and Witkovský, 2004), but also in references dealing with the methodology of LMMs (e.g. Rao and Molina, 2015; Searle et al., 2006).

With regard to statistical properties, we remind some general results from the estimation theory. Standardly used estimators of $\boldsymbol{\beta}$ in linear regression models like least squares estimators (ordinary – OLSE or weighted – WELSE) or maximum likelihood estimators (MLE or REMLE) are linear with respect to a time series observation $\boldsymbol{X}$. For our FDSLRM (3), OLSE of the regression parameters $\boldsymbol{\beta}$ give us[6] $\widehat{\boldsymbol{\beta}} = (44.38, -3.15, -3.52)$.

In connection with estimating $\boldsymbol{\nu}$, standard least-squares methods of estimation in FDSLRM in many cases lead to quadratic estimators which are invariant quadratic forms[7] in $\boldsymbol{X}$. Variance parameters can be estimated e.g. by double ordinary least squares estimators (DOOLSE) or by their modified unbiased version (MDOOLSE) as it is described in Remark 2.

Remark 2 (DOOLSE and MDOOLSE)

The double least squares method is based on two following steps. First of all, we find OLSE $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$, then we can compute empirical residuals $\widehat{\boldsymbol{\varepsilon}} \equiv \boldsymbol{X} - \mathrm{F}\widehat{\boldsymbol{\beta}}$. Then matrix $\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}' = (\boldsymbol{X} - \mathrm{F}\widehat{\boldsymbol{\beta}})\,(\boldsymbol{X} - \mathrm{F}\widehat{\boldsymbol{\beta}})'$ represents

[5] The periodogram can be computed in the base R package e.g. by function `spec.pgram{}`.

[6] In the R environment, OLSE can be found via function `lm{}` in the base R package.

[7] It means that estimators of variances can be written as $\boldsymbol{X}'\mathrm{A}\boldsymbol{X}$, where A is some $n \times n$ real symmetric matrix and values of $\boldsymbol{X}'\mathrm{A}\boldsymbol{X}$ do not depend on $\boldsymbol{\beta}$. In FDSLRM, it is equivalent with the condition AF = 0.

the well-known estimation matrix S($\boldsymbol{X}$) for a covariance matrix Σ of $\boldsymbol{X}$ which is equal to $Cov(\boldsymbol{X}) \equiv \Sigma_v = \sigma^2 I_n + VDV'$, $D = diag\{\sigma_j^2\}$. Using ordinary least squares method a second time, i.e. minimizing the square of a norm (distance) between Σ_v and S($\boldsymbol{X}$) with respect of $\boldsymbol{v} \in \Upsilon$, we find OLSE estimates $\widehat{\boldsymbol{v}}$ which are called DOOLSE. As for calculation methods, two approaches are usually applied: multi variable calculus and geometrical projection theory, since the least squares problems can be expressed in terms of orthogonal projections.

Requiring unbiasedness, we can modify DOOLSE to unbiased estimators (MDOOLSE). The exact formal definition of these type of estimators in linear regression models can be found e.g. in Štulajter (2002, p. 25). Moreover, it can be shown (Štulajter, Witkovský, 2004) that in any orthogonal Gaussian FDSLRM as it is in our case (3), DOOLSE are identical with maximum likelihood estimators (MLE) and MDOOLSE are equal to restricted MLE (REMLE).

If we apply corresponding DOOLSE/MDOOLSE formulas based on the geometrical theory of orthogonal projectors (Štulajter and Witkovský; 2004, p. 107, last two formulas) in our FDSLRM, we get the following results:

projectors for DOOLSE $\quad \widehat{\boldsymbol{v}} = (3.00, 0.12, 1.61, -0.24, 1.02) \notin \Upsilon^*$
projectors for MDOOLSE $\quad \tilde{\boldsymbol{v}} = (3.53, 0.08, 1.57, -0.29, 0.97) \notin \Upsilon^*$

We see that in both cases projection formulas for the standard estimation methods fail. Variance components can never be negative. However, it is very important to realize that the DOOLSE/MDOOLSE estimates can be based on the projection method only if the method provides values for $\boldsymbol{v}$ belonging to the parametric space Υ^* or Υ. Therefore in our case we had to use other methods of computation, e.g. numerical iterative methods (Štulajter, 2002), which indicate that DOOLSE give us an estimate of $\boldsymbol{v}$ with zero component $\tilde{\sigma}_3^2$ lying on the boundary of Υ or no estimate of $\boldsymbol{v}$, if we consider restricted space Υ^*. What action should be done in this situation?

In the framework of LMM, such estimation problem with negative or zero values of estimates for variance components has a long and rich history (Searle et al., 2006). It is a well-known problem at least 40 years, especially in using ANOVA, MLE and REMLE estimators for LMMs. Inspiring by section 4.4 in Searle at al. (2006, p.130), there are several possibilities how to solve it, if we speak about FDSLRM: (i) understand it as a consequence of insufficient data and collect more time series data; (ii) accept zero estimates and ignore the zero variance components in the model, if it is reasonable; (iii) interpret negative or zero results as indication of a wrong model and build a new, but still adequate FDSLRM model for considered data; (iv) use a modified or new method of estimation leading to positive estimates. In the case of FDSLRM, only last two possibilities were already studied.

Building a new adequate FDSLRM model was done in Štulajter and Witkovský (2004). During the spectral analysis authors replaced the third most significant Fourier frequency $2\pi/6$ by the fourth one $2\pi/12$. However, simulation results of the next section will show that this approach is not fully satisfactory since it does not work in relatively frequent circumstances.

The last above-mentioned solution (iv) is to use new estimators with always positive or almost sure positive values. Such new estimators also based on least squares (Remark 3), called natural estimators (NE), were proposed and studied in Hančová (2008). Statistically these estimators are biased invariant quadratic forms.

Remark 3 (Invariant quadratic biased NE)

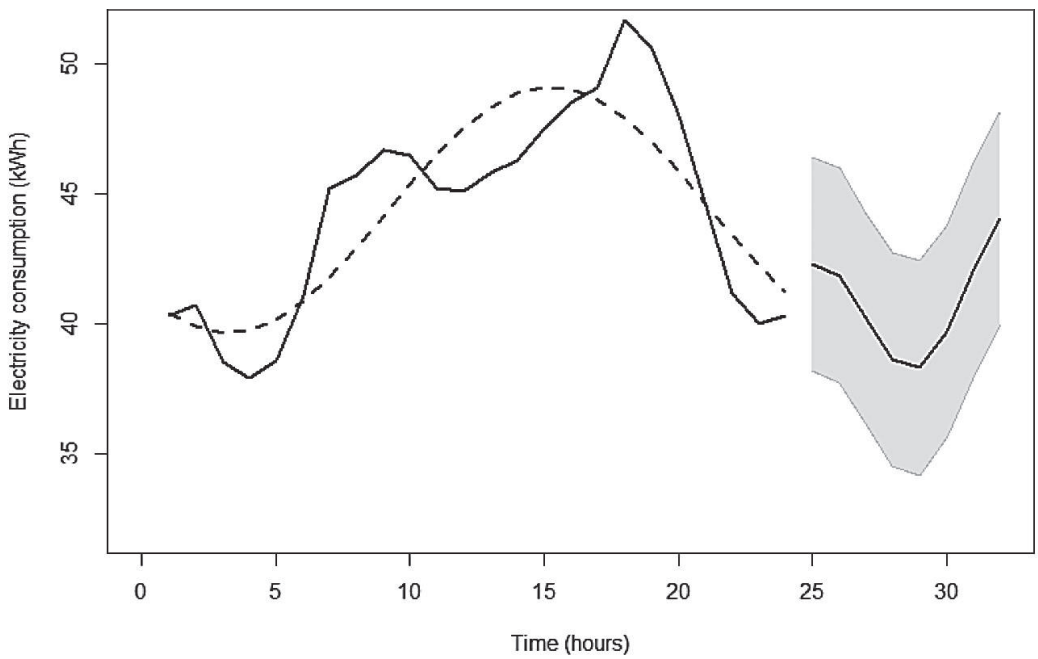
The main idea behind NE comes from the fact that $\sigma_j^2 = E(Y_j^2), j = 1, \ldots, l.$ Then, it is reasonable to estimate an l-vector of unobservable realization $\boldsymbol{y}$ of $\boldsymbol{Y}$ in model (2) by ordinary least squares and use squares of these estimates $\tilde{y}_j^2, j = 1, \ldots, l$, as estimators of σ_j^2 (see more details and the exact formal

definition in section 2.1, Hančová, 2008). Geometrically, these so-called natural estimators can be expressed by oblique projectors.

If we are interested in non-negative and simultaneously unbiased estimators, then generally there do not exist any non-negative unbiased estimators for $\sigma_j^2, j = 1, \ldots, l$. Twenty years ago, Ghosh (1996) formulated an elegant proof of two general important results connected with incompatibility between non-negativity and unbiasedness of random effects estimators in LMM: (i) in any LMM with $\boldsymbol{\nu} \in \Upsilon = (0, \infty) \times [0, \infty)^l$ and $\boldsymbol{X}$ having an absolutely continuous probability distribution with respect to some σ-finite measure, if $\nu_j^*(\boldsymbol{X})$ is an unbiased estimator of σ_j^2, then there is always non-zero probability for $\nu_j^*(\boldsymbol{X})$ to be negative with respect to some $\boldsymbol{\beta}$, $\boldsymbol{\nu}$; (ii) the same is true, $P\{\nu_j^*(\boldsymbol{X}) < 0 | \boldsymbol{\beta}, \boldsymbol{\nu}\} > 0$ for some $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$, if we suppose $\boldsymbol{\nu} \in \Upsilon^* = (0, \infty)^{l+1}$ and $\boldsymbol{X}$ having a probability density function continuous in all $\boldsymbol{\nu} \in \Upsilon^*$.

Computing NE numerically (Hančová, 2008, p. 268, formulas 2.2, 2.3), we get $\tilde{\boldsymbol{\nu}} = (3.53, 0.37, 1.86, 0.004, 1.27) \in \Upsilon$. In practice these estimates are suitable for computation of BLUPs (Štulajter, 2003 the first formula on p. 129) for future values $X(n + d), d \in \mathbb{N}$. These „plug-in" BLUPs are called empirical BLUPs (EBLUPs). At the same time NE estimates can be used for computation of corresponding „plug-in" MSEs (Štulajter, 2003, the second formula on p. 129) and 95% prediction intervals,[8] which are commonly used in displaying the uncertainty in time series forecasting (Hyndman and Athanasopoulos, 2014). Figure 2 is a summary graphical representation of obtained predictions for electricity consumption during the next eight hours.

Figure 2 Kriging forecasting of the electricity consumption for next 8 hours with 95% forecast intervals (solid line in the gray shaded region) and the time series trend (dashed line)



Source: Authors' figure based on their calculations, created in R software (R Development Core Team, 2016)

[8] A formula for $\mathcal{I}_{0,95}(n + d) = [X^*(n + d) - 1{,}96\sqrt{MSE\{X^*(n + d)\}}; X^*(n + d) + 1{,}96\sqrt{MSE\{X^*(n + d)\}}]$.

The plug-in step replacing true value of parameters (for which the BLUP was derived) by NE causes a certain deviation in the mean squared error. Therefore, the main task of the last fifth stage of kriging is to study statistical properties of EBLUPs based on NE which are invariant quadratic biased estimators. However, such research has not yet been made. So, the third section of the paper (after the simulation study) will be devoted to our analysis of the last stage of kriging using NE in the broader context of theoretical developments in kriging methodology.

2 SIMULATION STUDY

To answer our research question, how efficiently building a new FDSLRM can solve problems with negative values of projectors for standard variance estimates (DOOLSE, MDOOLSE), we planned four possible FDSLRM simulation designs whose structure can be based on three significant frequencies chosen by Štulajter and Witkovský (2004): $\lambda_1 = 2\pi/24, \lambda_2 = 2\pi/8, \lambda_3 = 2\pi/12$. The considered designs differ with respect to possible number $m \in \{0,1,2,3\}$ of given frequencies included in the FDSLRM trend (remaining 3-m frequencies are in the random component – shortly RC). Due to easier, more compatible notation with spectral analysis, we wrote their forms by the following compact formula ($m \in \{0,1,2,3\}$):

$$X_m(t) = \alpha + \sum_{i=1}^{m} (\beta_i \cos \lambda_i t + \gamma_i \sin \lambda_i t) + \sum_{j=1}^{3-m} \left(Y_j \cos \lambda_j t + Z_j \sin \lambda_j t\right) + w(t), \tag{4}$$

for $m = 1$ we get the identical model with the original one applied by Štulajter and Witkovský (2004). OLSE for regression parameters $\alpha, \beta_i, \gamma_j$ and NE for variance parameters calculated from the real dataset (Figure 1) were assigned as true parameters for simulation designs (Table 1). We also mention that in this case, NE values are evidently nonzero and they are also close to DOOLSE and MDOOLSE for the dataset.

Table 1 Vectors of regression and variance parameters for considered model designs

Model design	Regression parameters $\boldsymbol{\beta}$	Variance parameters $\boldsymbol{v}$
$m = 0$ (3 frequencies in RC)	(44.38)′	(1.09, 9.93, 12.43, 2.97, 1.76, 0.37, 1.86)′
$m = 1$ (2 frequencies in RC)	(44.38, –3.15, –3.52)′	(1.09, 2.97, 1.76, 0.37, 1.86)′
$m = 2$ (1 frequency in RC)	(44.38, –3.15, –3.52, –1.72, –1.33)′	(1.09, 0.37, 1.86)′
$m = 3$ (0 frequencies in RC)	(44.38, –3.15, –3.52, –1.72, –1.33, 0.61, 1.36)′	(1.09)′

Source: Authors' calculations based on real data from Štulajter, Witkovský (2004) using R (R Development Core Team, 2016)

Using R, we simulated $N = 5\,000$ time series realizations for each design (values of $Y_j, Z_j, w(t)$ were generated from normal distributions with zero means and variance parameters given by Table 1). Then for each realization (a time series dataset) estimates via corresponding orthogonal or oblique projectors (for DOOLSE, MDOOLSE and NE) were computed and simultaneously a relative occurrence of the projections with negative values was counted. Complete results dealing with a relative occurrence of negative values in the four evaluated simulation designs are reported in Table 2 (NE are not included since they really led only to positive estimates).

As for distributions, Figure 3 presents typical results in the form of histograms for projectors dealing with MDOOLSE and NE (as examples) in the case of simulation design $m = 2$. Table 2 clearly manifests that in all designs projection methods for computing estimations (DOOLSE = MLE, MDOOLSE = REMLE) give

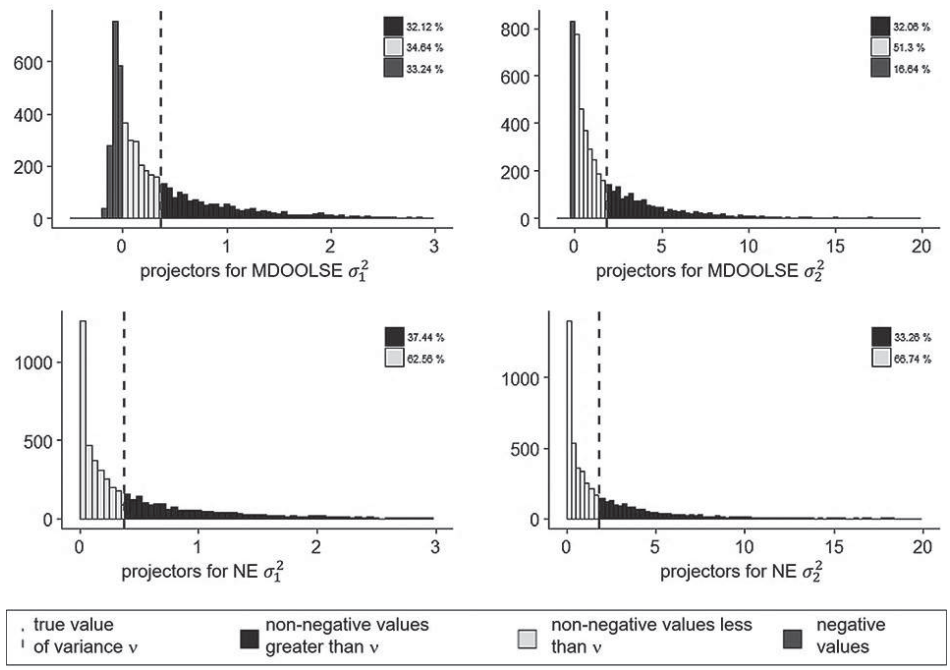
Table 2 Relative occurrence of negative values (as results of projection methods) for estimation of variance parameters $\boldsymbol{\nu}$ in $N = 5\,000$ simulated replications for each model design

Model design	Projectors for estimators	Relative occurrence of negative values for:					
$m = 0$ (3 frequencies in RC)	DOOLSE	6.48 %	6.76 %	13.44 %	17.16 %	32.68 %	16.98 %
	MDOOLSE	6.62 %	6.92 %	13.84 %	17.64 %	33.60 %	17.30 %
$m = 1$ (2 frequencies in RC)	DOOLSE	13.16 %	16.36 %	32.10 %	15.54 %	x	x
	MDOOLSE	14.14 %	17.82 %	34.56 %	16.60 %	x	x
$m = 2$ (1 frequency in RC)	DOOLSE	29.54 %	14.66 %	x	x	x	x
	MDOOLSE	33.24 %	16.64 %	x	x	x	x
$m = 3$ (0 frequencies in RC)	DOOLSE	x	x	x	x	x	x
	MDOOLSE	x	x	x	x	x	x

Source: Authors' simulation results based on parameters in Table 1 using R (R Development Core Team, 2016)

us from 6% up to 35% of wrong negative values for estimates which are not small, insignificant numbers. If we are interested in actual values of the estimators, alternative numerical methods (Štulajter 2002; also applied by us in R) show that zero values of DOOLSE = MLE and MDOOLSE = REMLE for $\boldsymbol{\nu}$ correspond to the negative values of projectors. Since $\boldsymbol{\nu}$ has only nonzero components, from a theoretical point of view, zero estimates also mean a failure. Therefore, we can conclude that changing or rebuilding the model (3) as it was done by Štulajter and Witkovský (2004) will not work in relatively frequent circumstances.

Figure 3 Typical results of the simulation study showing relative occurrence of values (as results of projection methods) for estimation of variance parameters in model design $m = 2$



Source: Authors' figure based on their calculations, created in R software (R Development Core Team, 2016)

3 KRIGING IN PRACTICE – MSE OF EMPIRICAL BLUPS

During the fourth stage of kriging, we saw that in the next stage, there is a need to study effects of estimating unknown variance parameters of FDSLRM on statistical properties of BLUPs, if these true parameters (to avoid any misunderstanding in this section we denote them $\boldsymbol{\nu}_{\mathrm{T}}$ instead of $\boldsymbol{\nu}$ are replaced by variance estimates $\check{\boldsymbol{\nu}}$ (e.g. invariant quadratic biased NE). Special attention must be paid to MSE of EBLUPs which determines not only quality of the obtained empirical predictors but allows us to express corresponding forecast intervals or to test statistical hypotheses.

However, general theory of empirical linear unbiased predictors (Harville, 2008; Štulajter, 2002; Witkovský, 2012; Żądło, 2009) says that explicit expression for the MSE of an empirical predictor (EBLUP) is not known. One of the reasons why it is still open research problem consists of the fact that EBLUP is a nonlinear function of the observation $\boldsymbol{X}$. Therefore, finding such expression is a very difficult mathematical task. On the other hand, the theory gives us an approximation for the correction (adjustment) of MSE of EBLUPs with respect to the original BLUP using Taylor's series (see e.g. Harville, 2008; or Štulajter, 2002):

$$E\left[X^*_{\check{\boldsymbol{\nu}}}(n+d) - X^*_{\boldsymbol{\nu}_T}(n+d)\right]^2 \approx E\left[\frac{\partial X^*_{\boldsymbol{\nu}}(n+d)}{\partial \boldsymbol{\nu}'}\Big|_{\boldsymbol{\nu}=\boldsymbol{\nu}_T} \cdot (\check{\boldsymbol{\nu}} - \boldsymbol{\nu}_T)\right]^2 =$$
$$= \sum_{a,b=0}^{l} E\left[\frac{\partial X^*_{\boldsymbol{\nu}}(n+d)}{\partial \nu_a}(\check{\nu}_a - \nu_a)(\check{\nu}_b - \nu_b)\frac{\partial X^*_{\boldsymbol{\nu}}(n+d)}{\partial \nu_b}\right] \tag{5}$$

Although we know explicit forms of $\check{\nu}_a, a = 0,1, \ldots, l$ for NE (Hančová, 2008), the direct use of these quadratic forms in theoretical and corresponding computational study of the approximation (5) would lead to cumbersome and uselessly complicated mathematical work. In this case, more abstract and general approach paradoxically makes the problem more tractable and understandable, stripping away non-essential features. In addition, such generalization allows us to use a new arsenal of mathematical techniques.

Therefore mathematically, it is more useful to describe NE $\check{\nu}_a$ in the approximation (5) only as invariant quadratic biased estimators of the general form $Q_{\mathrm{A}} \equiv \boldsymbol{X}'\mathrm{A}\boldsymbol{X}, \mathrm{AF} = 0$. Partial derivatives $\partial X^*_{\boldsymbol{\nu}}(n+d)/\partial \nu_a$ have the general form $\boldsymbol{c}'\boldsymbol{X} + d$, $\boldsymbol{c}' \in \mathbb{E}^n, d \in \mathbb{E}^1$. If we introduce a concept of the so-called parameter centered quadratic form $Q^*_{\mathrm{A}} \equiv Q_{\mathrm{A}} - \nu_a = \boldsymbol{X}'\mathrm{A}\boldsymbol{X} - \nu_a = \check{\nu}_a - \nu_a$, then it is easy to see that the approximation (5) depends on expressions such as $E(Q^*_{\mathrm{A}}), E(\boldsymbol{X}Q^*_{\mathrm{A}}\boldsymbol{X}'), E(Q^*_{\mathrm{A}}Q^*_{\mathrm{B}}), E(\boldsymbol{X}Q^*_{\mathrm{A}}Q^*_{\mathrm{B}}), E(\boldsymbol{X}Q^*_{\mathrm{A}}Q^*_{\mathrm{B}}\boldsymbol{X}')$.

These moments are up to sixth order with respect to $\boldsymbol{X}$. However, as our next theoretical results demonstrate, if the finite time series observation $\boldsymbol{X}$ (model (2)) comes from Gaussian FDSLRM (1) and consequently has a multivariate normal distribution $\boldsymbol{X} \sim N(\mathrm{F}\boldsymbol{\beta}, \Sigma)$ with the positive definite covariance $n \times n$ matrix Σ ($\Sigma > 0$), then all moments up to sixth order can be expressed as functions depending only on the second-order (not higher) properties of $\boldsymbol{X}$ given by mean value parameters $\boldsymbol{\beta}$ and variance parameters $\boldsymbol{\nu}$. Under the assumption of normality for $\boldsymbol{X}$, using the modern algebraic apparatus of advanced multivariate statistics (Ghazal and Neudecker, 2000; Kollo and Rosen, 2005) which includes vectorization, commutation matrices, the Kronecker product and relations among them, we derived the explicit form of mentioned expressions. Our results are summarized by the following general theorem which contains the moments for any invariant quadratic biased estimators (NE are a special case). Due to higher mathematical sophistication and technicalities, its proof is explained in the Appendix.

Theorem (the explicit form of moments)

Let a random vector $\boldsymbol{X} \sim N(\mathrm{F}\boldsymbol{\beta}, \Sigma)$ be a given finite observation of time series, where $\mathrm{F} \in \mathbb{E}^{n\times k}, \boldsymbol{\beta} \in \mathbb{E}^k$ and $\Sigma \in \mathbb{E}^{n\times n}, \Sigma > 0$. Let $Q_{\mathrm{A}} \equiv \boldsymbol{X}'\mathrm{A}\boldsymbol{X}, Q_{\mathrm{B}} \equiv \boldsymbol{X}'\mathrm{B}\boldsymbol{X}, \mathrm{A}, \mathrm{B} \in \mathbb{E}^{n\times n}$ be the invariant quadratic forms, i.e. $\mathrm{AF} = \mathrm{BF} = 0, E(Q_{\mathrm{A}}) = \mathrm{tr}(\mathrm{A}\Sigma), E(Q_{\mathrm{B}}) = \mathrm{tr}(\mathrm{B}\Sigma)$ and $Cov(Q_{\mathrm{A}}, Q_{\mathrm{B}}) = 2\mathrm{tr}(\mathrm{A}\Sigma\mathrm{B}\Sigma)$. Then for parameter-centered quadratic forms $Q^*_{\mathrm{A}} \equiv Q_{\mathrm{A}} - \nu_{\mathrm{A}}$ and $Q^*_{\mathrm{B}} \equiv Q_{\mathrm{B}} - \nu_{\mathrm{B}}$; $\nu_{\mathrm{A}}, \nu_{\mathrm{B}} \in \mathbb{E}^1$ the following properties hold:

(i) $E(Q_{\mathrm{A}}^{*}) = E(Q_{\mathrm{A}}) - \nu_{\mathrm{A}}, E(\boldsymbol{X}Q_{\mathrm{A}}^{*}) = E(Q_{\mathrm{A}}^{*})\mathrm{F}\boldsymbol{\beta}$,

(ii) $E(\boldsymbol{X}Q_{\mathrm{A}}^{*}\boldsymbol{X}') = E(Q_{\mathrm{A}}^{*})[\Sigma + \mathrm{F}\boldsymbol{\beta}(\mathrm{F}\boldsymbol{\beta})'] + 2\Sigma\mathrm{A}\Sigma$,

(iii) $E(Q_{\mathrm{A}}^{*}Q_{\mathrm{B}}^{*}) = Cov(Q_{\mathrm{A}}, Q_{\mathrm{B}}) + E(Q_{\mathrm{A}}^{*})E(Q_{\mathrm{B}}^{*})$,

(iv) $E(\boldsymbol{X}Q_{\mathrm{A}}^{*}Q_{\mathrm{B}}^{*}) = E(Q_{\mathrm{A}}^{*}Q_{\mathrm{B}}^{*})\mathrm{F}\boldsymbol{\beta}$,

(v) $E(\boldsymbol{X}Q_{\mathrm{A}}^{*}Q_{\mathrm{B}}^{*}\boldsymbol{X}') = 2\Sigma[E(Q_{\mathrm{A}}^{*})\mathrm{B} + E(Q_{\mathrm{B}}^{*})\mathrm{A} + 2\mathrm{A}\Sigma\mathrm{B} + 2\mathrm{B}\Sigma\mathrm{A}]\Sigma +$
$+ E(Q_{A}^{*}Q_{B}^{*})[\Sigma + \mathrm{F}\boldsymbol{\beta}(\mathrm{F}\boldsymbol{\beta})']$.

CONCLUSIONS AND FURTHER DEVELOPMENTS

One of the most important areas of time series theory application is forecasting time series providing a crucial knowledge for effective and efficient planning or decision making. In the paper, we have presented a general framework of forecasting methodology for econometric time series, called kriging. Our kriging application deals with a recently introduced family of linear regression time series models named FDSLRM, which apply regression not only to a trend, but also to a random component of the observed time series.

Using a real data example dealing with electricity consumption, we have also investigated one of the current research problems of kriging – a problem of negative or zero estimates which leads to kriging failures in empirical prediction. Performing a simulation study, we manifested that this problem occurs in relatively frequent circumstances and therefore cannot be neglected. Simultaneously we pointed out inadequacy of rebuilding the model as used problem solution. If computational methods using a dataset of time series observation give failing negative or zero values for standard estimates, then we can apply one of possible solutions – using alternative estimators like natural estimators (NE) which are invariant quadratic biased estimators.

Our consequent analysis in the broader context of kriging methodology developments allowed us to derive explicitly moments of a finite Gaussian time series observation for any invariant quadratic biased estimators of time series variances. Confronting with other research, we have found that our theoretical results were a direct extension of the results of the previous research (Prasad and Rao, 1990; Srivastava and Tiwari, 1976). In comparison with these references, our use of the matrix approach of advanced modern multivariate statistics in proving our results seems more elegant and conceptually simpler than the original cumbersome multiple use of sums with many indices.

As for further research and kriging developments, these moments will allow a theoretical study of properties of empirical predictors and corresponding approximations of MSE based on any invariant quadratic biased estimators (e.g. according to Harville, 2008; Štulajter, 2007). Since our results are written in the recurrent matrix form, they are also very suitable for checking or conducting an effective computational research (statistical computing environments like R are essentially matrix algebra processors) with real empirical data using simulations or bootstrap methods for time series and kriging (Kreiss and Lahiri, 2012; Schelin and Sjöstedt-de Luna, 2010; Sjöstedt-de Luna and Young, 2003). Such computational research could also be applied to study effects of MLE and REMLE, in general FDSLRM not expressible in a closed analytic form, on statistical properties of BLUPs and their MSEs.

Our last conclusion deals with a corresponding implementation of FDSLRM in R. Although any finite FDSLRM observation satisfies a linear mixed model (LMM), according to our inspection it seems that no current package in R for LMM methodology[9] is directly suitable for FDSLRM. Therefore, one

[9] There are many packages in R fitting various forms of LMM, e.g. amer, gamm, glmmAK, lme4.0, lmec, lmm, MASS, MCMCglmm, nlme or PSM (more details in Galecki and Burzykowski, 2013).

of the tasks of future computational FDSLRM research is to create a fully functioning R package using the current object-oriented programming. We also assume that the O-O programming approach which is now standard in the context of statistical modeling (Galecki and Burzykowski, 2013) allows us to use some classes of objects and methods operating on them from existing R packages for LMM.

ACKNOWLEDGMENTS

Our research has been supported by the Scientific Grant Agency of the Slovak Republic (VEGA) – grants VEGA 1/0073/15, VEGA 1/0344/14 and the Internal Research Grant System of Faculty of Science, P. J. Šafárik University in Košice (VVGS PF UPJŠ) – project VVGS-PF-2016-72616.

References

ANDERSEN, T. G., DAVIS, R. A., KREISS, J.-P., MIKOSCH, T. V. eds. *Handbook of Financial Time Series.* Berlin: Springer, 2009.

BOX, G. E. P., JENKINS, G. M., REINSEL, G. C. *Time Series Analysis: Forecasting and Control.* 4th Ed. Hoboken, NJ: Wiley, 2008.

BROCKWELL, P. J. AND DAVIS, R. A. *Time Series: Theory and Methods.* 2nd Ed. New York: Springer-Verlag, 2006.

CHAMBERS, J. M. *Software for Data Analysis: Programming with R.* 1st Ed. New York: Springer, 2008.

CHATTERJEE, S. AND HADI, A. S. *Regression Analysis by Example.* 5th Ed. Hoboken, NJ: John Wiley & Sons, 2012.

CHRISTENSEN, R. *Advanced Linear Modeling: Multivariate, Time Series, and Spatial Data; Nonparametric Regression and Response Surface Maximization.* 2nd Ed. New York: Springer, 2001.

CIPRA, T. *Finanční ekonometrie.* 2nd Ed. Prague: Ekopress, 2013.

CRESSIE, N. *Statistics for Spatial Data.* Rev. Ed. New York: Wiley-Interscience, 1993.

CRESSIE, N. AND WIKLE, C. K. *Statistics for Spatio-Temporal Data.* 1st Ed. Hoboken, N.J: Wiley, 2011.

CRONE, S. F., HIBON, M., NIKOLOPOULOS, K. Advances in Forecasting with Neural Networks? Empirical Evidence from the NN3 Competition on Time Series Prediction. *International Journal of Forecasting,* 2011, 27(3), pp. 635–660.

DEMIDENKO, E. *Mixed Models: Theory and Applications with R.* 2nd Ed. Hoboken, NJ: Wiley, 2013.

ENDERS, W. *Applied Econometric Time Series.* 4th Ed. Hoboken, NJ: Wiley, 2014.

ESCOBARI, D. AND NGO, T. Preface: Special Issue on Time Series Econometric Applications in Finance. *American Journal of Economics,* 2014, 4(2A), pp. 0–0.

FOMBY, T. B., TERRELL, D. eds. *Econometric Analysis of Financial and Economic Time Series Part B.* 1st Ed. Book series: Advances in Econometrics, Vol. 20, Oxford: Emerald Group Publishing Limited, 2006.

GALECKI, A. AND BURZYKOWSKI, T. *Linear Mixed-Effects Models Using R: A Step-by-Step Approach*, 2013 Ed. New York, NY: Springer, 2013.

GHAZAL, G. A. AND NEUDECKER, H. On Second-Order and Fourth-Order Moments of Jointly Distributed Random Matrices: A Survey. *Linear Algebra and Its Applications*, 2000, 321(1), pp. 61–93.

GHOSH, M. On the Nonexistence of Nonnegative Unbiased Estimators of Variance Components. *Sankhyā: The Indian Journal of Statistics, Series B (1960–2002)*, 1996, 58(3), pp. 360–362.

HANČOVÁ, M. Empirical Predictors in Finite Discrete Spectrum Linear Regression Models. In: HARMAN et al., eds. *PROBASTAT 2011 – Abstracts*, Bratislava, Slovak Republic: Institute of Measurement Science, Slovak Academy of Science, 2011, pp. 24–25.

HANČOVÁ, M. Natural Estimation of Variances in a General Finite Discrete Spectrum Linear Regression Model. *Metrika*, 2008, 67(3), pp. 265–276.

HANČOVÁ, M., HANČ, J., GAJDOŠ, J. A Simulation Study of Bootstrap Methods for Kriging in Time Series Forecasting. In: WITKOVSKÝ et al., eds. *PROBASTAT 2015 – Abstracts*, Bratislava, Slovak Republic: Institute of Measurement Science, Slovak Academy of Science, 2015, pp. 26–27.

HARMAN, R. AND ŠTULAJTER, F. Optimal Prediction Designs in Finite Discrete Spectrum Linear Regression Models. *Metrika*, 2010, 72(2), pp. 281–294.

HARVILLE, D. A. Accounting for the Estimation of Variances and Covariances in Prediction under a General Linear Model: An Overview. *Tatra Mountains Mathematical Publications*, 2008, 39(1), pp. 1–15.

HYNDMAN, R. J. AND ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. Print Ed. Melbourne, Australia: OTexts, 2014.

KOLLO, T. AND ROSEN, D. VON. *Advanced Multivariate Statistics with Matrices*. Berlin: Springer, 2005.

KREISS, J.-P. AND LAHIRI, S. N. Bootstrap Methods for Time Series. In: RAO, RAO, RAO, eds. *Handbook of Statistics, Vol. 30: Time Series Analysis: Methods and Applications*, Amsterdam: Elsevier, 2012, pp. 3–26.

MCCULLOCH, C. E., SEARLE, S. R., NEUHAUS, J. M. *Generalized, Linear, and Mixed Models*. 2nd Ed. Hoboken, N.J: Wiley-Interscience, 2008.

MCLEOD, A. I., YU, H., MAHDI, E. Time Series with R. In: RAO, RAO, RAO, eds. *Handbook of Statistics, Vol. 30: Time Series Analysis: Methods and Applications*, Amsterdam: Elsevier, 2012, pp. 661–712.
MOORE, M. eds. *Spatial Statistics: Methodological Aspects and Applications.* New York: Springer Science & Business Media, 2001.
MURPHY, K. P. *Machine Learning: A Probabilistic Perspective.* 1st Ed. Cambridge, MA: The MIT Press, 2012.
PANKRATZ, A. *Forecasting with Dynamic Regression Models.* 1st Ed. New York: John Wiley & Sons, 1991.
POŠTA, V. AND PIKHART, Z. The Use of the Sentiment Economic Indicator for GDP Forecasting: Evidence from EU Economies [online]. *Statistika: Statistics and Economy Journal*, 2012, 92(1), pp. 41–55.
PRASAD, N. G. N. AND RAO, J. N. K. The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American Statistical Association*, 1990, 85(409), pp. 163–171.
PRIESTLEY, M. B. *Spectral Analysis and Time Series.* Amsterdam: Elsevier Acad. Press, 2004.
R DEVELOPMENT CORE TEAM. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing, 2016.
RAO, J. N. K. AND MOLINA, I. *Small Area Estimation*. 2nd Ed. Hoboken, New Jersey: Wiley, 2015.
ROBINSON, G. K. That BLUP Is a Good Thing: The Estimation of Random Effects. *Statist. Sci.,* 1991, 6(1), pp. 15–32.
SALAMAGA, M. Testing the Effectiveness of Some Macroeconomic Variables in Stimulating Foreign Trade in the Czech Republic, Hungary, Poland and Slovakia [online]. *Statistika: Statistics and Economy Journal,* 2015, 95(1), pp. 47–59.
SCHELIN, L. AND SJÖSTEDT-DE LUNA, S. Kriging Prediction Intervals Based on Semiparametric Bootstrap. *Mathematical Geosciences,* 2010, 42(8), pp. 985–1000.
SEARLE, S. R., CASELLA, G., MCCULLOCH, C. E. Variance Components. John Wiley & Sons, 2006.
SHUMWAY, R. H. AND STOFFER, D. S. *Time Series Analysis and Its Applications: With R Examples.* 3rd Ed. New York: Springer, 2011.
ŠIMPACH, O. Fertility of Czech Females Could Be Lower than Expected: Trends in Future Development of Age-Specific Fertility Rates up to the Year 2050 [online]. *Statistika: Statistics and Economy Journal,* 2015, 95(1), pp. 19–37.
SJÖSTEDT-DE LUNA, S. AND YOUNG, A. The Bootstrap and Kriging Prediction Intervals. *Scandinavian Journal of Statistics,* 2003, 30(1), pp. 175–192.
SRIVASTAVA, V. K. AND TIWARI, R. Evaluation of Expectations of Products of Stochastic Matrices. *Scandinavian Journal of Statistics,* 1976, 3(3), pp. 135–138.
STEIN, M. L. *Interpolation of Spatial Data: Some Theory for Kriging.* New York: Springer, 1999.
ŠTULAJTER, F. Mean Squared Error of the Empirical Best Linear Unbiased Predictor in an Orthogonal Finite Discrete Spectrum Linear Regression Model. *Metrika,* 2007, 65(3), pp. 331–348.
ŠTULAJTER, F. *Predictions in Time Series Using Regression Models.* New York: Springer, 2002.
ŠTULAJTER, F. The MSE of the BLUP in a Finite Discrete Spectrum LRM. *Tatra Mountains Mathematical Publications,* 2003, 26(1), pp. 125–131.
ŠTULAJTER, F. AND WITKOVSKÝ, V. Estimation of Variances in Orthogonal Finite Discrete Spectrum Linear Regression Models. *Metrika,* 2004, 60(2), pp. 105–118.
TSAY, R. S. *Analysis of Financial Time Series.* 3rd Ed. Cambridge, Mass.: Wiley, 2010.
VERZANI, J. *Getting Started with RStudio.* Sebastopol, Calif.: O'Reilly, 2011.
WITKOVSKÝ, V. Estimation, Testing, and Prediction Regions of the Fixed and Random Effects by Solving the Henderson's Mixed Model Equations. *Measurement Science Review,* 2012, 12(6), pp. 234–248.
ŻĄDŁO, T. On MSE of EBLUP. *Statistical Papers*, 2009, 50(1), pp. 101–118.

APPENDIX: PROOF

Since used arguments are very similar in proofs of all items (i)–(v), we explain ideas of the proof only for the first two items (i), (ii). We achieve the first simplification, when we concentrate on deriving moments $E(Q_A), E(\boldsymbol{X}Q_A), E(\boldsymbol{X}Q_A\boldsymbol{X}')$. The second, essential simplification of the proof arises from introducing the residual vector $\boldsymbol{\varepsilon} \equiv \boldsymbol{X} - E(\boldsymbol{X}) \sim N(0, \Sigma)$, where $\Sigma = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = E(\boldsymbol{X}\boldsymbol{X}') - \mathrm{F}\boldsymbol{\beta}(\mathrm{F}\boldsymbol{\beta})'$, using linearity of mean value $E(.)$, invariance of Q_A and rewriting considered moments as functions of $\boldsymbol{\varepsilon}$:

$$E(Q_A) = E(\boldsymbol{\varepsilon}'A\boldsymbol{\varepsilon}), E(\boldsymbol{X}Q_A) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'A\boldsymbol{\varepsilon}) + \mathrm{F}\boldsymbol{\beta}E(\boldsymbol{\varepsilon}'A\boldsymbol{\varepsilon})$$

$$E(\boldsymbol{X}Q_A\boldsymbol{X}') = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'A\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') + \mathrm{F}\boldsymbol{\beta}E(\boldsymbol{\varepsilon}'A\boldsymbol{\varepsilon})(\mathrm{F}\boldsymbol{\beta})'.$$

Assumptions of the theorem about Q_A give us immediately $E(\boldsymbol{\varepsilon}'A\boldsymbol{\varepsilon}) = E(Q_A) = \mathrm{tr}\,(A\Sigma)$.

At this moment, we recall needed expressions and properties of multivariate statistics apparatus (Ghazal, Neudecker, 2000; Kollo, Rosen, 2005):

a) vec $\boldsymbol{a}'$ = vec $\boldsymbol{a}$ = $\boldsymbol{a}$ for any column vector $\boldsymbol{a}$ where vec is defined as follows:
let A be a $m \times n$ matrix and $A_{.j}$ the jth column of A; then vec A is the mn-column

vector vec $A = \begin{pmatrix} A_{.1} \\ \vdots \\ A_{.n} \end{pmatrix}$,

b) vec $\boldsymbol{ab}' = \boldsymbol{b} \otimes \boldsymbol{a}$ for any pair of column vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ where $\otimes$ is the Kronecker product (also known as the direct or tensor product) defined in general for arbitrary $k \times l$ matrix A with elements A_{ij} and $m \times n$ matrix B by the formula:

$$A \otimes B = \begin{pmatrix} A_{11}B & \cdots & A_{1l}B \\ \vdots & \ddots & \vdots \\ A_{k1}B & \cdots & A_{kl}B \end{pmatrix},$$

c) vec (ABC) = (C′ ⊗ A)vec B = vec[(C′ ⊗ A)vec B] = (vec B ⊗ Imp)′vec(C′ ⊗ A) for compatible matrices A,B and C, where mp is the row order of C′ ⊗ A,
d) tr (A′B) = (vec A)′vec B for compatible matrices A and B,
e) $K_{mn} := \Sigma_{ij}(E_{ij} \otimes E'_{ij})$, $i \in \{1,2,\ldots, m\}$, $j \in \{1,2,\ldots, n\}$ is called commutation matrix, where Σ_{ij} is a double summation symbol, E_{ij} is a $m \times n$ matrix with a unity in its i,j-th position and zeroes elsewhere,
f) K_{mn} vec A = vec A′,
g) (A ⊗ B)(C ⊗ D) = AC ⊗ BD for compatible matrices A,B,C and D,
h) $K'_{mn} = K^{-1}_{mn} = K_{nm}$,
i) $K_{m1} = K_{1m} = I_m$,
j) vec (A ⊗ B) = $(I_n \otimes K_{qm} \otimes I_p)$(vec A ⊗ vec B) for $m \times n$ matrix A and $p \times q$ matrix B,
k) $K_{rs,m} = (I_r \otimes K_{sm})(K_{rm} \otimes I_s)$.

Now employing property c) we easily conclude about term $E(\boldsymbol{\varepsilon\varepsilon}'A\boldsymbol{\varepsilon})$ that:

$E(\boldsymbol{\varepsilon\varepsilon}'A\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}$ vec $\boldsymbol{\varepsilon}'A\boldsymbol{\varepsilon}) = E[\boldsymbol{\varepsilon}(\boldsymbol{\varepsilon}' \otimes \boldsymbol{\varepsilon}')$vec A] = $E[\boldsymbol{\varepsilon}(\boldsymbol{\varepsilon}' \otimes \boldsymbol{\varepsilon}')]$vec A .

An expression for $E[\boldsymbol{\varepsilon}(\boldsymbol{\varepsilon}' \otimes \boldsymbol{\varepsilon}')]$ (corollary 2.2.7.2 (ii) in Kollo, Rosen, 2005, p. 204) and $E(\boldsymbol{\varepsilon}) = 0$ definitively lead to:

$E(\boldsymbol{\varepsilon\varepsilon}'A\boldsymbol{\varepsilon}) = 0$.

The most sophisticated part of the proof is the calculation of $E(\boldsymbol{\varepsilon\varepsilon}'A\boldsymbol{\varepsilon\varepsilon}')$. Using properties c) and j), we can write:

vec $\boldsymbol{\varepsilon\varepsilon}'A\boldsymbol{\varepsilon\varepsilon}'$ = $(\text{vec}'A \otimes I_{n^2})(I_n \otimes K_{nn} \otimes I_n)(\boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon})$.

Taking the mean value of $\boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon}$, using c) and the expression for $E[\boldsymbol{\varepsilon}(\boldsymbol{\varepsilon}' \otimes \boldsymbol{\varepsilon}' \otimes \boldsymbol{\varepsilon}')]$ (Corollary 2.2.7.2 (iii) in Kollo, Rosen, 2005, p. 204) we find that:

$E(\boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon} \otimes \boldsymbol{\varepsilon}) = \text{vec}[\Sigma \otimes \text{vec}'\Sigma + (\text{vec}'\Sigma \otimes \Sigma)(I_{n^3} + I_n \otimes K_{nn})]$.

Then three expressions from the last equation need to be treated separately. Using appropriate relations from a)–k) and results from preceding steps, it is possible to show that the following equalities hold:

$(\mathrm{vec}'\ \mathrm{A} \otimes \mathrm{I}_{n^2})(\mathrm{I}_n \otimes \mathrm{K}_{nn} \otimes \mathrm{I}_n)\mathrm{vec}(\Sigma \otimes \mathrm{vec}'\Sigma) = \mathrm{vec}\ \Sigma\mathrm{A}\Sigma,$
$(\mathrm{vec}'\ \mathrm{A} \otimes \mathrm{I}_{n^2})(\mathrm{I}_n \otimes \mathrm{K}_{nn} \otimes \mathrm{I}_n)\mathrm{vec}(\mathrm{vec}'\ \Sigma \otimes \Sigma) = \mathrm{vec}\ \Sigma\mathrm{A}\Sigma,$
$(\mathrm{vec}'\ \mathrm{A} \otimes \mathrm{I}_{n^2})(\mathrm{I}_n \otimes \mathrm{K}_{nn} \otimes \mathrm{I}_n)\mathrm{vec}[(\mathrm{vec}'\ \Sigma \otimes \Sigma)(\mathrm{I}_n \otimes \mathrm{K}_{nn})] = \mathrm{tr}(\mathrm{A}\Sigma)\mathrm{vec}\ \Sigma.$

All partial results together directly provide the final form for $E(\boldsymbol{\varepsilon\varepsilon}'\mathrm{A}\boldsymbol{\varepsilon\varepsilon}')$:

$$E(\boldsymbol{\varepsilon\varepsilon}'\mathrm{A}\boldsymbol{\varepsilon\varepsilon}') = 2\Sigma\mathrm{A}\Sigma + \mathrm{tr}(\mathrm{A}\Sigma)\Sigma.$$

Combining obtained results for moments $E(Q_\mathrm{A}), E(\boldsymbol{X}Q_\mathrm{A}), E(\boldsymbol{X}Q_\mathrm{A}\boldsymbol{X}')$ with $Q_\mathrm{A}^* = Q_\mathrm{A} - \nu_\mathrm{A}$, we finally get required moments in (i), (ii).

Cluster Analysis of World's Airports on the Basis of Number of Passengers Handled (Case Study Examining the Impact of Significant Events)

Žambochová Marta [1] | *J. E. Purkyně University, Ústí nad Labem, Czech Republic*

Abstract

Nowadays, the air transportation is one of key means of transport. Unfortunately, there are many factors influencing its popularity and intensity of its use. There are many studies investigating these factors. The paper investigates the possibility of classifying the world's airports in terms of the trend in the number of handled passengers as it is one of the main economic indicators for airports. For this classification we chose cluster analysis. The paper focuses on an important aspect of the process, which chooses the appropriate number of clusters. It turned out that in terms of interpretation of the results, it may not always be the most efficient to set this number at the theoretically best and recommended value. As a result of our analysis, several groups of airports with similar trend of post-event reactions are found. Therefore, this may bring better understanding of the intensity and the range of the impact of particular events on air transportation.[2]

Keywords

Cluster analysis, number of clusters, occupancy of airports, Bayesian Information Criterion, Akaike Information Criterion, silhouette coefficient

JEL code

C38

INTRODUCTION

Nowadays, air transport is the fastest growing transport sectors. In order to operate successfully, it is necessary to care not only for its means of transport, i.e. aircraft, but also for the facilities and background – airports and airfields. Assuming an airport to be a normal economic entity, its success is evaluated

[1] Faculty of Social and Economic Studies, Jan Evangelista Purkyne University in Usti nad Labem, Pasteurova 3544/1, 400 96 Ústí nad Labem, Czech Republic. E-mail: marta.zambochova@ujep.cz.

[2] This article is based on contribution from the conference *Robust 2016*.

according to operational and economic indicators. The basic indicators include performance indicators such as the number of aircraft movements, the number of tons of cargo handled, the number of passengers handled, etc. In this paper we deal with the last of these factors – the number of passengers.

The paper (Akamai et al., 2015) focuses on the importance of the amount of passengers for the operation of airports. The paper (Lu et al., 2014) deals with changes in traveller's behaviour during extreme weather conditions such as strong wind. Stability of air traffic at selected airports within a particular time period is reviewed in paper (Grenčíková et al., 2011). In the paper of ours, the stability of air traffic is examined globally. At the same time, the paper searches for factors influencing the possible instability at a certain moment.

This paper focuses on facts influencing the the number of passengers handled at particular airports around the world. The main task of the analysis is a data classification using cluster analysis. Several authors dealt with cluster analysis in the field of aviation before. In paper (Kraft, 2012) authors use cluster analysis to examine the key factors affecting the transport important for settlement of the Czech Republic. The paper is focused on road transport. Similarly, in the paper (Grabbe et al., 2014) cluster analysis is performed when the input variables are particular weather data at given times. Based on this analysis, the authors focus on the impact of weather on air traffic delays. However, in our contribution we used cluster analysis differently. Our main goal is to show the way enabling to find analytically a group of world airports which exhibit the same trend in the number of passengers handled. Based on this or a similar analysis, it would be possible to understand better effects which influence air transport.

1 METHOD OF ANALYSIS

Cluster analysis is based on finding similarities of data objects. It divides sets of objects into several previously unspecified groups (clusters) so that objects within an individual cluster are the most similar and objects from different clusters are the least similar.

Statistical software systems typically include, among other things, both the hierarchical algorithm with the possibility of the result shown in the form of so-called dendrogram, and non-hierarchical iterative *k*-means algorithm. The SPSS statistical system has included the TwoStep method since the version 11.5.

1.1 *K*-means method

The *k*-means method and its variants belong among the most important representatives of *k*-centroid algorithms, which form an important subset of divisive methods. This method is a very popular and widely used iterative clustering process which is suitable for analysis of quantitative data. The principal idea of the algorithm is to divide objects into a predetermined number of clusters so that the sum of distances of component objects from the centre of the cluster is minimal. In other words, its objective is to find minimum of the function:

$$Q = \sum_{\mathbf{x}\in\mathbf{X}} \left\|\mathbf{x} - c(\mathbf{x})\right\|^2 = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{il} \sum_{j=1}^{d} \left(x_{ij} - c_{lj}\right)^2, \tag{1}$$

where $\mathbf{X}$ represents the set of all analysed objects, n represents the number of objects, d represents the number of dimensions, k is the number of clusters, $\mathbf{x}$ represents an object with coordinates x_{ij}, $c(\mathbf{x})$ is the nearest centroid of the object $\mathbf{x}$, w_{il} is indicator of belonging i-th object to the l-th cluster, c_{lj} is j-th coordinate of the centroid of l-th cluster. Many variations of the basic *k*means procedure are described in literature under different names.

The algorithm is composed of the following steps:

- Step 0: An initial partition of the data file into k clusters,
- Step 1: The counting of every cluster's centroid,

- Step 2: The assignment of every object to the closest centroid,
- Step 3: Repeating Steps 1 and 2 until the centroids no longer change.

1.2 TwoStep method

This method uses the BIRCH algorithm (Balanced Iterative Reducing and Clustering using Hierarchies) which is explained in detail in (Zhang et al., 1996), or (Zhang et al., 1997). The algorithm creates first a so-called CF-tree, which is progressively filled by incoming data. The advantage of this principle is that it goes through the data file only once. The disadvantage is the sensitivity to the entry data ordering.

The TwoStep clustering procedure consists of the following steps:

- Step 1: Pre-clustering,
- Step 2: Outlier handling (optional),
- Step 3: Clustering.

In the first phase the CF-tree is created and the entering objects are progressively organized. An entry in the leaf node represents a sub-cluster. The non-leaf nodes and their entries are used for entering a new object quickly into a correct leaf node. Each entry is characterized by its CF that consists of the entry's number of objects, mean and variance of all data points belonging to the node. In the second phase the CF-tree is condensed and optimized due to its threshold adjustment. The outliers are eliminated with the help of the proper tree re-designing. In the third phase the impact of entry data order sensitivity is minimized. The leaf nodes of the CF tree are then grouped using an agglomerative hierarchical clustering algorithm. The cluster step takes sub-clusters resulting from the pre-cluster step as input and then groups them into the desired number of clusters.

1.3 Criteria for determining the optimal number of clusters

There are many information criteria for determining the optimal number of clusters (Řezanková et al., 2009). Among them, three information criteria are implemented in the SPSS. The first is the Bayesian Information Criterion, (*BIC),* which is used to determine the optimal number of clusters in the TwoStep cluster analysis. For our purpose it is calculated by the following formula:

$$BIC(k) = -2\sum_{i=1}^{k} \lambda_i + w_k \ln(n)\,, \tag{2}$$

where λ_i is the characteristic for the i-th cluster determined by the formula:

$$\lambda_i = -n_i \sum_{j=1}^{m_1} \frac{1}{2} \ln\left(s_j^2 + s_{ij}^2\right) + \sum_{j=1}^{m_2} H_{ij}\,, \tag{3}$$

n_i is the number of objects in the i-th cluster, m_1 is a number of quantitative continuous variables, m_2 is the number of categorical variables, s_j^2 is the sample variance of the j-th continuous variable, s_{ij}^2 is the sample variance of the j-th continuous variable in the i-th cluster. H_{ij} is the entropy given by the relation:

$$H_{ij} = -\sum_{l=1}^{p_j} \frac{n_{ijl}}{n_i} \ln\left(\frac{n_{ijl}}{n_i}\right), \tag{4}$$

p_j is the number of categories of the j-th categorical variables and n_{ijl} is the frequency of the l-th category of the j-th categorical variables in the i-th cluster. Furthermore, w_k is calculated according to the formula:

$$w_k = k\left(2m_1 + \sum_{j=1}^{m_2} p_j - 1\right)\cdot \tag{5}$$

When determining the optional number of clusters, the values of *BIC* are calculated. The estimated initial number of clusters is ruled by the minimum value of *BIC*.

The second criterion is called the Akaike Information Criterion (*AIC*) and is calculated according to the formula:

$$AIC(k) = -2\sum_{i=1}^{k}\lambda_i + 2w_k \ . \tag{6}$$

The optimal number of clusters is determined by the same principle as in the case of *BIC*.

For the evaluation of resulting clusters obtained by divisive methods we use the silhouette coefficient (*SC*), which expresses the silhouette measure of cohesion and separation. The silhouette coefficient for individual *i*-th object from the *h*-th cluster is calculated according to the formula:

$$SC(i) = \frac{\mu_i - \eta_i}{\max\{\mu_i;\eta_i\}}, \tag{7}$$

where n_i is the average distance of the individual *i*-th object with all other objects within the same cluster and:

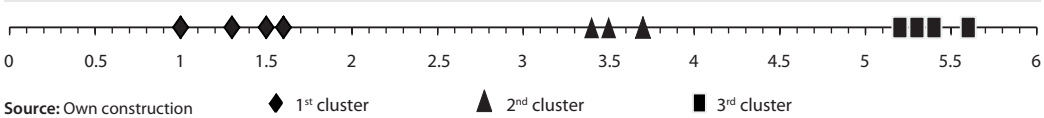
$$\mu_i = \min_{g\neq h}\left(\frac{\sum_{j\in C_g} D_{ij}}{n_g}\right), \tag{8}$$

where C_g is the *g*-th cluster and D_{ij} is the distance between the *i*-th and *j*-th objects.

Using Formula (7), average values for individual clusters are determined as well as the average value for the whole decomposition. The higher the average value is, the more compact the cluster is.

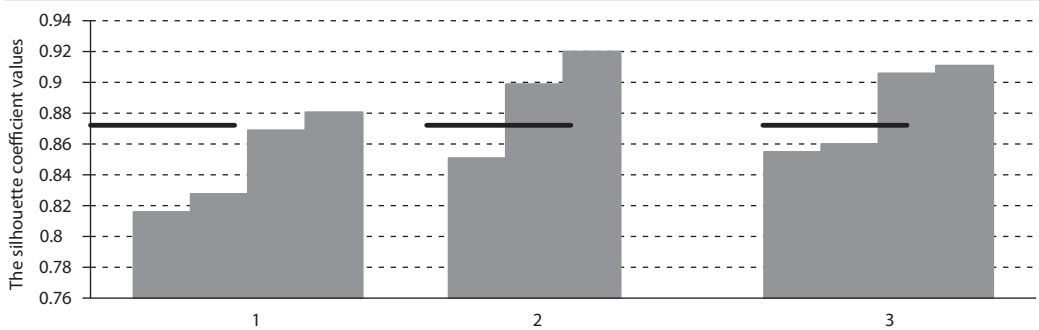
The following three figures show a simple and illustrative example of silhouette coefficient. Figure 1 presents the situation of the eleven objects divided into three clusters. In Figure 2 can be seen graphical representation of both all individual values *SC* (gray bars) and the resulting average *SC* (black dashed line). Figure 3 shows *SC*, which is the output of the system when applying SPSS TwoStep method.

Figure 1 A data file containing 11 objects divided into three clusters



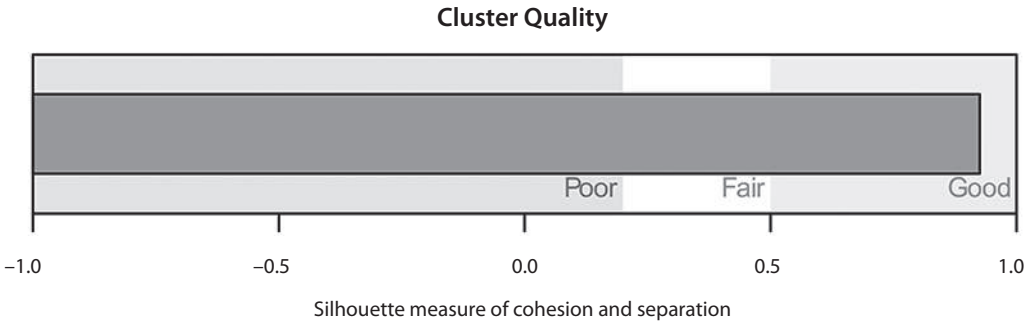
Source: Own construction

Figure 2 Silhouette coefficient for objects in three clusters



Source: Own construction

Figure 3 Silhouette coefficient – the output from SPSS



Source: Own construction

2 CASE STUDY

In our work, we focused on the segmentation of airports using cluster analysis. Each airport stands for an object to be clustered. We analysed data of 838 airports from a total of 977 monitored ones. The data consisted of numbers of passengers who passed through the particular airport per month. Data were collected in the thesis work (Darda, 2014), individual data were obtained partly from the Institute of Civil Aviation (Service technique de l'aviation civile) with headquarter in Paris and from the French Ministry for ecology, sustainable development and energy (Ministère de l'écologie, du Développement durable et de l'Énergie), headquartered in Paris.

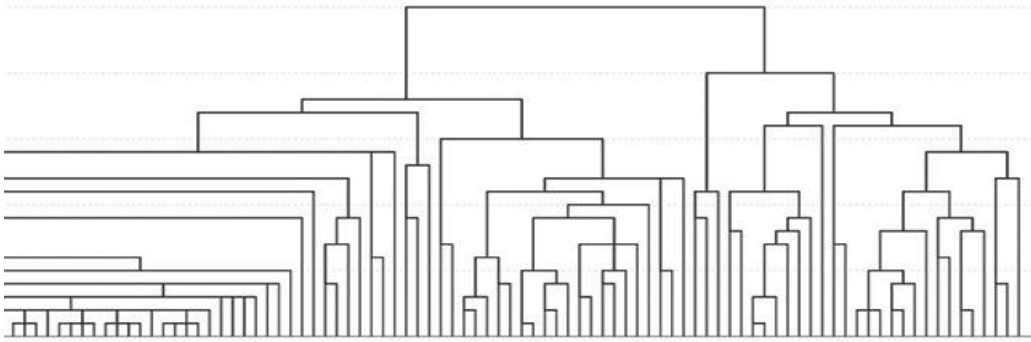
Data were monitored in the period from January 2000 to April 2014. Some airports (mainly Asian) publish data for up to year-end summary, therefore, we restricted our analysis to the period at the end of 2013. World airports, about which we were not able to provide all required information, were not included in the processing. The annual throughput of passengers through each of airports was another factor considered in processing. Airports with the annual throughput lower than 100 000 passengers were excluded. Airports where the statistical data on a monthly basis are published only once per year are also not included in our dataset. This is mainly the case of Asian, particularly Chinese airports where statistics are always published in early April of the following year. Data about several airports were not available since 2000, therefore, we could not incorporate them into the analysis. Complete data about 838 airports were collected from the beginning of 2000 until the end of 2013, thus the input data matrix contains 838 rows and 168 columns.

It should be recalled that the aim of the analysis is first of all to compare the trends in number of handled passengers, so that the absolute values of passengers handled become irrelevant. Therefore, the data for each airport were standardized, subtracting from data on each row corresponding row mean and dividing them by corresponding row standard deviation. We assume these new transformed data to be representatives of quantitative continuous random variables and further as the input for our cluster analysis.

For segmentation, we selected three methods implemented in SPSS. These were the hierarchical method, the TwoStep method and the *k*-means method. The first two methods were used to determine the optimal number of clusters, the third method for the analysis itself.

The final dendrogram was the output of hierarchical clustering (with use Average Linkage method and Euclidean measure). The entire dendrogram was very confusing due to the large number of objects. Its interesting part is shown in Figure 4. Still, it was clearly visible that the suitable number of clusters is two or three.

Figure 4 The interesting part of a dendrogram – the output from SPSS

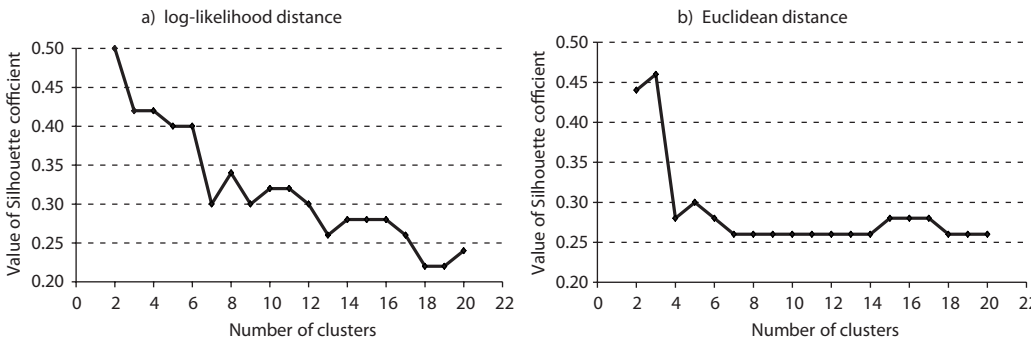


Source: Own construction

Processing with the use of the TwoStep method showed similar results. We used all three indexes implemented in SPSS to monitor the quality of clustering, namely *BIC*, *AIC* and *SC*.

When selecting the log-likelihood distance, the TwoStep method showed three clusters to be the optimum number, for both *BIC* and *AIC* criteria. When the Euclidean distance was selected, the optimal number of clusters turned to be two. Further, we used TwoStep method with both the log-likelihood and Euclidean distance for fixed number k of clusters, $k \in \{2, ..., 20\}$, and calculated corresponding silhouette coefficients SC_k. The values SC_k for each reached distance are plotted in Figure 5.

Figure 5 Graph of silhouette coefficient for TwoStep method



Source: Own construction

It is clear from Figure 5a) and 5b) that the maximum value of *SC* was achieved in case of three and two clusters. The SC value was never lower than 0.2, which means that even in the worst case the quality of clustering was fair. In case of using the Euclidean distance the resulting clusters were highly unbalanced in terms of the number of objects in different clusters. Therefore, for further processing the log-likelihood distance was selected as more appropriate one.

Summarizing, the choice of either two or three clusters appears to be the best choice while using various indicators. Unfortunately, the resulting clusters were not satisfactorily interpretable in either of these theoretically recommended cases. However, we received interpretatively interesting results using k-means method or TwoStep method when choosing a parameter determining the required number of clusters equal to four and then eight. Either the values of the three indicators of quality or the dendrogram does not condemn this solution. Therefore, we will discuss these cases below.

2.1 Analysis of the results for *k*-means method with Euclidean distance and four clusters

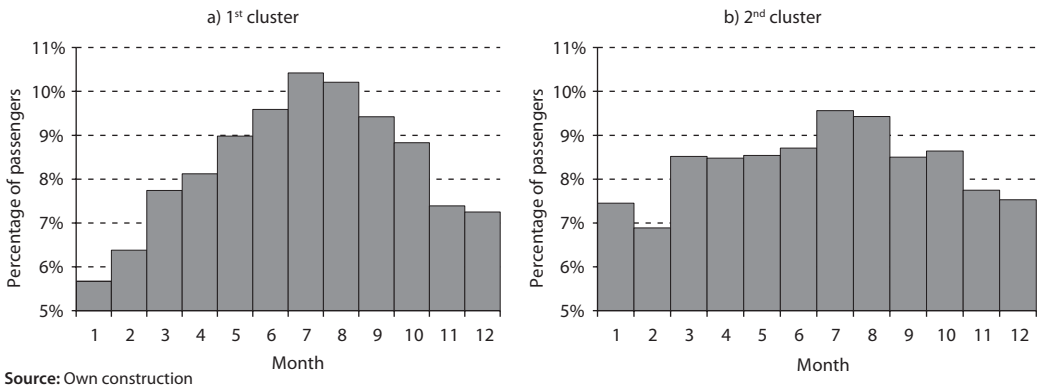
The airports which handled the most passengers during the summer months are included in the first cluster. This cluster contains 33.65% of the monitored airports, and is thus the second largest one. In this category, the number of passengers at the airports increases every month since the beginning of the year until the end of June. The number of passengers reaches the maximum values during July and August. Then again, the number of handled passengers gradually decreases. The end and the beginning of the year have the same or similar values. Therefore, the trend has a recurring character. Many European airports are represented in this category. Many airports from this cluster are also located in North America.

Almost all the airports from cluster one are located in the Northern Hemisphere, where summer culminates in June-September. Demand for air travel increases dramatically in this period, since there is the summer holiday period in many countries. Several airlines operate out of these airports so-called charter and seasonal flights, which carry large numbers of passengers to tourist destinations. This leads to the fact that some airports in the Mediterranean region handle more than 80% of passengers during the summer. During the rest of the year, they handle the remaining 20%. This is demonstrated, for example, on the aggregated group of Greek airports, where 69% of passengers are handled during the summer. Similar indicators can be found also in other Mediterranean airports that experience the greatest rush of passengers during the summer months, such as Spain, France, Italy, Montenegro, Turkey, Egypt and Tunisia. The proportional distribution of handled passengers during the year is shown in Figure 6a).

The second cluster of airports includes 20.88% of the monitored airports. Trends in the number of passengers handled during a year at these airports are similar to those from the first cluster, but with the difference that the increase of passengers in summer is not that dramatic. This means that the traffic and operations in these airports are more balanced during the year. The months of July and August represent the highest percentage of passengers handled, which transcend the boundaries of 9.5%. At the beginning and the end of the year, the percentage is much lower, ranging between 6% and 8%. Proportional distribution of passengers handled during the year is shown in Figure 6b).

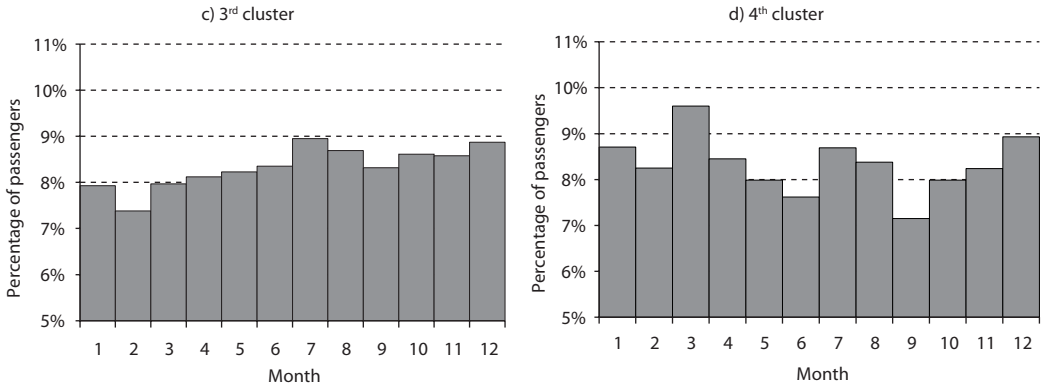
In this cluster, airports from any continent are not predominant as in the case of the previous cluster, three continents being more or less equally represented, namely North America, Europe and Asia. As in the first group, the airports included here operate with the most traffic in the summer. We suppose that one of the main reasons why these months prevail again is the summer season in the Northern Hemisphere, that brings increased tourism activity. This is certainly true in the case of Europe and North America, but we do not think that it would be possible to be applied to airports located in Japan.

Figure 6 Proportional distribution of handled passengers during the year which is characteristic for the airports of the clusters



Source: Own construction

Figure 6 Proportional distribution of handled passengers during the year which is characteristic for the airports of the clusters – continuation



Source: Own construction

Airports with relatively balanced year-round operation represent the third cluster. It is the strongest cluster in the number of airports, precisely 33.77% of the monitored airports. These airports handled 36.35% of passengers of the total number of passengers.

Considering the number of passengers handled per month, these airports are relatively stable during the whole year. Compared with two previous groups, there are not any significant fluctuations in this cluster. In this case, the number of passengers handled per month compared to the total annual number oscillates between 7% and 9%. Therefore, the fluctuation is only about 2%. Proportional distribution of handled passengers during the year is shown in Figure 6c). These airports are located worldwide. We suppose that there exist two main explanations for interpretation of such a distribution.

The first explanation is the transitivity of these airports. There are several important airports belonging to this group such as the one in Dubai, Beijing, Hong Kong, Bangkok, Singapore, Istanbul, Shanghai, Seoul and American Charlotte. Most of these airports have lines serving all inhabited continents and most of countries in these continents. In our opinion, the balanced character of their operation lies in a dense network of destinations. For example, the airport in Singapore used to be a key transit point between Australia and the UK until March 2013 as it was used by the Australian airlines QANTAS. QANTAS have selected a new transit airport – Dubai after the termination of cooperation. Nowadays, Dubai became one of the largest transit airports in the world. If we look at the exact statistics on the number of passengers handled at these airports, we find essentially no difference since this line of company QANTAS had a negligible share of passenger transport between Australia and the UK. In other words, it is obvious that a transit airport gain passengers from dozens of lines. Accordingly, the sudden or forward known loss of one or several airlines does not have considerable importance to the fluctuation of handled passengers.

In the case of airports where geographical conditions make it difficult or even impossible to travel by other means of transport are the second group in this category of airports with a balanced operation. Examples which demonstrate this cluster well are airports in India or Brazil, where travelling by train or car from one end of the country to another is very time consuming. Furthermore, in this cluster, there are also airports located in island countries, such as Japan, the Philippines, Indonesia and South-east Asian countries.

The airports that handle larger numbers of passengers during the months at the beginning and at the end of the calendar year form the last cluster. This is the smallest cluster of airports generated by our analysis. The number of passengers reached only 6.95% of the total number of handled passengers at all examined airports. The greatest number of passengers at these airports occurs in the first quarter

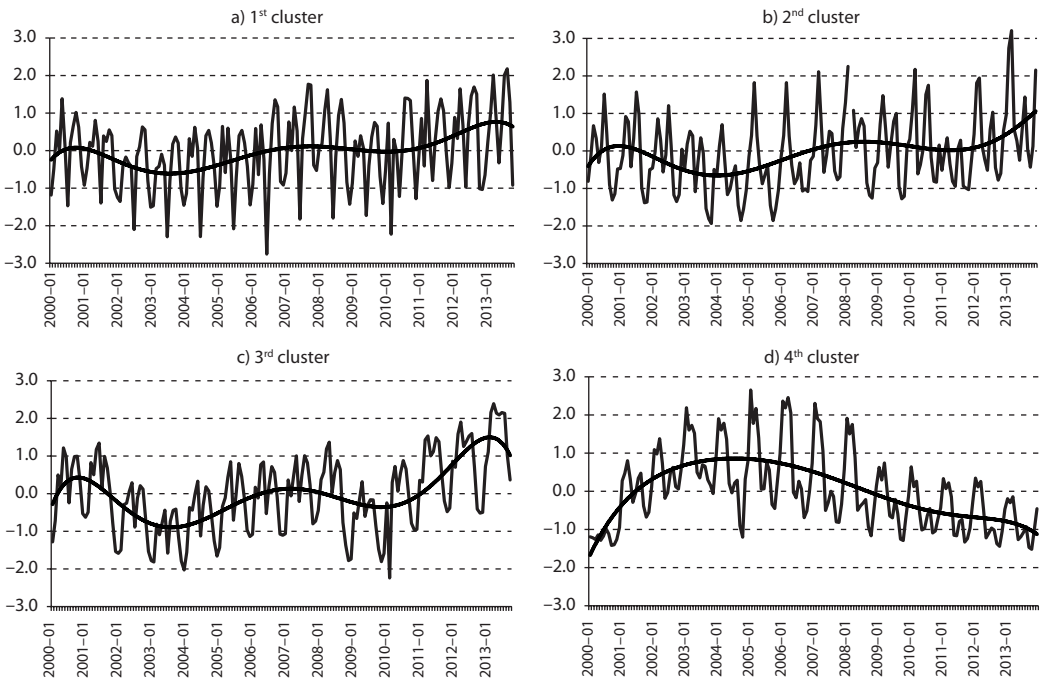
of the year. Then the number slightly decreases. It achieves very small numbers in the months of May and June compared with the previous month. Number of passengers increases slightly during the summer months and culminates its growth in the last quarter of the year.

If we disregard at this moment the months of July and August, which are relatively rich in passengers in each category due to the position of most major airports in the Northern Hemisphere and tourism, we find that the months from January to April and October to December provide a large percentage of handled passengers. From 8% to 10% of passengers are handled during these months. Proportional distribution of handled passengers during the year is shown in Figure 6d).

In this group, there are the airports in the Southern Hemisphere, particularly airports in Australia, New Zealand or South America. Trends in the number of passengers handled in this category can be described similarly as we characterized trends of the first and second category. In the Southern Hemisphere, summer culminates in months around the turn of the year. Conversely, there is winter in the Northern Hemisphere. Tourism brings an increased number of passengers into these thermopile areas. A group of airports in the Caribbean Sea and the Gulf of Mexico have the same tendency. The principle of airports functioning in the Southern Hemisphere is nearly symmetrical compared to the Northern Hemisphere.

There is a second significant group of airports which belong to this cluster. These are airports situated very close to the two poles of Earth. These are mainly airports in Scandinavia, Canada and the southern part of South America. We think that very low temperatures and frozen water make it difficult to use land or water transport. Therefore, air transport prevails in these months.

Figure 7 Time series of normalized monthly values of the number of handled passengers in the airports belonging to different clusters



Source: Own construction

SUMMARY

During processing, we wiped out all four time series resulting from averaging the values for all airports of the cluster using polynomials of sixth grade. The courses of these four regression functions were similar (except 4th cluster). It is seen from Figures 7a) to 7d). It can be concluded that the clusters do not differ too much in terms of long term evolution. Substantial difference was demonstrated in terms of seasonality.

In the first group, there are airports with a significant increase in passengers during summer months. The second group of airports shows a similar situation as the first group, but the summer increase in passengers is not as significant. The third group of airports has a balanced number of transported passengers during the whole year. The fourth group consists of airports, where the number of transported passengers is the highest in winter months. The main factor determining this division is therefore seasonal development during the year. The most important characteristics of individual clusters are summarized in the Table 1.

Table 1 Characterization of the clusters

	1st cluster	2nd cluster	3rd cluster	4th cluster
Number of airports	282	175	283	98
Cluster proportion (%)	33.65	20.88	33.77	11.69
Proportion of transported passengers (%)	44.68	12.02	36.35	6.95
Prevailing geographic location	Europe, North America, Japan	Japan, North America, Europe	Asia, Africa, Australia	Mexico, Scandinavia, SE Asia, New Zealand

Source: Author's calculations

2.2 Analysis of the results for *k*-means method with Euclidean distance and eight clusters

We also received interpretatively interesting results when we used the *k*-means method and selected the parameter determining the required number of clusters equal to eight. In all 8 clusters the significant impact of the world economic crisis 2008 was obvious.

There were 49 airports in the first cluster. Of these, 35 were from Asia, and more than half of them were from South Korea (the largest of representatives was the Gimpo Airport), also from Thai (the largest representative was the Phitsanulok airport), but also from Japan (Kansai airport). Among major airports from other continents, there were for instance the New Orleans airport from the USA, or Swedish Växjö.

All airports in this cluster suffered a significant decrease in the number of checked passengers after September 11, 2001. There is another large decline at the end of 2003. This decline began to mitigate until the end of 2005. It can be deduced from knowledge of world events that airport operations were in part influenced by the terrorist attacks of September 11, 2001, but even more by period of SARS epidemic, which erupted at the end of 2002. The impact of the SARS epidemic on aviation is described in Loh and Elaine (2006), see Figure 8a.

There were 29 airports forming the second cluster. Of these, 18 came from Europe, especially from the tourist centres. The Spanish airports Grand Canaria and Tenerife South, the Austrian airports in Innsbruck and Salzburg and Italian Turin, belong among the largest of them. Since airports in this cluster are characterized by their distinct seasonality, they experience a strong increasing trend of handled passengers in the observed period. Events of September 11, 2001 are slightly noticeable (Figure 8b).

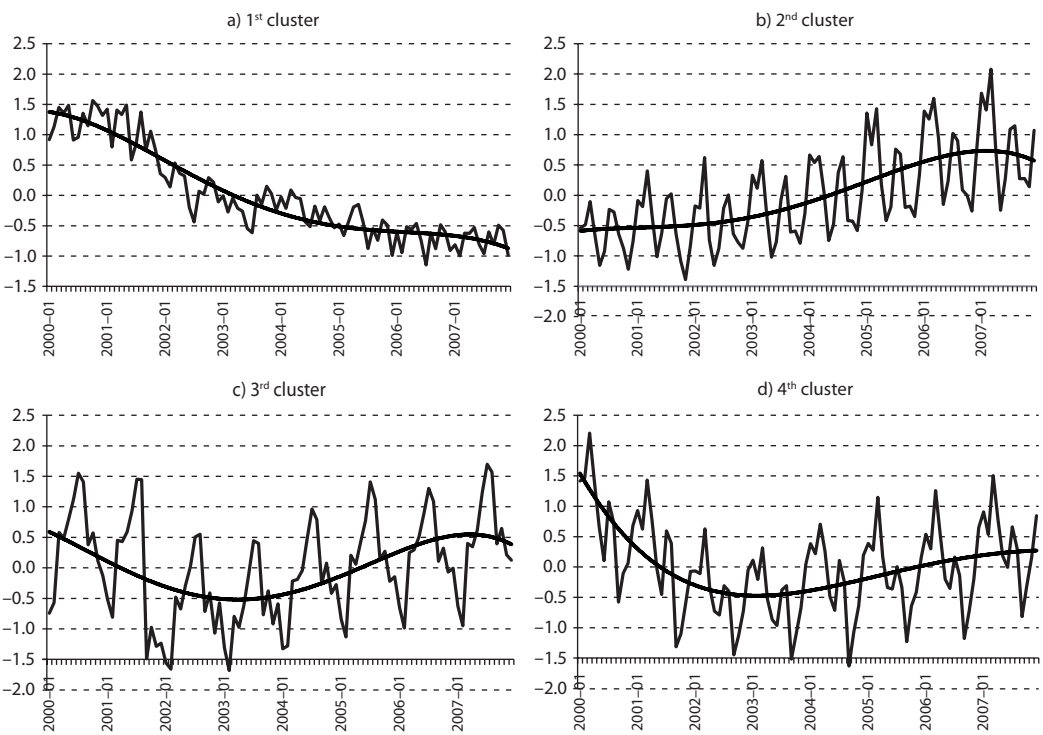
There were 98 airports included in the third cluster. The vast majority of them come from the US. Chicago, Los Angeles and Dallas-Fort Worth are the most important representatives. In this cluster, there are also 12 European airports. The airport Paris Orly and the airport in Brussels, but also Nordic airports, are among the most important ones.

The very strong decline since September 2001 at airports in this cluster is obvious in Figure 8c). The number of handled passengers decreased rapidly compared to the year 2000. After the slump, the same number as in 2000 was not achieved until 2007. Many publications deal with the influence of terrorist attacks of 11 September 2001, namely publication of airlines as (IATA) and scientific papers, such as (Dempsey 2003; Chen, C.-C et al., 2009; Cui and Li, 2015).

The fourth cluster consisted of 29 airports, 15 of them is located in Mexico (Acapulco is the largest representative). This cluster includes e.g. the US Miami airport or the Puerto airport in the Dominican Republic. In this cluster, the decrease in the number of checked passengers since September 2001 was not that dramatic compared with the previous one. However, the decline at the end of 2003 is more significant. Unlike airports from the second cluster, these airports failed to restore the status of early 2000, yet by this time (see Figure 8d).

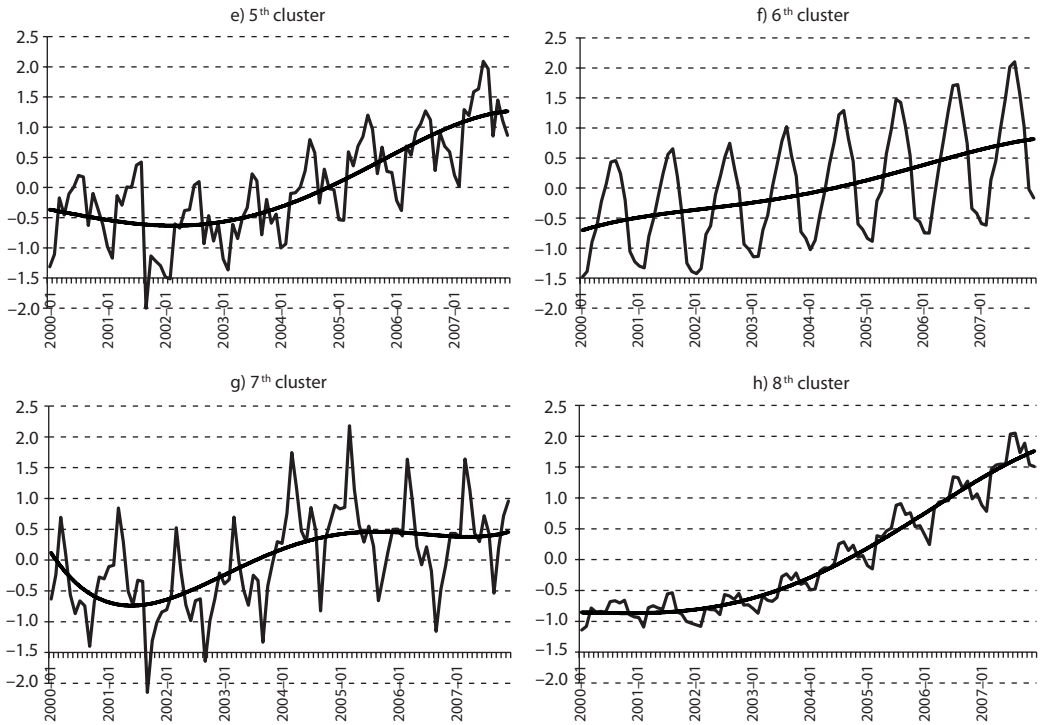
The fifth cluster included 121 airports, of which 104 were located in North America, primarily in the USA. Among the most important these were airports in Atlanta, Denver, Houston and Las Vegas, Pheonix. Airports in Mexico City or Canadian Montreal were another important representatives of this cluster. It is very strong decrease in the number of passengers handled after September 2001 which is characteristic

Figure 8 Time series of normalized monthly values of the number of handled passengers at the airports belonging to different clusters



Source: Own construction

Figure 8 Time series of normalized monthly values of the number of handled passengers at the airports belonging to different clusters – continuation



Source: Own construction

for airports in this cluster. Thenceforth, a growing trend is evident. As seen from Figure 8e), the current condition is at a higher level than at the beginning of the observed period.

There were 213 airports forming the sixth cluster. The vast majority of them are European. A slightly increasing trend of handled passengers during the whole period is evident for all these airports. In addition, strong seasonal behaviour is manifested here. There is above-average amount of passengers checked during holidays and vice versa strongly below-average amount around the turn of the year (Figure 8f).

In the seventh cluster, there were 26 airports included. Most of them are located in North America. Tampa and Fort Lauderdale in the USA and Mexican Cancun are the most important representatives. The growing trend throughout the incriminated period with a noticeable downturn after September 2001 is characteristic for representatives of this cluster (see Figure 8g).

The eighth cluster consisted of 268 airports. Almost half of them are located in Europe. Most Euro pean airports are from Spain (Madrid Barajs is the largest one) and Italy (Milan Linate). Strongly represented are also Australian airports (Sydney and Melbourne) as well as airports in Asia (Hong Kong, Beijing, Mumbai). Some of Canadian and Mexican airports are included here, too. This cluster also contains African and South American airports and airports in the Middle East. Significantly increasing trend of handled passengers is typical for airports included in this cluster. This is clearly seen in Figure 8h. he description of the cluster is in accordance with the description of development, for example air traffic in China, which is described in the paper (Wang J., Mo and Wang F., 2014).

Table 2 Characterization of the clusters

	1st cl.	2nd cl.	3rd cl.	4th cl.	5th cl.	6th cl.	7th cl.	8th cl.
Number of airports	49	30	98	33	121	213	26	268
Cluster proportion (%)	5.85	3.58	11.69	3.94	14.44	25.42	3.10	31.98
Proportion of transported passengers (%)	2.66	1.66	24.09	1.65	22.89	28.51	2.85	15.68
Prevailing geographic location	South Asia, North America	Europe	USA	Mexico	North America	Europe	North America	Europe, North America

Source: Author's calculations

Table 3 Cross numbers of airports belonging to the individual clusters

	1st cl.	2nd cl.	3rd cl.	4th cl.	5th cl.	6th cl.	7th cl.	8th cl.	Suma
1st cl.	0	1	41	0	118	122	0	0	282
2nd cl.	47	4	19	0	0	91	14	0	175
3rd cl.	0	21	7	0	3	0	0	252	283
4th cl.	2	4	31	33	0	0	12	16	98
Suma	49	30	98	33	121	213	26	268	838

Source: Author's calculations

SUMMARY

Result of our clustering indicates that various global disasters are important factors affecting the number of handled passengers. Within individual clusters it is possible to distinguish airports, which has not been much affected by these disasters, from the airports on which these disasters had either short-term or even long-term impact. The most important characteristics of individual clusters are summarized in Table 2. Cross numbers of airports belonging to the individual clusters are shown in Table 3.

3 DISCUSSION

It is obvious that from the interpretative point of view there is no optimal number of clusters. The resulting interpretation of division into 8 clusters led us to the idea that the other key accident may occur in some clustering results. In literature, the impact of the eruption of the Eyjafjallajökull volcano in Iceland on air transport (April–May 2010) is often quoted. The volcanic cloud created after the explosion, gradually closed European airports between 15 and 23 April as the cloud progressed through Europe. The paper (O'Regan, 2011) describes serious consequences for the operation of air transport in Europe after the eruption of Eyjafjallajokull.

Unfortunately, in this case, the classification was not very successful. Cluster which could be identified as groups of airports affected by this event appeared only when the value of the parameter determining the number of clusters was set at 15. However, in such a large number, clusters cannot be unequivocally interpreted.

What may justify this failure in identification using cluster analysis? The time period affected by the event was very short. Restrictions in aviation lasted only a month. The trend of development of numbers of passengers in neighbouring periods overrode this difference. Therefore, airport affected by the volcanic eruption were not separated well into a separate cluster.

CONCLUSION

We show that cluster analysis can be used to classify airports on the basis of the number of handled passengers per each individual month. It turned out that this classification has quite interesting interpretation. In one case, we managed to classify the world's airports in terms of seasonal development in the number of handled passengers. In a more detailed division we managed to classify the world's airports in terms of their reactions on world events which had an impact on air traffic. It turned out that an important factor resulting in event classification is the sufficient length of the period during which the consequences of this event persisted at particular airports. Conversely, it proved that the regional arrangement is not important for classification in the first place. Moreover, it turned out that if the cluster analysis was used for closer examination of the data structure from different perspectives, it is not good to restrict itself just to division into "ideal" number of clusters.

A side clustering of the airports here requires another type of analyses of available data, especially of trends and changes in them, and the reasons behind eventual changes. Unfortunately, these problems are behind the scope of this paper. On the other hand, our preliminary results based on so called change point analysis as described, e.g. in (Antoch et al., 2007; Antoch et al., 2004; Antoch et al., 2008; Antoch et al., 1997; Antoch and Jarušková, 2017), show very promising results. It appears that if we take into account the fact that these types of data are obtained sequentially in time, the cluster analysis with time series analysis and change point analysis can lead to more profound explanation of studied problems. We will cover this approach elsewhere.

ACKNOWLEDGMENT

This article was supported by Jan Evangelista Purkyně University in Ústí nad Labem.

References

AKAMAVI, R. K., MOHAMED, E., PELLMANN, K. et al. Key determinants of passenger loyalty in the low-cost airline business. *Tourism management*, 2015, 46, pp. 528–545.

ANTOCH, J., GREGOIRE, G., HUŠKOVÁ, M. Tests for continuity of regression functions. *Journal of Statistical Planning and Inference*, 2007, 137(3), pp. 753–777.

ANTOCH, J., GREGOIRE, G., JARUŠKOVÁ, D. Detection of structural changes in generalized linear models. *Statistics & Probability Letters*, 2004, 69(3), pp. 315–332.

ANTOCH, J., HUŠKOVÁ, M., JANIC, A., LEDWINA, T. Data driven rank test for the change point problem. *Metrika*, 2008, 68(1), pp. 1–15.

ANTOCH, J., HUŠKOVÁ, M., PRÁŠKOVÁ, Z. Effect of dependence on statistics for determination of change. *Journal of Statistical Planning and Inference*, 1997, 60(2), pp. 291–310.

ANTOCH, J. AND JARUŠKOVÁ, D. Detection of breaks in capital structure. A case study [online]. *Statistika: Statistics and Economy Journal*, 2017, 97(1), pp. 32–43.

CHEN, C.-C., CHEN, J., LIN, P.-C. Identification of significant threats and errors affecting aviation safety in Taiwan using the analytical hierarchy process. *Journal of Air Transport Management*, 2009, 15(5), pp. 261–263.

CUI, Q. AND LI, Y. The change trend and influencing factors of civil aviation safety efficiency: The case of Chinese airline companies. *Safety Science*, 2015, 75, pp. 56–63.

DARDA, P. *The economic importance of passengers for airports.* Master's thesis (in Czech), Univerzita J. E. Purkyně, ESF: Ústí nad Labem, 2014.

DEMPSEY, P., S. Aviation security: The role of law in the war against terrorism. *Columbia Journal of Transnational law*, 2003, 41(3), pp. 649–733.

GRABBE, S., SRIDHAR, B., MUKHERJEE, A. Clustering days and hours with simile airport traffic and weather conditions. *Journal of Aerospace Information Systems*, 2014, 11(11), pp. 751–763.

GRENČÍKOVA, J., KRIŽAN, F., TOLMÁČI, L. Stability and actuality of aviation networks in Bratislava and Prague. *Moravian Geographical Reports*, 2011, 19(1), pp. 17–31.

IATA, *The Impact of September 11 2001 on Aviation* [online]. 2014. [cit. 04.03.14]. <http://www.iata.org/pressroom/documents/impact-9-11-aviation.pdf>.

KRAFT, S. A transport classification of settlement centres in the Czech Republic using cluster analysis. *Moravian Geographical Reports*, 2012, 20(3), pp. 38–49.

LOH, E. The impact of SARS on the performance and risk profile of airline stock. *International Journal of Transport Economics*, 2006, 33(3), pp. 401–422.

LU, Q., CH., ZHANG, J., PENG, Z., R., et al. Inter-city travel behaviour adaptation to extreme weather events. *Journal of Transport Geography*, 2014, 41, pp. 148–153.

O'REGAN, M. On the edge of chaos: European aviation and disrupted mobilities. *Mobilities*, 2011, 6(1), pp. 21–30.

ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V. Clusters number determination and statistical software packages. *DEXA 2008: 19th International Conference on Database and Expert Systems Applications*, 2008, pp. 549–553.

WANG, J., MO, H., WANG, F., Evolution of air transport network of China 1930–2012. *Journal of Transport Geography*, 2014, 40 (October), pp. 145–158.

ZHANG, T., RAMAKRISHNAN, R., LIVNY, M. BIRCH: An efficient data clustering method for very large databases. *ACM SIGMOD Record*, 1996, 25(2), pp. 103–114.

ZHANG, T., RAMAKRISHNAN, R., LIVNY, M. BIRCH: A new data clustering algorithms and its applications. *Journal of Data Mining and Knowledge Discovery*, 1997, 1(2), pp. 141–182.

Steel Augmented Production Function: Robust Analysis for European Union

Bilal Mehmood [1] | *GC University, Lahore, Pakistan*
Muhammad Aleem [2] | *GC University, Lahore, Pakistan*
Marwah Rafaqat [3] | *GC University, Lahore, Pakistan*

Abstract

This study contributes to the empirical literature on augmented neo-classical production function. It is done by introducing steel production into macro-production function of the European Union. The data is collected from the World Development Indicators and the World Steel Association from the period of 1980–2014. We apply second generations of unit root tests to examine stationarity and panel cointegration with cross-sectional dependence to analyze long run relationship between national income and steel production. Robustness of tests is also reached by using 23 estimators and country specific slopes. Whereas, to detect the cause and effect, Granger and Dumitrescu-Hurlin causality tests are applied. Uni-directional causality from national income to steel production is found. Recommendations are made on the basis of empirical results.

Keywords

Steel production, national income, augmented mean group, panel causality

JEL code

D24, E01, C23

INTRODUCTION

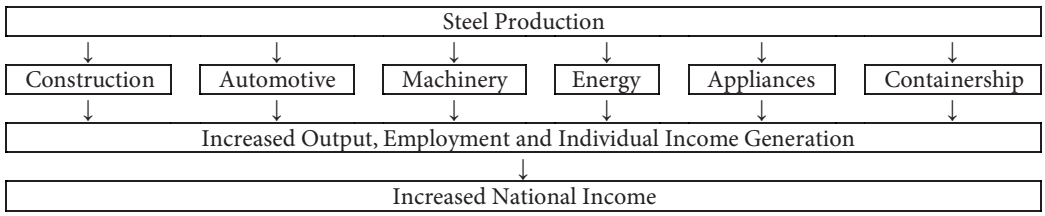
Steel is not a nascent alloy, however, its manufacturing at commercial scale and organized trade started only after industrial revolution. Events related to changes of fuel industry and collection of technological advances during 1600s, 1700s and 1800s laid the foundation of the contemporary steel industry. During 1830 to 1860, steel was used as a semi-precious metal in expensive products. 'The Great Transformation' era (1860–1900) was mainly attributed to by low cost 'open hearth' and 'Bessemer' methods of steel production that spurred the growth of steel production by seventeen fold. Establishment of US steel companies led to consolidation period during 1900–1920. However, steel industry also felt depression during 1930s due to 'Great Depression' and the rise of labor movement. The revival of steel industry was triggered by World War-II during 1940-1945. Warring European countries used steel for manufacturing arsenals, tanks, trucks, ships and other war weapons rendering steel industry a giant industry. The post-war period (1946 to 1970) is called the 'period of prosperity' due to growth of steel industry, evolution of recycling segment of industry and development of substitute, such as aluminum.

[1] Assistant Professor, GC University, 1 Ketchry Road, 540 00 Lahore, Punjab, Pakistan. E-mail: dr.bilalmehmood@gcu.edu.pk.
[2] MPhil Scholar, GC University, 1 Ketchry Road, 540 00 Lahore, Punjab, Pakistan. E-mail: aleemshaikh16@gmail.com.
[3] Graduate, GC University, 1 Ketchry Road, 540 00 Lahore, Punjab, Pakistan. E-mail: marwahrafaqat@yahoo.com.

Alongside prosperity, steel also witnessed 'troubled times' during the period of 1970–1989, when many plants were closed, production declined and layoffs took place. The intensity of this trouble was mitigated during period of 1990–2001 which is called the era of 'uneasy trouble'.

The usage of steel is universal as the World Steel Association states, "steel is everywhere in your life". Related benefits include employment generation and infrastructural effects of providing infrastructure to industrial and modern sectors. In 1970 steel industry employed 531.196 people and even after its decline in 2000, it still had 225.000 on its payroll. Its backward & forward linkages play a significant role in development via providing critical inputs like machinery for developmental projects. In addition to assistance to developmental projects, public sector also gains from tax revenue by steel industry. Steel industry can also indirectly contribute to foreign exchange reserves by assisting industries in the production of exports. In addition, steel industry also contributes to agriculture sector by providing tractors, aerial spray, and harvesters etc. that increase the per acre yield of crops which will ultimately increase the national income.[4]

Figure 1 Mechanism of Contribution of Steel Industry in National Income



Source: Authors' formulation

Based on the pictorial explanation in Figure 1, this paper attempts to quantify the relationship between steel industry and national income in the European Union.

Hypothesis

Based on the objective, following hypothesis shall be tested:

$\mathbf{H_A}$: There exists a causal and long run relationship between steel production and national income of the European Union.

1 LITERATURE REVIEW

Subject of this research has not been chosen by many of the researchers and it is due to the fact that there exists limited literature on it. Jeferrson (1990) using Chinese data investigated the productivity variation among enterprises within China's steel and iron industry. He found enhanced productivity growth during reform period within the industry. Labson and Crompton (1993) studied relationship between income and five industrial metals for Japan, OECD, USA and UK for the period of 1960–1987. However, they proposed slight explanation to support the existence of long run relationship between two variables. Hoechle (2007) studied the energy efficiency of China's steel and iron sector for the time span of 1994–2003 using Malmquist decomposition index. Provincial panel data was used permitting various energy inputs and outputs. Results revealed that empirical productivity of China's steel and iron sector increased by 60% from 1994 to 2003 which is a sign of technological progress.

Evans (2011) analyzed the long run relationship between crude steel and economic activity production in United Kingdom. He used integrated processes and allowed for the possibility of changes in equilibrium

[4] For more on sectoral contribution of steel industry, visit website of the World Steel Association.

path. Evidence is found in support the long term relationship. Huh (2011) studied the short run and long run relationship between steel production and GDP in Korea from 1975 to 2008. He used vector error correction and vector autoregressive models. He found a long term causal relationship, running from GDP to total steel production. He also found the bi-causal relationship between flat product consumption and GDP. Ozkan (2011) analyzed the relation between steel production consumption, import & export, GDP and industrial production. They applied error correction model, Engle-Granger cointegration and Granger causality test. Their results revealed a positive relation of steel export and production with GDP. A positive relationship was also found between industrial production and steel export. Both relations showed causality effects.

Siddique, Mehmood and Ilyas (2016) analyzed the relationship between economic growth and steel production in Pakistan. They used time series data to apply Philip Person (PP) and Augmented Dickey Fuller (ADF) test and Cointegration with Bai-Perron structural breaks test to check the long run relationship between the two variables. Their results show a positive relationship between steel production and economic growth with causality from economic growth to steel production.

Review shows that majority of studies are limited to a single country not allowing the benefits of panel data analysis. Moreover, possible effects of common shocks are not incorporated either. Important variables like capital and labor that play a critical role in any production are also missing in till-date empirical literature. Though steel industry is capital intensive, yet labor employment is also substantial due to need for manual labor in mega-structures. European Union (EU) is the largest producer of steel after China. Research on this sample should allow for improved policy directions. Current paper does so by choosing a sample of EU.

2 ESTIMABLE PRODUCTION FUNCTION

The estimable production function for testable prediction that steel production and national income have nexus in European Union countries is given as follows:

$$NI_{i,t} = f(ST_{i,t}, CP_{i,t}, LB_{i,t}), \tag{1}$$

where:

$NI_{i,t}$ = GDP (constant 2005 US$),
$ST_{i,t}$ = Steel production (thousand tones),
$CP_{i,t}$ = Gross fixed capital formation (constant 2005 US$),
$LB_{i,t}$ = Labor force, total,
i and t stand for cross-sections and time periods, respectively.

2.1 Methodology – Data Sources

Depending on the availability of data, 28 EU countries are selected while the number of years is 35 (1980–2014). Sample countries are Austria, Belgium, Bulgaria, Croatia, Republic of Cyprus, Czech Republic, Denmark, Estonia, Finland, France Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden and UK. Collection of data is done from World Development Indicators (WDI) and World Steel Association.

3 EMPIRICAL ANALYSIS

3.1 Static Estimations

In order to examine the empirical relationship of national income and steel production, following analysis is conducted. We estimated static models that are devoid of any lagged dependence

of dependent variable. These include pooled OLS (POLS: $NI_{it} = \alpha + \beta_{ST}.ST_{it} + \beta_{CP}.CP_{it} + \beta_{LB}.LB_{it} + \varepsilon_{it}$), fixed effects (FE: $NI_{it} = \alpha_i + \beta_{ST}.ST'_{it} + \beta_{CP}.CP'_{it} + \beta_{LB}.LB'_{it} + \varepsilon_{it}$), random effects (RE: $NI_{it} = \alpha_i + \beta_{ST}.ST'_{it} + \beta_{CP}.CP'_{it} + \beta_{LB}.LB'_{it} + \beta_0 + \varepsilon_{it}$) and first differenced fixed effect (FD: $NI_{it} = \beta_{ST}.\Delta ST'_{it} + \beta_{CP}.\Delta CP'_{it} + \beta_{LB}.\Delta LB'_{it} + \Delta\varepsilon_{it}$). The estimated coefficients are statistically significant at 1% in POLS, FE, RE and at 5% in FD estimations, respectively. The range of statistically significant coefficients is from 0.0018 to 0.0731. Capital and labour also show desirable signs of their coefficients. R^2 also falls in reasonable range.

Table 1 Static Analysis – POLS, FE, RE and FD-FE Estimates

	POLS	FE	RE	FD
$ST_{i,t}$	0.0221[b]	0.0293[a]	0.0018[c]	0.0731
	(0.011)	(0.004)	(0.001)	(0.049)
$CP_{i,t}$	0.0023[b]	0.0223[a]	0.0024[b]	0.1324[a]
	(0.020)	(0.006)	(0.001)	(0.038)
$LB_{i,t}$	0.7169[a]	0.8099[a]	0.7169[a]	0.1321[a]
	(0.024)	(0.027)	(0.024)	(0.029)
Constant	–0.0731	0.9204[a]	–0.0731	0.0693[a]
	(0.076)	(0.142)	(0.076)	(0.005)
Observations	887	910	887	884
Countries	26	26	26	26
R^2	0.50	0.35	0.50	0.24
CD[‡]	87.60[a]	87.60[a]	83.03[a]	17.06[a]

Note: [‡]CD is the cross-sectional dependence test by Pesaran (2004) and is calculated as $CD = \sqrt{\frac{TN(N-1)}{2}}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{ij}\right)$. [a], [b] and [c] represent statistical significance at 1%, 5% and 10% respectively, whereas standard errors are in parentheses.

Source: Authors' estimates

3.2 Dynamic Analysis

3.2.1 Unit Root Test Results

Table 2 reports the results of unit root tests meant for investigating stationarity in the series, selection of the appropriate lag length was made using the Schwarz Bayesian Information Criterion.

Table 2 Unit Root Tests

	NI_{it}	ΔNI_{it}	ST_{it}	ΔST_{it}	CP_{it}	LB_{it}	ΔLB_{it}
LLC	–0.8	–6.4[a]	1.1	–12.0[a]	–5.4[a]	1.1	–12.0[a]
IPS	–0.6	–10.2[a]	–0.4	–17.9[a]	–2.2[a]	–0.4	–17.9[a]
	NI_{it} is I(1)		ST_{it} is I(1)		CP_{it} is I(0)	LB_{it} is I(1)	

Note: LLC and IPS stand for Levin, Lin and Chu (2002) and Im, Pesaran and Shin (2003) tests respectively. [a] represents statistical significance at 1%.

Source: Authors' compilation

3.3 Cointegration Tests

Results of LLC tests in Table 2 show that NI_{it}, ST_{it}, CP_{it} and LB_{it} have a mixed order of integration, i.e. *I(0)* and *I(1)*. Eberhardt and Teal (2010) suggest the use of macro-panel data techniques when time span is more than 20 years. Here $t = 35$, so we can resort to macro-panel data techniques. Since the series involved in our analysis is not integrated of same order, Pedroni and Kao tests cannot be applied. Therefore, we employ three econometric techniques that allow for mixed order of integration i.e. Mean Group (MG), Dynamic Fixed Effects (DFE) and Pooled Mean Group (PMG). Pesaran and Smith (1995) provided MG

estimator of dynamic panels for large number of time observations and large number of groups. In this method separate equations are estimated for each group and distribution of coefficients of these equations across groups is examined. It provides parameter estimates by taking means of coefficients calculated by separate equations for each group. It is one extreme of estimation because it just makes use of averaging in its estimation procedure. It does not consider any possibility of same parameters across groups. For MG estimator, each parameter is taken as:

$$\overline{\hat{u}_i} = \frac{1}{N} \sum_{i=1}^{N} u_i \qquad \overline{\hat{\theta}_i} = \frac{1}{N} \sum_{i=1}^{N} \theta_i \qquad \overline{\hat{\phi}_i} = \frac{1}{N} \sum_{i=1}^{N} \phi_i \,, \tag{2}$$

where u_i, θ_i and ϕ_i denotes intercept, long run integrating vector and error correction term respectively.

For the averages of the parameters MG estimator will give consistent estimates. Thus allows all parameters to vary across countries, but it does not consider the fact that certain parameters may be the same across groups.

Pesaran and Smith (1997) suggested PMG estimator of dynamic panels for large number of time observations and large number of groups. Pesaran et al. (1997, 1999) added further in PMG and extended it. Pooled mean group estimator considers both averaging and pooling in its estimation procedure, so it is considered as an intermediate estimator. PMG allows variation in the intercepts, short-run dynamics and error variances across the groups, but it does not allow long-run dynamics

Table 3 Dynamic Analysis – Cointegration Estimation

	Mean Group	Dynamic Fixed Effects	Pooled Mean Group
	Long Run Parameters		
ST_{it}	0.0290 (0.046)	6.3398[a] (1.357)	1.0571[a] (0.089)
CP_{it}	0.4412 (0.275)	0.3295[b] (0.135)	1.7952[a] (0.239)
LB_{it}	0.4339 (1.492)	0.5074[c] (0.294)	7.5817[a] (0.952)
	Average Convergence Parameter		
φ_i	–0.4523[a] (0.035)	–0.0447[a] (0.008)	–0.0387[a] (0.011)
S.o.A	2.2 years	22.4 years	25.9 years
	Short Run Parameters		
ΔST_{it}	–0.0147 (0.015)	0.0061[a] (0.001)	0.0533[a] (0.012)
ΔCP_{it}	–0.0263 (0.038)	0.0039[a] (0.001)	0.0259 (0.016)
ΔLB_{it}	0.0325 (0.533)	–0.0628[a] (0.021)	–0.2920[a] (0.092)
C	4.1490[a] (1.588)	10.5039[a] (1.173)	3.2218[a] (1.008)
Observations	451	451	451
Groups	26	26	26
p-value	**(Hausman)$_{MG/DFE}$ = 0.978**		
	(Hausman)$_{MG/PMG}$ = 0.746		
Remarks	**PMG is efficient & consistent**		
CD (MG)	**20.99[a]**		

Note: In parenthesis, standard errors of parameters are given while [a], [b] and [c] represent statistical significance at 1%, 5% and 10%, respectively. φ_i is the error correction term. S.o.A is the speed of adjustment.

Source: Authors' estimates

to differ across the groups. Adopting from Pesaran et al. (1997, 1999), PMG estimable model has an adjustment coefficient φ_i that is known as the error-correction term (ECT). In fact, explains what percentage of adjustments take place in each period. In addition to MG and PMG, DFE is also used to estimate the cointegrating vector. DFE specification controls the country specific effects, estimated through least square dummy variable (LSDV) or generalized method of moment (GMM). DFE relies on pooling of cross-sections. Like the PMG, DFE estimator also restricts the coefficient of cointegrating vector to be equal across all panels.

Results in the Table 3 reveal the comparison of panel cointegration estimation using MG, DFE and PMG. All three alternative methods of cointegration (MG, DFE and PMG) show the long run relationship between the national income and steel production. It is evident from error correction terms (φ_i), which are less than unity and negative in terms of sign with statistical significance at 1% level of significance. However, the most efficient of the three estimators should be relied upon. Its selection is done by employing the Hausman test. The results in Table 5 show statistical insignificance which implies superiority of PMG over MG and DFE. Therefore, the relationship is established under the assumption of absence of cross-sectional dependence.

3.4 Cross-Sectional Dependence

Results of CD test in Table 1 show the presence of cross-sectional dependence in the estimable model. Values of CD test are 87.60, 87.60, 83.03 and 17.06 for POLS, FE, RE and FD respectively. All are statistically significant at 1%, affirming cross-sectional dependence (CD) in residuals of the estimable models. In real life, CD is due to reasons like oil price shock, global financial crisis and local spill over and is common in most of panels.

We examined the CD in residuals and variables using further tests. Friedman (1937) proposed a non-parametric test (R_{ave}) based on Spearman's rank correlation coefficient. It helps in determining cross-sectional dependence. One of the most well-known cross-section dependence diagnostic is the Breusch-Pagan (1980) Lagrange Multiplier (LM) test statistic. Frees (1995) proposed a statistic (R^2_{ave}) which is based on the sum of the squared rank correlation coefficients. Pesaran (2004) proposed a standardized version of Breusch-Pagan LM test (LM_s) , suitable for large N samples. Since (LM) and (LM_s) are likely to exhibit worsening size distortion for small T_{ij} for larger N, Pesaran (2004) proposed an alternative statistic (CD_p) based on the average of the pairwise correlation coefficients. This test is already used in Table 1. The null hypothesis of this test is cross-sectional independence against the alternative hypothesis of cross-sectional dependence. More recently, Baltagi, Feng, and Kao (2012) presented a simple

Table 4 Tests for Cross-Sectional Dependence in Residuals of Estimable Model

Test	Statistic	Value
R_{ave}	$\frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{r}_{ij}$	322.62[a]
LM	$\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}T_{ij}\hat{\rho}_{ij}^2 \rightarrow \chi^2\frac{N(N-1)}{2}$	2804.49[a]
R^2_{ave}	$\frac{2}{N(N-1)}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{r}_{ij}^2$	5.30[a]
LM_s	$\sqrt{\frac{1}{N(N-1)}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(T_{ij}\hat{\rho}_{ij}^2-1)\rightarrow N(0,1)$	97.25[a]
CD_P	$\sqrt{\frac{TN(N-1)}{2}}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{ij}\right)$	39.62[a]
LM_{BC}	$\sqrt{\frac{1}{N(N-1)}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(T_{ij}\hat{\rho}_{ij}^2-1)-\frac{N}{2(T-1)}\rightarrow N(0,1)$	96.87[a]

Note: [a] represents statistical significance at 1%.
Source: Authors' estimates

asymptotic bias corrected scaled LM test (LM_{BC}). In Table 4, six statistics are estimated to scrutinize the presence of cross-sectional dependence in residuals of estimable model. All are statistically significant at 1% supporting the assumption of cross-sectional dependence in the residuals of estimable model.

Table 5 delves deeper by estimating four statistics, while considering the presence of cross-sectional dependence, in estimable model. All four tests are statistically significant at 1% showing cross-sectional dependence in the variables of estimable model.

Table 5 Tests for Cross-Sectional Dependence in Variables

		Value for			
Test	**Statistic**	$NI_{i,t}$	$ST_{i,t}$	$CP_{i,t}$	$LB_{i,t}$
LM	$\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} T_{ij}\hat{\rho}_{ij}^2 \rightarrow \chi^2 \frac{N(N-1)}{2}$	8726.90[a]	3099.25[a]	6654.54[a]	10943.10[a]
LM_s	$\sqrt{\frac{1}{N(N-1)}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(T_{ij}\hat{\rho}_{ij}^2-1) \rightarrow N(0,1)$	329.55[a]	108.81[a]	248.27[a]	416.48[a]
CD_P	$\sqrt{\frac{TN(N-1)}{2}}\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\hat{\rho}_{ij}\right)$	91.75[a]	14.51[a]	79.80[a]	104.58[a]
LM_{BC}	$\sqrt{\frac{1}{N(N-1)}}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(T_{ij}\hat{\rho}_{ij}^2-1)-\frac{N}{2(T-1)} \rightarrow N(0,1)$	329.17[a]	108.43[a]	247.88[a]	416.09[a]

Note: [a] represents statistical significance at 1%.
Source: Authors' estimates

3.5 Stationarity Tests in Presence of Cross-sectional Dependence

Cross-sectional dependence has a strong presence in residuals as tested in Table 4 and Table 5. It calls for checking stationarity using second generation of unit root tests since first generation of unit root tests (Im et al., 2003; Levin et al., 2002) do not account for cross-sectional dependence in testing for stationarity.

Considering the evident cross-sectional dependence, we use second generation unit root tests proposed by Pesaran to shed light on the findings. Mathematically:

$$\Delta y_{i,t} = a_i + b_i y_{i,t-1} + c_i \bar{y}_{t-1} + d_i \Delta \bar{y}_t + \varepsilon_{i,t}\ , \tag{3}$$

where a_i is a deterministic term, $\bar{y}_t$ is the cross-sectional mean at time t and ρ is the lag order. $t_i(N,T)$ denotes the corresponding t-ratio of a_i and is known as cross-sectional ADF [CADF, attributed to Pesaran (2003)]. The average of the t-ratios gives the cross-sectional IPS [CIPS, attributed to Pesaran (2007)]. In Table 6, these tests are estimated with a constant term at level and first difference. Mutual consensus of both, CADF and CIPS tests, reveals that variables are stationary at level and at first difference i.e. *I(0) and I(1).*

3.6 Dynamic Analysis with Cross-sectional Dependence

Dynamic analysis is suitable in case of relationships where current values of the explained variable are inclined by past ones. Growth regressions, such as in this paper, are mostly characterized by a lagged term of explained variable ($NI_{i,t-1}$).

Table 6 Second Generation Unit Root Tests for Individual Variables

Cross-Sectional ADF (CADF) Test						
$NI_{i,t}$	$\Delta NI_{i,t}$	$ST_{i,t}$	$\Delta ST_{i,t}$	$CP_{i,t}$	$LB_{i,t}$	$\Delta LB_{i,t}$
–1.371	–3.300[a]	–2.362[a]	–	–3.352[a]	–0.691	–3.884[a]
Cross-Sectional IPS (CIPS) Test						
$NI_{i,t}$	$\Delta NI_{i,t}$	$ST_{i,t}$	$\Delta ST_{i,t}$	$CP_{i,t}$	$LB_{i,t}$	$\Delta LB_{i,t}$
–1.778	–2.857[a]	–1.858	–3.672[a]	–2.067[c]	–1.007	–2.953[a]
$NI_{i,t}$ **is I(1)**		$ST_{i,t}$ **is I(1)**		$CP_{i,t}$ **is I(0)**	$LB_{i,t}$ **is I(0)**	

Note: By definition: $CIPS = \frac{\sum_{i=1}^{N} t_i(N,T)}{N} = \frac{\sum_{i=1}^{N} CADF_i}{N}$, [a] and [c] represent statistical significance at 1% and 10%.

Source: Authors' estimates

In case of dynamic analysis, presence of CD requires implementation of improved versions of MG approach. In Table 1 and Table 3, CD tests have shown the presence of cross-sectional dependence in POLS, FE, RE, FD and MG estimates, respectively. Therefore, it is logical to deploy estimation techniques that cater cross-sectional dependence. Pesaran (2006) forwarded Common Correlated Effects Mean Group (CCEMG) model with estimator $\beta_j(= \beta + \omega_j)$ which implies a common parameter β across the countries while $\boldsymbol{\omega}_j \sim IID(0, V_\omega)$. CCEMG has the tendency to asymptotically eliminate CD. Moreover, it allows heterogeneous slope coefficients across group members that are captured simply by taking the average of each country's coefficient.

Attributed to Eberhardt and Teal (2010), Augmented Mean Group (AMG) is a surrogate for CCEMG, which also captures the unobserved common effect in the model. Moreover, AMG estimator also measures the group-specific estimator and takes a simple average across the panel. The highlight of AMG is that it follows first difference OLS for pooled data and is augmented with year dummies.

The estimable model can be written as follows:

$$NI_{it} = \alpha_i + c_i t + d_i \hat{\mu}_t^{va\bullet} + \beta_{i,1}(ST_{i,t}) + \beta_{i,2}(CP_{i,t}) + \beta_{i,3}(LB_{i,t}) + \varepsilon_{i,t}\,, \tag{4}$$

where, i stands for cross-sectional dimension $i = 1,\ldots,n$ and time period $t = 1,\ldots,t$ and α_i represents country specific effects and $d_i t$ denotes heterogeneous country specific deterministic trends. α_i is related with the coefficient of respective independent variables $\beta_{i1} = \frac{\alpha_{i1}}{1-\alpha_{i1}}$, $\beta_{i2} = \frac{\alpha_{i2}}{1-\alpha_{i2}}$ and $\beta_{i2} = \frac{\alpha_{i2}}{1-\alpha_{i2}}$ that are considered as heterogeneous across the countries. It is also assumed that the short run dynamics and their adjustment towards long run take place via error term $u_{i,t}(= \acute{\Gamma}_i f_t + \varepsilon_{i,t})$. f_t characterizes the vector of unobserved common shocks. f_t can be either stationary or nonstationary, which does not influence the validity of the estimation (Kapetanios, Pesaran, and Yamagata, 2011). AMG estimation finds an explicit estimate for f_t which renders $\hat{\mu}_t^{va\cdot}$ (common dynamic process) economic meaningfulness. Total factor productivity (TFP) is one of the plausible interpretations of $\hat{\mu}_t^{va\cdot}$. Its coefficient d_i represents the implicit factor loading on common TFP. In addition, the cross-sectional specific errors $\varepsilon_{i,t}$ are permissible to be serially correlated over time and weakly dependent across the countries (Cavalcanti, Mohaddes, and Raissi, 2011). However, the regressors and unobserved common factor have to be identically distributed.

3.6.1 Interpretation

In Table 7, the main variable of concern i.e. steel production shows statistically significant positive relationship using augmented mean group (AMG) as well as under common correlated effects mean group (CCEMG) estimation. CCEMG is estimated with 'without and with country specific trend' assumption.

Whereas AMG is estimated with an additional assumption of 'with and without common dynamic process (CDP)'. This allows for 4 variants of AMG. The significant positive relationship holds true for all variants 6 of CCEMG and AMG in Table 7. AMG being the most sophisticated is to be relied on.

Table 7 Dynamic Analysis with Cross-Sectional Dependence

Estimator	Common Correlated Effects Mean Group		Augment Mean Group			
Dependent variable	$NI_{i,t}$	$NI_{i,t}$	$NI_{i,t}$	$NI_{i,t}$	$NI_{i,t} - \widehat{\mu}_t^{va\bullet}$	$NI_{i,t} - \widehat{\mu}_t^{va\bullet}$
Trend Assumption	WoT	WT	WoT	WT	WoT	WT
$ST_{i,t}$	0.5479[b] (0.234)	0.5626[a] (0.199)	0.3353[a] (0.124)	0.4275[b] (0.174)	0.3964[a] (0.113)	0.3353[b] (0.150)
$CP_{i,t}$	0.1407[c] (0.077)	0.1572[b] (0.073)	0.2205[a] (0.058)	0.1867[b] (0.085)	0.0980 (0.060)	0.2040[b] (0.092)
$LB_{i,t}$	0.3826 (0.692)	0.3270 (0.468)	0.3528[a] (0.043)	0.3130[a] (0.057)	0.4679[a] (0.072)	0.3914[a] (0.047)
CDP	–	–	0.9682[a] (0.099)	0.7367[a] (0.200)	–	–
Country Trend	–	0.0005 (0.025)	–	0.0203 (0.019)	–	-0.00001 (0.010)
Constant	–3.4966[c] (2.040)	–3.2822 (3.627)	–5.4127[a] (1.953)	–7.3500[b] (2.964)	–6.3176[a] (1.806)	–5.6000[b] (2.440)
NST	–	13	–	13	–	21
RMSE	0.1468	0.1235	0.2207	0.1882	0.2439	0.2102
Observations	910	910	910	910	910	910
Groups	26	26	26	26	26	26
CD	3.84[a]	2.90[a]	4.30[a]	5.14[a]	7.38[a]	6.58[a]

Notes: WoT and WT stand for estimation without and with country specific trends. *CDP* is the common dynamic process. In parenthesis, standard errors are given whereas [a], [b] and [c] show statistical significance at 1%, 5% and 10%, respectively. NST stand for Number of Significant Trends. RMSE stands for root mean squared error and uses residuals from group-specific regression.

Source: Authors' estimates

3.7 Robustness Check

In Table 8, twenty-three (23) slopes are estimated using difference estimators and their variants and compared in order to check the robustness of results of hypothesis. These include Pooled Ordinary Least Squares (POLS), Fixed Effects (FE), Fixed Effects with Driscoll & Kraay standard errors (FE-DK), Random Effects (RE), Generalized Least Squares (GLS), First Differenced-Fixed Effects (FD), Pooled-Fully Modified Ordinary Least Squares (P-FMOLS), Weighted Pooled-Fully Modified Ordinary Least Squares (WP-FMOLS), Group Mean-Fully Modified Ordinary Least Squares (GM-FMOLS), Pooled-Dynamic Ordinary Least Squares (P-DOLS), Weighted Pooled- Dynamic Ordinary Least Squares (WP-DOLS), Group Mean-Dynamic Ordinary Least Squares (GM-DOLS), Difference Generalized Method of Moments (DIF-GMM), System Generalized Method of Moments (SYS-GMM), Dynamic Fixed Effects (DFE), Mean Group (MG), Pooled Mean Group (PMG), Common Correlated Effects Mean Group (CCEMG) and Augmented Mean Group (AMG).

CCEMG and AMG are further estimated with and without country specific trends (WoT and WT). In addition, AMG is further estimated without common dynamic process under the assumptions of with and without country specific trends {$(WoT)_{CDP}$ and $(WT)_{CDP}$}. In case of steel production, majority (83%) 19 out of 23 estimators give desirable results in terms of expected sign and statistical significance that adds to the robustness of the Steel production-growth relationship analyzed in this paper. Moreover, AMG – the most sophisticated of estimators – shows desirable results with all of its variants (with and without country specific trends and common dynamic process).

Table 8 Robustness Slope Parameters

Technique		Statistic of Estimator	Value	S.E
POLS		$\beta_{OLS} = \left(\sum_i X_i' X_i\right)^{-1}\left(\sum_i X_i' Y_i\right)$	0.0221[b]	0.011
FE		$\beta_{FE/DK} = \left(\sum_{i=1}^{N} X_i' Q X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i' Q Y_i\right)$	0.0293[a]	0.004
FE-DK			0.0293[a]	0.005
RE		$\beta_{RE/GLS} = \left(\sum_{i=1}^{N} X_i' \Omega_M^{-1} X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i' \Omega_M^{-1} Y_i\right)$	0.0018[c]	0.001
GLS			0.0012[b]	0.001
FD		$\beta_{FD} = (\Delta X' \Delta X)^{-1} \Delta X' \Delta Y$	0.0731	0.049
P-FMOLS		$\beta_{FP} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} X_{it} X_{it}'\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(X_{it}\overline{Y}_{it}^{+} - \lambda_{12}^{+\prime}\right)$	0.1600	0.111
WP-FMOLS		$\beta_{FW} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \overline{X}_{it}^{*} \overline{X}_{it}^{*\prime}\right)^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\left(\overline{X}_{it}^{*}\overline{Y}_{it}^{*} - \lambda_{12i}^{*\prime}\right)$	0.3308[a]	0.015
GM-FMOLS		$\beta_{FG} = \frac{1}{N}\sum_{i=1}^{N}\left\{\left(\sum_{t=1}^{T} \overline{X}_{it}\overline{X}_{it}'\right)^{-1}\sum_{t=1}^{T}\left(\overline{X}_{it}\overline{Y}_{it} - \lambda_{12i}'\right)\right\}$	1.0818[a]	0.215
P-DOLS		$\begin{bmatrix}\beta_{DP}\\ \gamma_{DP}\end{bmatrix} = \left(\sum_{i=1}^{N}\sum_{t=1}^{T} \overline{W}_{it}\overline{W}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\sum_{t=1}^{T} \overline{W}_{it}\overline{Y}_{it}'\right)$	0.1666	0.149
WP-DOLS		$\begin{bmatrix}\beta_{DW}\\ \gamma_{DW}\end{bmatrix} = \left(\sum_{i=1}^{N}\widehat{\omega}_{1.2i}^{-1}\sum_{t=1}^{T} \overline{W}_{it}\overline{W}_{it}'\right)^{-1}\left(\sum_{i=1}^{N}\widehat{\omega}_{1.2i}^{-1}\sum_{t=1}^{T} \overline{W}_{it}\overline{Y}_{it}'\right)$	0.2806[a]	0.099
GM-DOLS		$\begin{bmatrix}\beta_{DG}\\ \gamma_{DG}\end{bmatrix} = \frac{1}{N}\sum_{i=1}^{N}\left\{\left(\sum_{t=1}^{T} \overline{W}_{it}\overline{W}_{it}'\right)^{-1}\sum_{t=1}^{T} \overline{W}_{it}\overline{Y}_{it}'\right\}$	1.3661[a]	0.263
DIF-GMM		$\beta_{GMM} = argmin_{\beta}(\bar{v}'Z)A_N(Z'\bar{v}) = \frac{\bar{y}'_{-1}ZA_NZ'\bar{y}}{\bar{y}'_{-1}ZA_NZ'\bar{y}_{-1}}$	0.0247[a]	0.006
SYS-GMM			0.0198[a]	0.005
DFE		$\beta_{DFE} = \left(\sum_{i=1}^{N} X_{i,t-1}' Q X_{i,t-1}\right)^{-1}\left(\sum_{i=1}^{N} X_{i,t-1}' Q Y_i\right)$	6.3398[a]	1.357
MG		$\beta_{MG} = \frac{1}{N}\sum_{i=1}^{N}\theta_i$	0.029	0.046
PMG		$\beta_{PMG} = -\left(\sum_{i=1}^{N}\frac{\hat{\phi}_i^2}{\hat{\sigma}_i^2} X_i' H_i X_i\right)^{-1}\left\{\sum_{i=1}^{N}\frac{\hat{\phi}_i}{\hat{\sigma}_i^2} X_i' H_i\left(\Delta Y_i - \hat{\phi}_i Y_{i,t-1}\right)\right\}$	1.0571[a]	0.089
CCEMG	**WoT**	$\beta_{CCEMG} = J^{-1}\sum_{i=1}^{J}\hat{\beta}_j$	0.5479[b]	0.234
	WT		0.5626[a]	0.199
AMG	**$(WoT)_{CDP}$**	$\beta_{AMG} = \frac{\alpha_{i1}}{1-\alpha_{i1}}$	0.3353[a]	0.124
	$(WT)_{CDP}$		0.4275[b]	0.174
	WoT		0.3964[a]	0.113
	WT		0.3353[b]	0.15

Notes: WoT and WT show estimates without common dynamic process 'without trend' and 'with trend' argument. $(WoT)_{CDP}$ and $(WT)_{CDP}$ show estimates with explicit common dynamic process 'without trend' and 'with trend' argument. [a], [b] and [c] show statistical significance at 1%, 5%, and 10% respectively. S.E stands for standard error.

Source: Authors' estimates

3.8 Impetus of Relationship

At country level, robustness of the results is also affirmed by estimating country specific slopes ($\theta_i = \frac{\delta_i}{1-\lambda_i}$). Majority of countries show highly significant positive relationship between steel production and national income. Whereas remaining countries either give unexpected sign and/or statistical insignificance.

In similar veins, country specific error correction terms (ECT) are also estimated. Ones listed in the Table 9 fulfill the following conditions:

$$ECT_i < 1, \left|ECT_i\right| > 0 \; and \; (p - value)_{ECT_i} < 0.05. \tag{5}$$

These countries are major contributors to overall statistical long run relationship.

Table 9 Imputes of Relationship

Country Specific Slopes (θ_i)					
Country	θ_i	**S.E**	**Country**	θ_i	**S.E**
Croatia	0.1524[b]	0.060	Latvia	0.1152[a]	0.039
Czech Republic	0.4503[a]	0.102	Lithuania	0.1616[a]	0.041
Estonia	0.2168[a]	0.027	Netherlands	0.5482[b]	0.270
Finland	0.3160[b]	0.147	Romania	1.0623[a]	0.270
Germany	1.0492[a]	0.395	Slovakia	0.2631[a]	0.065
Greece	0.7258[a]	0.168	Slovenia	0.1895[b]	0.076
Italy	1.2976[a]	0.390	Spain	0.8162[a]	0.184
Country Specific Error Correction Terms (ECT_i)					
Country	ECT_i		**Country**	ECT_i	
Austria	–0.0315[a]		Italy	–0.0448[a]	
Belgium	–0.0116[a]		Latvia	–0.0402[a]	
Croatia	–0.0621[a]		Lithuania	–0.0525[a]	
Czech Republic	–0.0633[a]		Luxembourg	–0.0861[a]	
Denmark	–0.0340[a]		Netherlands	–0.0182[a]	
Estonia	–0.0423[a]		Poland	–0.0344[a]	
Finland	–0.0118[a]		Portugal	–0.0296[a]	
France	–0.0196[a]		Romania	–0.0366[a]	
Germany	–0.0219[a]		Slovakia	–0.0240[a]	
Greece	–0.0365[a]		Slovenia	–0.0559[a]	
Hungary	–0.0703[a]		Spain	–0.0153[a]	
Ireland	–0.0167[a]		United Kingdom	–0.0190[a]	

Note: [a] and [b] show statistical significance at 1% and 5%. S.E stands for standard error. ECT_i are the country specific error correction terms.
Source: Authors' estimates

Countries including Croatia, Czech Republic, Estonia, Finland, Germany, Greece, Italy, Latvia, Lithuania, Netherlands, Romania, Slovakia, Slovenia and Spain show both expected significant slope as well as country specific significant ECT. These countries contribute to the overall positive sign and significance of relationship between national income and steel production.

3.9 What Causes What?

3.9.1 Panel Granger Causality Test

Work of Granger (1969) laid the foundation of causality test that uses the bivariate regressions in a panel data context:

$$\begin{aligned} y_{i,t} &= \alpha_{0,i} + \alpha_{1,i}\, y_{i,t-1} + \cdots + \alpha_{p,i}\, y_{i,t-p} + \beta_{1,i}\, x_{i,t-1} + \cdots + \beta_{p,i}\, x_{i,t-p} + \epsilon_{i,t} \\ x_{j,t} &= \alpha_{0,j} + \alpha_{1,j}\, x_{j,t-1} + \cdots + \alpha_{p,j}\, y_{j,t-p} + \beta_{1,j}\, y_{j,t-1} + \cdots + \beta_{p,j}\, y_{j,t-p} + \varepsilon_{j,t} \end{aligned} \tag{6}$$

Depending on the assumptions about homogeneity of the coefficients across cross-sections, there are two forms of panel causality test. First and conventional type treats the panel data as one large stacked set of data and performs the causality test in the standard way, that assumes all coefficients same across all cross-sections.

$$\begin{aligned} &\alpha_{0,i} = \alpha_{0,j}, \alpha_{1,i} = \alpha_{1,j}, \ldots, \alpha_{p,i} = \alpha_{p,i}, \forall_{i,j} \\ &\beta_{1,i} = \beta_{1,j}, \ldots, \beta_{p,i} = \beta_{p,i}, \forall_{i,j} \end{aligned} \tag{7}$$

Results of panel Granger causality are shown in Table 10.

Table 10 Panel Granger Causality Test Results

Causality	F-Statistic	Remarks
$ST_{i,t} \rightarrow NI_{i,t}$	0.1444	Uni-causal Relationship from macroeconomic performance to steel production.
$NI_{i,t} \rightarrow ST_{i,t}$	24.0905[a]	

Note: [a] shows statistical significance at 1%.
Source: Authors' estimates

Uni-causality from national income to steel production is evident from results in Table 10. Under the hypothesis in the Introduction, causal relationship is set for investigation. Siddique, Mehmood & Ilyas (2016) who explain the mechanism of causal linkages from the national income to steel production (demand following view) is showed in Figure 2. 'Demand following view' holds in case of EU since results in Table 10 show evidence of causality from national income to steel production. Due high growth rates of national income the need for innovation, industrialization and mechanization increases. Such raises the demand for steel that causes increased steel production. Same seems to be case of EU countries during the time span under consideration.

Figure 2 Demand Following Hypothesis for Steel Production and National Income

$$\boxed{Higher\ growth\ rates\ of\ national\ income} \rightarrow$$
$$\boxed{Need\ for\ developmental\ projects\ (innovation, industrialization\ and\ mechanization)} \rightarrow$$
$$\boxed{Increased\ demand\ for\ steel} \rightarrow \boxed{Increased\ production\ in\ Steel\ Industry}$$

Source: Authors' formulation

3.9.2 Rationale for Dumitrescu-Hurlin Causality

However, one of the main issues specific to panel data models refers to the specification of the heterogeneity between cross-sections. To consider the heterogeneity across cross-sections, Dumitrescu-Hurlin (2012) made an assumption of allowing all coefficients to be different across cross-sections. In this causality context, the heterogeneity can be between the heterogeneity of the regression model and/or in terms of causal relationship from x to y. Indeed, the model considered may be different from an individual to another, whereas there is a causal relationship from x to y for all individuals. The simplest form of regression model heterogeneity takes the form of slope parameters' heterogeneity. More precisely, in a '*p*' order linear vectorial autoregressive model, four kinds of causal relationships are defined. Under the Homogeneous Non-Causality (HNC) hypothesis, no individual causality from x to y occurs. On the contrary, in the Homogeneous Causality (HC) and Heterogeneous Causality (HEC) cases, there is a causality relationship for each individual of the sample. To be more precise, in the Homogeneous Causality (HC) case, the same regression model is valid (identical parameters' estimators) for all individuals, whereas this is not the case for the HEC hypothesis. Finally, under the Heterogeneous Non-Causality (HENC) hypothesis, the causality relationship is heterogeneous since the variable x causes y only for a subgroup of $N–N_1$ units.

Authors based their version of causality test on the Granger (1969) and extended to non-causality test for heterogeneous panel data models with fixed coefficients.

Considering linear model:

$$y_{i,t} = \alpha_i + \sum_{k=1}^{K} \gamma_i^{(k)} y_{i,t-k} + \sum_{k=1}^{K} \beta_i^{(k)} x_{i,t-k} + \varepsilon_{i,t} \qquad i = 1, 2, \ldots, N;\ t = 1, 2, \ldots, T, \tag{8}$$

where x and y are two stationary variables observed for N individuals in T periods. $\beta_i = (\beta_i^{(1)}, \ldots, \beta_i^{(K)})'$ and the individual effects α_i are assumed to be fixed in the time dimension. It is assumed that there are lag orders of K identical for all cross-section units of the panel. Moreover, autoregressive parameters $\gamma_i^{(k)}$ and

the regression coefficients $\beta_i^{(k)}$ are allowed to vary across groups. Under the null hypothesis, it is assumed that there is no causality relationship for any of the units of the panel. This assumption is called the Homogeneous Non-Causality (HNC) hypothesis, which is defined as:

$$\mathrm{H}_0: \beta_i = 0 \; \forall i = 1, \ldots, N. \tag{9}$$

The alternative is specified as the Heterogeneous Non-Causality (HENC) hypothesis. Under this hypothesis, two subgroups of cross-section units are allowed. There is a causality relationship from x to y for the first one, but it is not necessarily based on the same regression model. For the second subgroup, there is no causality relationship from x to y. A heterogeneous panel data model with fixed coefficients (in time) in this group is considered. This alternative hypothesis is expressed as follows:

$$\begin{aligned} &\mathrm{H}_1: \beta_i = 0 \; \forall i = 1, \ldots, N_1, \\ &\beta_i \neq 0 \; \forall i = N_1 + 1, \ldots, N. \end{aligned} \tag{10}$$

It is assumed that β_i may vary across groups and there are $N_1 < N$ individual processes with no causality from x to y. N_1 is unknown but it provides the condition $0 \leq N_1/\mathrm{N} < 1$.

The average statistic $W_{N,T}^{HNC}$, which is related with the null Homogeneous non-causality (HNC) hypothesis are proposed:

$$W_{N,T}^{HNC} = \frac{1}{N}\sum_{i=1}^{N} W_{i,T}\,, \tag{11}$$

where $W_{i,t}$ indicates the individual Wald statistics for the i[th] cross-section unit corresponding to the individual test $\mathrm{H}_0: \beta_i = 0$.

Let $Z_i = [e : Y_i : X_i]$ be the *(T, 2K+1)* matrix, where *e* indicates *a (T, 1)* unit vector and $Y_i = Y_i = \left[y_i^{(1)} : y_i^{(2)} : \ldots : y_i^{(K)}\right]$, $X_i = \left[x_i^{(1)} : x_i^{(2)} : \ldots : x_i^{(K)}\right]$. $\theta_i = (\alpha_i \gamma_i' \beta_i')$ is the vector of parameters of the model. Also let $R = [0: I_K]$ be a *(K, 2K+1)* matrix.

For each $i = 1, \ldots, N$, the Wald statistic $W_{i,t}$ corresponding to the individual test $\mathrm{H}_0: \beta_i = 0$ is defined as:

$$W_{i,T} = \hat{\theta}_i' R' [\hat{\sigma}_i^2 R (Z_i' Z_i)^{-1} R']^{-1} R \hat{\theta}_i\,. \tag{12}$$

Under the null hypothesis of non-causality, each individual Wald statistic converges to a chi-squared distribution with K degrees of freedom for $T \to \infty$.

$$W_{i,T} \to \chi^2(K), \forall i = 1, \ldots, N. \tag{13}$$

The standardized test statistic $W_{N,T}^{HNC}$ for $T, N \to \infty$ is as follows:

$$Z_{N,T}^{HNC} = \sqrt{\frac{N}{2K}}\left(W_{N,T}^{HNC} - K\right) \to N(0,1). \tag{14}$$

Also, the standardized test statistic $\tilde{Z}_N^{HNC}$ for fixed T samples is as follows:

$$\tilde{Z}_N^{HNC} = \sqrt{\frac{N}{2K} \times \frac{(T-2K-5)}{(T-K-3)} \times \left[\frac{(T-2K-5)}{(T-K-3)} W_{N,T}^{HNC} - K\right]} \to N(0,1), \tag{15}$$

where $W_{N,T}^{HNC} = \left(\frac{1}{N}\right)\sum_{i=1}^{N} W_{i,T}$.

In addition to presence of heterogeneity among cross-sections, if cross-sectional dependence exists in panel, Dumitrescu-Hurlin causality is suitable. Results of CD tests in Table 1, Table 3, Table 4 and Table 5 show the presence of cross-sectional dependence. Whereas, stationarity is a basic requirement

of Dumitrescu-Hurlin causality test. Second generation unit root test named as Pesaran's CADF (2003) and CIPS (2007) statistic fulfills the objective of checking for stationarity in presence of cross-sectional dependence. Therefore, Dumitrescu-Hurlin causality test should be applied. Its results are as follows:

Table 11 Dumitrescu-Hurlin Causality Test Results

Causality	$W_{N,T}^{HNC}$	$\widetilde{Z}_{N}^{HNC}$	p-value	Remarks
$ST_{i,t} \rightarrow NI_{i,t}$	2.6132	1.0068	0.314	Homogeneous Uni-causal relationship from macroeconomic performance to steel production.
$NI_{i,t} \rightarrow ST_{i,t}$	6.0121	8.4564	0.000	

Source: Authors' estimates

Table 11 represents statistical significance of first $\tilde{Z}_N^{HNC}$ test statistic showing that null hypothesis cannot be rejected that $ST_{i,t}$ do not homogeneously cause $NI_{i,t}$, whereas it gets rejected in reverse causality. It implies that the causality is homogeneous from national income to steel production. This specialized form of causality provides the insights into the causal relationship without contradicting the primary result of bi-causal Granger causality in Table 10. Homogenous causality can be attributed to 'uniform growth effects' of economic growth on steel industries in economies that are 'integrated' in a union known as European Union.

CONCLUSION

European Union was chosen for investigating relationship between steel production and national income. Using sophisticated econometric techniques, the relationship is found to be robust. The causality gives support to 'Demand Following Hypothesis'. Feedback effect of steel industry on national income can amplify the macroeconomic contribution of steel production. However, it is missing or too weak at this stage. Firm level studies can help in understanding the microeconomic foundations of causal linkage from steel production to national income. Such firm/industry specific studies are suggested for future. Role of substitute metals e.g. aluminum can also be investigated in terms of their macroeconomic contribution. In addition, to spur efficiency, state may increase the incentive and proportion of private sector in steel industry. Moreover, it may also re-allocate subsidies for steel industry and infrastructure sector. For reducing monopoly power, pricing policy can be effective.

References

BALTAGI, B. H., FENG, Q. AND KAO, C. A Lagrange Multiplier test for cross-sectional dependence in a fixed effects panel data model. *Journal of Econometrics*, 2012, 170(1), pp. 164–177.

BREUSCH, T. S., AND PAGAN, A. R. The Lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies*, 1980, 47(1), pp. 239–253.

CAVALCANTI, T. V. D. V., MOHADDES, K. AND RAISSI, M. Growth, development and natural resources: New evidence using a heterogeneous panel analysis. *The Quarterly Review of Economics and Finance*, 2011, 51(4), pp. 305–318.

DUMITRESCU, E.-I. AND HURLIN, C. Testing for Granger non-causality in heterogeneous panels, *Economic Modeling*, 2012, 29, pp. 1450–1460.

EBERHARDT, M. AND TEAL, F. *Productivity analysis in global manufacturing production.* University of Oxford, Department of Economics, Economics Series Working Papers, 2010, p. 515.

EVANS, M. Steel consumption and economic activity in the UK: The integration and cointegration debate. *Resources Policy*, 2011, 36(2), pp. 97–106.

FREES, E. W. Assessing cross-sectional correlation in panel data. *Journal of Econometrics*, 1995, 69, pp. 393–414.

FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 1937, 32, pp. 675–701.

GRANGER, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 1969, pp. 424–438.

HOECHLE, D. Robust standard errors for panel regressions with cross-sectional dependence. *Stata Journal*, 2007, 7(3), p. 281.
HUH, K.-S. Steel consumption and economic growth in Korea: Long-term and short-term evidence. *Resources Policy*, 2011, 36(2), pp. 107–113.
IM, K. S., PESARAN, M. H. AND SHIN, Y. Testing for unit roots in heterogeneous panels. *Journal of Econometrics*, 2003, 115(1), pp. 53–74.
JEFFERSON, G. H. China's iron and steel industry: Sources of enterprise efficiency and the impact of reform. *Journal of Development Economics*, 1990, 33(2), pp. 329–355.
KAPETANIOS, G., PESARAN, M. H. AND YAMAGATA, T. Panels with non-stationary multifactor error structures. *Journal of Econometrics*, 2011, 160(2), pp. 326–348.
LABSON, B. S. AND CROMPTON, P. L. Common trends in economic activity and metals demand: Cointegration and the intensity of use debate. *Journal of Environmental Economics and Management*, 1993, 25(2), pp. 147–161.
LEVIN, A., LIN, C.-F. AND CHU, C.-S. J. Unit root tests in panel data: asymptotic and finite-sample properties. *Journal of Econometrics*, 2002, 108(1), pp. 1–24.
OZKAN, F. Steel industry and the sector's impact on economic growth in Turkey. *Regional and Sectoral Economic Studies*, 2001, 11(2).
PESARAN, M. H. A simple panel unit root test in the presence of cross section dependence. *Cambridge Working Papers in Economics 0346*, Faculty of Economics (DAE), University of Cambridge, 2003.
PESARAN, M. H. General diagnostic tests for cross section dependence in panels. *Cambridge Working Papers in Economics 0435*, Faculty of Economics, University of Cambridge, 2004.
PESARAN, M. H. Estimation and inference in large heterogeneous panels with a multifactor error structure. *Econometrica*, 2006, 74(4), pp. 967–1012.
PESARAN, M. H., SHIN, Y. AND SMITH, R. P. *Pooled estimation of long-run relationships in dynamic heterogeneous panels*. University of Cambridge, Department of Applied Economics, 1997.
PESARAN, M. H., SHIN, Y. AND SMITH, R. P. Pooled mean group estimation of dynamic heterogeneous panels. *Journal of the American Statistical Association*, 1999, 94(446), pp. 621–634.
PESARAN, M. H. A simple panel unit root test in the presence of cross-section dependence. *Journal of Applied Econometrics, 2007,* 22(2), pp. 265–312.
PESARAN, M. H. AND SMITH, R. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of Econometrics*, 1995, 68(1), pp. 79–113.
SIDDIQUE, F., MEHMOOD, B. AND ILYAS, S. Steel production and macroeconomic performance in Pakistan: What does long-run data show? *Science International*, 2016, 28(4), pp. 95–97.

Estimating the Economic Returns to Schooling: Restricted Maximum Likelihood Approach

Adelaide Agyeman [1] | *Crops Research Institute, Kumasi, Ghana*
Nicholas Nsowah-Nuamah [2] | *Kumasi Polytechnic, Kumasi, Ghana*

Abstract

The economic returns to schooling is a fundamental parameter of interest in many different areas of economics and public policy. The most common technique for estimating this parameter is based on the assumption that the 'true' coefficient of education in the earnings equation is constant across individuals. However, this may not often be wholly true and returns to schooling estimates may be biased and inconsistent. The objective of this study was to estimate the returns to schooling as a random coefficient and obtain accurate and reliable estimates that will be useful for policy recommendations. The restricted maximum likelihood (REML) method was used to estimate the parameters of a random coefficient model using data from a 2007/2008 Ghanaian twins' survey. The results revealed that the REML economic returns to schooling in three selected cities were between 7% and 9%. Significant (p<0.05) variances around the mean returns to schooling implied that returns to schooling might vary among individuals due to unobserved factors.

Keywords

Return to schooling, Random coefficient model, Maximum Likelihood, REML, Variance Heterogeneity

JEL code

I26, J31, C1

INTRODUCTION

The relationship between schooling and earnings is of key importance to the research community and policymakers in both the developed and developing countries. This is because studies have consistently confirmed that people with higher level of education earn more money, experience less unemployment, and work in more prestigious occupations than their less-educated counterparts (Card, 1999; Patrinos, 2006). An important parameter of interest frequently estimated in the schooling-earnings relationship is the economic returns to schooling. It is an indicator of schooling impact on levels of output per worker and a determinant of relative wages (Kaboski, 2007). In addition, studies of returns to schooling along

[1] Biometric Unit, Crops Research Institute, P. O. Box 3785, Kumasi, Ghana. Corresponding author: e-mail: nesiagyeman@yahoo.com, phone: (+233)322060389.
[2] Rector, Kumasi Polytechnic, P. O. Box 854, Kumasi, Ghana. E-mail: n3n_nuamah@yahoo.com.

with other research act as a guide for public policy decisions about the organization and financing of education reforms (Psacharopolous and Patrinos, 2004). In the empirical literature, the standard approach used in estimating the economic returns to schooling is the ordinary least squares (OLS) method on a simple Mincer type earnings function.

Two major issues associated with the estimation of the returns to schooling using the OLS method have been pointed out by (Card, 1999; Pfeiffer and Pohlmeier, 2012) and others. Firstly, an assumption made in most empirical studies when estimating the standard Mincerian wage equation states that the return to schooling is homogenous, (i.e., constant across individuals) making the OLS returns to schooling a fixed coefficient (i.e., a single parameter in the population). However, the return of an additional year of schooling may vary across schooling levels and across individuals of the same schooling level due to differences in observable factors (e.g. family background, school quality, level of schooling, etc.) as well as unobservable factors (e.g. cognitive and non-cognitive skills, peer group and network effects), Pfeiffer and Pohlmeier (2012). In such a situation, it may be better to regard the returns to schooling as a random coefficient subject to random variation (Hildreth and Houck, 1968). If this random coefficient is correlated with the schooling variable or the additive error term in the earnings equation, then standard OLS estimates of returns to schooling will be biased and inconsistent. Secondly, in the presence of nested and hierarchically structured data, such as individuals or twins within families, OLS techniques violate the assumption of independence of errors leading to imprecise parameter estimates and loss of statistical power, and subsequently increases the likelihood of rejecting a true null hypothesis (Raudenbush and Bryk, 2002).

Consequently, given these limitations an OLS estimation of schooling on earnings will fail to accurately identify the schooling earnings relationship and its usefulness with respect to policy recommendations will be limited. A number of economists have used the instrumental variable (IV) approach (Heckman, 1998) to address the inefficiency of OLS when returns to schooling vary across individuals. However, as noted by (Card, 2001) even the IV technique based on ideal instruments will produce estimates that are weighted averages of the returns to schooling for each individual with higher weight placed on those individuals most likely to have been affected by the instrument of choice. As a result, the IV will be a biased estimate of both the average return to schooling and the return to schooling of the group affected by the instrument if returns to schooling varies across individuals. They both concluded that in several instances the IV estimates are not precise and cannot effectively estimate policy relevant parameters.

The dominant approach to the random coefficient model estimate in recent years is based on the principle of maximum likelihood (ML) estimation (Bickel, 2007). The reason being that when the assumptions of independence of observations and residuals are violated as in the case of varying parameter estimates, maximum likelihood estimators provide parameter estimates that are relatively consistent, asymptotically normal and efficient (Card, 2001). However, the ML estimator of variance components in a linear model can be biased downwards because it does not adjust for the degrees of freedom lost by estimating the fixed regression coefficients. Patterson & Thompson (1971) introduced the restricted maximum likelihood (REML) estimator to address the limitations of the ML. REML in contrast to ML, adjusts for the degrees of freedom lost due to the estimation of the fixed effects parameters by maximizing the likelihood of linearly independent residual error contrasts to obtain unbiased estimates (Laird and Ware, 1982; Lindstrom and Bates, 1988). REML provides unbiased regression coefficients even with small samples by considering the number of parameters used in model estimation (Nunnally and Bernstein, 1994). Consistent with this trend, Ashenfelter and Krueger (1994) identified an income premium related to higher educational attainment by using data from an Ohio Twinsburg survey. Their REML returns to schooling estimate was about 16%. Mazumder (2004) also analyzed data from the 1979 National Longitudinal Survey (NLSY79) in the United States using the restricted maximum likelihood (REML) method.

His findings indicate that more than half the variation in the log of wages among men is due to differences in family and community background. Sadeq (2014) investigated differences in wage penalty between formal and informal employment using labor force survey data from three countries. His REML rate of return to required years of education for formal employees ranged from 7.8% to 8.4%. Likewise, Anger and Schnitzlein (2013) analyzed data from the German Socio-Economic Panel Study (SOEP), using a Restricted Maximum Likelihood (REML) model. They find substantial influence of family background on the skills of both brothers and sisters. Their sibling correlations of the personality traits range from 0.24 to 0.59 indicating that even for the lowest estimate, one fourth of the variance or inequality can be attributed to factors shared by siblings. Sibling correlations in cognitive skills were also higher than 0.50, indicating that more than half of the inequality in earnings could be explained by family characteristics.

The objectives of this paper are to (a) to estimate the return to schooling as a random regression coefficient, (b) to determine the influence of individual and family background characteristics on the returns to schooling and to (c) to decompose the variance around the mean return into family heterogeneity, individual heterogeneity and residual error.

1 MATERIALS AND METHODS

1.1 Data

There is no national twins database in Ghana and therefore primary data was collected by a team of five interviewers during a twins' survey in December 2007 and January 2008 in three cities in Ghana, namely Accra, Kumasi and Takoradi. Questionnaires were administered through face-to-face personal interviews to gainfully employed adult twins aged between 18 and 65. Twins were identified through various channels including twins registered at the twin's clubs, various work places, markets, shops, colleagues, friends, relatives, and households. In Kumasi 404 respondents were identified, whereas in Accra and Takoradi the total of 96 respondents were identified. Altogether, 500 respondents were identified. 50% of twins identified were randomly selected and interviewed giving a total of 250 respondents made up of 125 twin pairs. Out of the 250 respondents, 144 individuals were dizygotic (DZ) twins and 106 were monozygotic (MZ) twins. This data set provides a unique and rich source of information on the socio-economic characteristics (age, gender, marital status, earnings, education, family background characteristics such as sibling education, father's and mother's education etc.) of twins' in Ghana. Data analysis was performed using three samples (Pooled, Monozygotic and Dizygotic) in order to identify the comparative roles of genetics and family background as mediating influences in the returns to schooling.

1.2 Modeling Framework

The modeling technique used for estimating the return to schooling as a random coefficient was the hierarchical linear Model (HLM) by Raudenbush and Bryk, (2002). The multilevel characteristic of HLM captured the inherently hierarchical nature of the family-twins dataset (i.e. individuals/twins observations (level 1) nested within families (level 2)). The mean effect of education on earnings and the variance in returns around this mean were represented as fixed and random effects respectively. Observable differences in returns across individuals were controlled by the influence of siblings and family background characteristics (e.g. parental education) on earnings. Family-specific random returns were also estimated as deviations around the sample average return to schooling. An individual-specific random intercept was also introduced to control the unobserved heterogeneity which is usually interpreted as the return to an individual's innate ability or skill. The proportion of the total variation in earnings that lies "between" individuals in terms of an intra-class correlation (ICC or ρ) was also calculated to describe how strongly twins in the same family resemble each other.

As a first step in the HLM analysis of the returns to schooling, the ICC was determined using the unconditional or null model. The null model (contains no explanatory variables) expresses

the individual-level earnings Y_{ij} for the i^{th} sibling/twin in the j^{th} family (i = 1, 2; j = 1, 2, k) by combining two linked models: one at the individual level (level 1) and another at the family level (level 2) as:

Level-1 (sibling/twins-level) model is:

$$Y_{ij} = \beta_{0j} + e_{ij}, \text{ where } e_{ij} \sim N(0, \sigma_e^2). \tag{1.1}$$

Level-2 (family-level) model is:

$$\beta_{0j} = \beta_0 + \mu_{0j}, \text{ where } \mu_{0j} \sim N(0, \sigma_\mu^2). \tag{1.2}$$

The level-1 and level-2 equations are combined into a single model equation and represented as:

$$Y_{ij} = \beta_o + \mu_{0j} + e_{ij}, \text{ where } \mu_{0j} \sim N(0, \sigma_\mu^2), e_{ij} \sim N(0, \sigma_e^2), \tag{1.3}$$

$$Var(\mu_{0j} + e_{ij}) = \sigma_{\mu 0}^2 + \sigma_e^2,\ Cov(\mu_{0j}, e_{ij}) = 0,$$

where Y_{ij} refers to earnings for the i^{th} sibling/twin in the j^{th} family, β_0 is the overall mean, μ_{0j} is the random effect for the j^{th} family and e_{ij} is an individual-specific random error component with population variance σ_e^2. The intra-class correlation ρ is then specified as:

$$\rho = \frac{\sigma_{\mu 0}^2}{\sigma_{\mu 0}^2 + \sigma_e^2}, \tag{2}$$

where $\sigma_{\mu 0}^2$ captures the variance in annual earnings that is due to differences between families while the σ_e^2 captures the variance in annual earnings within families.

Secondly, a two-level hierarchical linear model which involves the estimation of fixed effects, random returns to schooling coefficients, the variance components and individual and family variables to explain differences in returns to schooling across individuals can be written as:

Level 1:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}, \text{ where } e_{ij} \sim N\left(0, \sigma_e^2\right), \tag{3.1}$$

Level 2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + \mu_{0j}, \tag{3.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + \mu_{1j}, \tag{3.3}$$

where (γ_{00} and γ_{10}) are the intercepts or overall means for (β_{0j} and β_{1j}) from the second-level models, (γ_{01} and γ_{11}) are the regression coefficients (slopes) from the second-level models, (μ_{0j} and μ_{1j}) are the random effects or residuals for (β_{0j} and β_{1j}), X and Z are matrices containing explanatory variables. X represents an explanatory variable for individual (twin) i nested in level 2 (family) unit j, and Z represents an explanatory variable for level 2 (family) unit j.

Substituting Equations (3.2) and (3.3) into Equation (3.1) gives the combined model as:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j (\mu_{0j} + \mu_{1j} X_{ij} + e_{ij}), \tag{3.4}$$

where $e_{ij} \sim N(0, \sigma_e^2)$ and $\begin{pmatrix} \mu_{0j} \\ \mu_{1j} \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right] = \mathrm{T},$

where $\mathrm{var}(\mu_{0j}) = \tau_{00}$, $\mathrm{var}(\mu_{1j}) = \tau_{11}$ and $\mathrm{cov}(\mu_{0j}, \mu_{1j}) = \tau_{01}$.

Y_{ij} is a function of the mean intercept (γ_{00}), the regression coefficient or mean slope (γ_{10}) for the first explanatory variable (e.g., education) at level 1, plus two random parameters (variation of the intercepts (μ_{0j}) and variation of the slopes (μ_{1j}) and the residual variation (e_{ij}). X_{ij} and Z_j are matrices containing individual and family level variables (e.g., education, age, marital status, parental education, etc.).

In the two-level HLM, the ($\gamma's$) are the fixed effects parameter estimates that are assumed to be constant across individuals from Equations (3.2) and (3.3), β_{0j} and β_{1j} are the random effects parameter estimates that vary across individuals from Equation (3.1). (e_{ij}) is the variance of the first-level residuals from Equation (3.1) and (μ_{0j} and μ_{1j}) are the variances of the second-level residuals.

The variance around the mean returns to schooling is decomposed into three components as:

$$Var(\mu_{0j} + \mu_{1j} + e_{ij}) = \sigma_{u0}^2 + \sigma_{\mu 1}^2 + \sigma_e^2,\ Cov(\mu_{0j}, e_{ij}) = 0,\ Cov(\mu_{1j}, e_{ij}) = 0,$$

where μ_{0j} is family heterogeneity (i.e., variance component common to all siblings in family j), μ_{1j} is individual or sibling heterogeneity (i.e., variance component unique to individual i in family j) and e_{ij} represents residual error due to measurement errors and other transient errors which are associated with earnings and age-related earnings differences.

1.3 Parameter Estimation

The two-level hierarchical model involves the estimation of three types of parameters, namely the fixed effects, random effects or random coefficients and the variance-covariance components. The restricted maximum likelihood (REML) estimator (Patterson and Thompson, 1971) was used to estimate the parameters. With the REML method only the variance components are included in the likelihood function and the regression coefficients are estimated in a second estimation step. The fixed effects are represented by (γ_{00}, γ_{01}, γ_{10} and γ_{11}) in Equation (3.4) and were estimated by the generalized Least Squares (GLS), Laird and Ware, (1982) given variance-covariance estimates calculated by the REML method (Raudenbush, Bryk, Cheong and Congdon, 2001, p.7). The random coefficients are represented by (β_{0j} and β_{1j}) in Equations (3.2) and (3.3) and were estimated by the empirical Bayes approach or the best linear unbiased prediction (BLUP) method. The variance-covariance components were estimated by REML method and they include (1) the covariance between level-2 error terms (i.e., cov(μ_{0j}, $\mu_{1j} = \tau_{01}$), (2) the variance in the level-1 error term (i.e.,var(e_{ij}) = σ_e^2) and (3) the variance in the level-2 error terms (i.e., var(μ_{0j}, μ_{1j}) = τ_{00} and τ_{11}, respectively). The three parameters in the HLM were estimated using the Statistical Analysis System (SAS) model notation of the two-level HLM in Equation (3.4) specified as follows:

$$Y_j = A_j\gamma + X_j\mu_j + e_j, \quad j=1, 2, J, \tag{4}$$

where $A_j = X_jZ_j$, A_j and X_j, and are known design matrices, Z_j is the level 2 covariate, γ is a vector of fixed effects, μ_j is a vector of random effects and e_j is a vector of random errors. The random effects and the random errors are normally distributed with:

$$e_j \sim N(0, R_j),\ R_j = \sigma^2 I_{nj},\ \mu_j \sim N(0, G),\ G = \begin{bmatrix} \tau_{00} & \tau_{10} \\ \tau_{01} & \tau_{11} \end{bmatrix}.$$

The fixed effects (γ's) were estimated using the GLS. The GLS estimator which provides weighted estimates of the second-level regression coefficients can be written as:

$$\hat{\gamma} = \left(A'\hat{V}^{-1}A\right)^{-1}\left(A'V^{-1}Y\right), \text{ where } V = \text{var}(Y) = XGX' + R\,. \tag{5.1}$$

The variance of $\hat{\gamma}$ is given as:

$$\text{var}\left(\hat{\gamma}\right) = (A'\hat{V}^{-1}A)^{-1}\,. \tag{5.2}$$

The random effects (μ's) were estimated using shrinkage estimators, namely the empirical Bayes method or the best linear unbiased prediction (BLUP) according to the equation below:

$$\hat{\mu} = \hat{G}X'\hat{V}^{-1}\left(Y - A\hat{\gamma}\right). \tag{5.3}$$

The variance-covariance components (σ_e^2, τ_{00} τ_{01} and τ_{11}) were estimated using the restricted maximum likelihood (REML) method. REML estimates of the variance-covariance components (G and R) were calculated by maximizing the REML log-likelihood function:

$$l_{REML}(G,R) = -\frac{1}{2}\log|V| - \frac{1}{2}\log\left|A'V^{-1}A\right| - \frac{N-p}{2}\log r'V^{-1}r - \frac{(N-p)}{2}\left[1 + \log\frac{2\pi}{(N-p)}\right], \tag{6}$$

where: $r = \text{Y} - A\left(A'V^{-1}A\right)' A'V^{-1}\text{Y}$ and $p = rank(A)$.

The maximization was carried out using a ridge-stabilized Newton-Raphson algorithm (Lindstrom and Bates, 1988). Tests of hypotheses about the fixed and random effects and the variance-covariance components were carried out using an approximate t-statistics, Wald Z test and chi-square statistics (Polit, 1996; Agresti, 1990; Verbeke and Molenbergs, 2000). Statistical analyses were conducted using SAS Version 9.1.3 PROC MIXED with REML option.

2 RESULTS

Overall, female twins slightly outnumbered male twins by about 2.4% (Table 1). MZ twins earned more on average than DZ twins and fathers acquired more education than mothers (Table 2).

Table 1 Total Number of Monozygotic (MZ) and Dizygotic (DZ) Twin Respondents in the Three Survey Areas

Area	MZ		DZ		Total
	Male	Female	Male	Female	
Kumasi	42	36	60	64	202
Takoradi	4	2	4	6	16
Accra	8	14	4	6	32
Total	54	52	68	76	250

Source: GTS authors' calculation

Table 2 Descriptive Statistics – Means and Standard Errors

Variable	Pooled sample	Monozygotic twins	Dizygotic twins
Own education (years)	12.576	14.009	11.521
	(0.343)	(0.535)	(0.427)
Co-twins education (years)	12.692	13.840	11.847
	(0.345)	(0.550)	(0.429)
Male (proportion)	0.488	0.509	0.472
	(0.032	(0.049)	(0.042)
Age (years)	32.816	31.887	33.500
	(0.649)	(0.905)	(0.907)
Married (proportion)	0.432	0.321	0.514
	(0.031)	(0.046)	(0.042)
Mother's education	5.776	6.189	5.472
	(0.408)	(0.638)	(0.530)
Father's education	8.288	9.557	7.354
	(0.462)	(0.723)	(0.591)
Log of annual income	GH¢7.184	GH¢7.368	GH¢7.049
	(0.054)	(0.084)	(0.068)
Sample size	250	106	144

Note: Standard errors in parentheses below means.
Source: GTS authors' calculation

2.1 The null model or unconditional model

Table 3 represents the parameter estimates and standard errors for the null model of Equation (1.3). Results of this model reveal that the fixed effects intercept terms are approximately 7.18, 7.37 and 7.05 for pooled, MZ and DZ twins, respectively. The variance of the twins-level residual errors denoted by σ_e^2 is estimated as 0.1544. Likewise, the variance of the family-level residual effect denoted by σ_μ^2 is estimated as 0.5714. All the parameter estimates are positive and the Wald Z-test indicates that they are also significant. The proportion of variance (i.e., the intra-class correlation coefficient (ICC)) in annual earnings that occurs between families for the pooled sample of twins is calculated as $p = 0.5714/(0.5714 + 0.1544) = 0.787$. This estimate which is very high tells us that about 80% of the total variation in earnings of twins can be accounted for by family background effect. Moreover, the ICC estimates (0.88 and 0.70) for MZ and DZ twins respectively (Table 3) indicate that about 12% and 30% of the variances in the two models are attributable to individual traits of MZ twins and DZ twins, respectively. These estimates show the extent to which observations are related within each family and therefore suggest that MZ twins are more closely genetically related than DZ twins. Overall, the correlations describe the proportion of variance associated with differences between families and indicate that family background effects contribute a sizable percentage of the variation in the returns to schooling for twins than individual effects.

Furthermore, the results of the ICC (which are greater than 10% of the total variance in the model) indicate that the HLM is an appropriate model for the estimation of the regression relationship that varies by family using multiple level data (siblings/twins nested within families, Table 3). The residual variance for all three samples are significant ($p<0.01$) and therefore supports the alternative hypothesis that average annual earnings may vary across individuals or twins with the same level of schooling.

Table 3 Results from the Null Hierarchical Linear Model (HLM)

Fixed Effects	Pooled	MZ twins	DZ twins
Family intercept, γ_{00}	7.1846**	7.3686**	7.0492**
	(0.0720)	(0.1158)	(0.0889)
Random Effects	Variance components		
Family mean, τ_{00}	0.5714**	0.6639**	0.4687**
	(0.0830)	(0.1396)	(0.0969)
Residual effect, σ_ε^2	0.1544**	0.0925**	0.2000**
	(0.0195)	(0.0180)	(0.0333)
ICC, ρ	0.7873	0.8777	0.7009
Model Fit			
–2 Res log likelihood	510.9	194.7	304.1
AICc	514.9	198.7	308.1
N	250	106	144

Note: * = $p < .05$, ** = $p < .01$. Standard errors in parentheses below means.
Source: GTS authors' calculation

The results of a second model which includes some demographic characteristics such as number of years spent schooling, age, gender, marital status, father's education and mother's education as explanatory variables are presented in Table 4. Average annual earnings (5.68, 5.02 and 5.88) for the pooled, MZ and DZ twins' samples respectively, are highly significant (p<0.01), suggesting that the effects of education on earnings vary from one family to another and among individuals. The findings also indicate that data used is dominated by a hierarchical structure, which may affect both the intercepts and the slopes (returns to education) of earnings functions. The results further indicate that expected earnings for the three data sets were similar irrespective of the type of model (null or random coefficient) used. However, the expected earnings of the different groups were higher for the null model compared to the second model. Apparently, accounting for the variation in sibling earnings by including demographic variables decreases expected earnings (intercepts) by about 1.5 and 1.2 points for MZ and DZ twins respectively when compared to the expected earnings of the null model which did not have any covariates. This suggests that demographic characteristics explain a proportion of the variation in annual earnings. The effect of an additional year spent schooling on individual earnings ranged from 7% to 9% for the three data samples (Table 4) and it differed significantly from zero (i.e., p<0.01). Returns to schooling estimates for MZ twins were lower than that of both the pooled and DZ twins. This may indicate the existence of some upward bias for MZ twins REML estimates due to omitted unobserved characteristics and also confirms the fact that failure to take account of unobserved heterogeneity leads to biased estimates on the returns to schooling. It may also suggest that high-ability MZ twins find it easier to acquire more education. Father's education significantly (p<0.05) affected REML returns to schooling for MZ twins, whiles mother's education had a significant impact on REML returns to schooling for DZ twins. The returns to schooling estimates (Pooled = –0.15, MZ = –0.10 and DZ = –0.16) for gender measured by the dummy male were negative for all three samples and significant at the 5% level for the pooled and DZ twins' samples. This inverse relationship implies negative average returns to education for male twins and suggests that an additional year of schooling has a higher pay-off for females than for males. This means that while females have lower wage levels than men, they have higher average returns to education. The effect of age on earnings for every additional life year was significant (p<0.05) for MZ twins but insignificant (p>0.05) for DZ twins. This finding may be associated with age being a better proxy for actual work

experience for MZ twins than it is for DZ twins. Moreover, the MZ twins sample are on average younger than the DZ twins and therefore a decline in earnings could come about at older ages. The effect of the proportion of those who are married on earnings was negative and not significant (p>0.05) for all three data samples, indicating that being married does not guarantee an individual an increase in earnings.

2.2 Variation around the mean returns to schooling

Comparison of the variance components corresponding to the random intercepts (family-level variance) between the null and second models (Tables 3 and 4) shows that family-level variance components for MZ twins decreased by 70% in the second model (Table 4). This indicates that individual and family characteristics explain a larger portion of the differences in the returns to schooling for MZ twins and that an earnings-education model that does not take into account these characteristics may overestimate the returns to schooling. Although, the family level variance for MZ twins in the second model is not significantly different from zero (p>0.05), the variance around the mean returns to schooling is, however, significant (p<0.05), Table 4.

Table 4 Results of the Hierarchical linear Model (HLM) including Covariates

Fixed Effects	Pooled	MZ twins	DZ twins
Family intercept, γ_{00}	5.6777**	5.0160**	5.8847**
	(0.2372)	(0.2999)	(0.3070)
Schooling (years) slope, γ_{10}	0.0878**	0.06801**	0.0886**
	(0.0113)	(0.0173)	(0.0133)
Age (years)	0.0141*	0.0399**	0.0085
	(0.0061)	(0.0084)	(0.0076)
Gender	–0.1488**	–0.1022	–0.1643**
	(0.0570)	(0.1315)	(0.0587)
Married	–0.0038	–0.0286	–0.0024
	(0.0896)	(0.1368)	(0.1034)
Father's schooling (years)	0.0061	0.0323**	–0.02051
	(0.0112)	(0.0111)	(0.0169)
Mother's schooling (years)	0.0137	–0.0093	0.0430*
	(0.0130)	(0.0132)	(0.0191)
Random Effects	Variance Components		
Family variance, τ_{00}	0.7652**	0.1835	0.9521**
	(0.2105)	(0.2632)	(0.2884)
Schooling slope, τ_{11}	0.0031**	0.0031*	0.0028**
	(0.0012)	(0.0019)	(0.0011)
Residual effect, σ^2_ε	0.07456**	0.0827**	0.0698**
	(0.0102)	(0.0163)	(0.0135)
Model Fit			
–2 Res log likelihood	398.1	164.8	242.0
AICc	406.1	172.8	250.0
N	250	106	144

Note: * = p < .05, *** = p < .001. Standard errors in parentheses below means.
Source: GTS authors' calculation

The variance components for the returns to schooling coefficient in the second model is denoted by τ_{11} and estimated as 0.0031 and 0.0028 with standard errors of 0.0019 and 0.0011 for the MZ and DZ twins' respectively (Table 4). These variances are higher than their standard errors suggesting that the second model picks up most of the variance in the returns to schooling that exist across families, though this variation is still significant (p<0.05). The significant variances of the regression slopes for the MZ and DZ twins data imply that returns to schooling varies across families and the values of 0.07 and 0.09 are just the expected returns across families (Table 4). This indicates that there could be some level of unobserved differences between MZ twins which may be attributed to individual characteristics. Similar random coefficient variance estimates are also associated with the returns to schooling for the pooled and DZ twins datasets and are significantly different from zero (p<0.05) using the Wald Z-test. This shows that the returns to schooling for these twins differ more than one could reasonably attribute to chance. REML returns to schooling results from Table 4 show significant (p<0.05) variation in the estimated intercepts and slope coefficients and therefore suggest that there exists heterogeneity in the returns to schooling. Since the random effects for the MZ and DZ twins are assumed to follow a normal distribution, about 67% of the returns to schooling regression coefficients for the MZ twins are expected to lie between an interval of (0.0123 and 0.1237) and about 95% are predicted to lie between (0.0411 and 0.1771). Similarly, about 67% of the returns to schooling regression coefficients for DZ twins are expected to lie between (0.0357 and 0.1415) and about 95% are predicted to lie between (–0.0151 and 0.1923). Thus, a return to schooling corresponding to the lower interval would indicate that if an employee is a DZ twin, annual family earnings is decreased by approximately 1.5% when compared with returns to schooling for non-DZ twins. Likewise the returns to schooling that corresponds to the upper limit of the interval would mean that annual family earnings for a DZ twin employee increased by 19% when compared with returns to schooling for non-DZ twins. A returns to schooling corresponding to the lower interval would indicate that if an employee is a MZ twin, annual family earnings are decreased by less than 5% when compared with returns to schooling for non-MZ twins. Likewise the returns to schooling that correspond to the upper limit of the interval would mean that annual family earnings for a MZ twin employee increased by 18% when compared with returns to schooling for non-MZ twins.

Furthermore, the Wald-Z test pointed out that the residual components which measured the variation not accounted for in the hierarchical linear models for both MZ and DZ twins in the null and second models were statistically significant (p<0.05). Interestingly, the residual variance associated with the returns to years of schooling for MZ twins in the second model decreased by about 11% whiles that of DZ twins decreased by about 65% (Table 4) when compared to the null model residual variances. This suggests that there is still some unobserved variation in returns to schooling for both MZ and DZ twins which could be attributable to measurement error in reported schooling levels and possible individual differences in inherent ability, among other reasons. Additionally, the significant REML residual variation is essentially due to the randomness of observed rates of returns to schooling and is an indication that returns to additional schooling varies randomly across individuals due to factors unknown to both the researcher and the individual at the time of their decisions.

According to the smaller-is-better rule for the information criteria, Model 2 has a smaller (AICc) (406.1) and a lower Restricted log likelihood (–2RLL) (398.1) compared to (AICc – 514.9) and (–2RLL – 510.9) of the null model and is therefore considered the best model. The probability chi-square of the difference in the log likelihood test of the models for the MZ, DZ and Pooled data sets, revealed that there were significant (p<0.01) differences between the null and the second model with explanatory variables (Table 5).

Table 5 Testing the significance of 2 Hierarchical linear Models for MZ and DZ Twins

Item	Difference in Log likelihood (–2LL)	Difference in *df*	*p*>chi-square
MZ – Model 1&2	29.9	2	3.21586E-07
DZ – Model 1&2	62.1	2	3.27459E-14
Pooled – Model 1&2	112.8	2	3.20473E-25

Source: GTS authors' calculation

3 DISCUSSION

Returns to schooling have been estimated as fixed coefficients using OLS methods in a number of labor economics studies. However, the OLS estimates may be inconsistent and biased when the returns to schooling vary across individuals as a result of observable factors as well as unobservable factors. Card, (1995) observed in a number of studies that different individuals acquire different returns to schooling and the same individual's returns to schooling vary with the level and type of schooling. In such situations the assumptions of non-varying slopes and intercepts, and uncorrelated residuals in standard OLS estimates are violated. Maximum likelihood estimators address the violation of the assumption of fixed coefficients by permitting intercepts and slopes to vary from one group to another. Moreover, in real life situations data collected are mostly of a hierarchical nature and statistical measures must be taken to exploit the opportunities offered by multilevel data structures. In order to obtain efficient and consistent estimates of the returns to schooling for a set of fully employed MZ and DZ twins, we used the REML estimation procedure in a hierarchical linear model to estimate the returns to schooling. Three types of parameters, namely the fixed effects, random effects and the variance-covariance components were estimated. REML estimated an unbiased variance around the mean returns to schooling parameters by accounting for the degrees of freedom lost by the estimation of the mean returns to schooling.

The estimated rates of return to schooling for the pooled, MZ and DZ twins ranged between 7% and 9%. These rates of returns to schooling are comparable to that of Conneely and Uusitalo (1999) who estimated a random coefficient model using Finish data that allowed for endogenous schooling and ability bias with an estimated maximum likelihood mean return to schooling of 6%. Similarly, the REML returns to schooling estimates of Sadeq (2014) using a hierarchical linear model varied between 7.8% and 8.4%. Moreover, Ashenfelter and Krueger (1994) found a higher restricted maximum likelihood estimate (16%) of the returns to schooling. Altogether, these results provide consistent and efficient estimates of the returns to schooling across individuals. The positive and somewhat large returns to schooling in the hierarchical linear model also indicate the importance of accounting for unmeasured ability and motivational factors that affect the returns to schooling.

Interestingly, MZ twins' earnings were significantly affected by fathers' education whiles mothers' education significantly influenced DZ twins' earnings. Thus, MZ twins' had better educated fathers who increased their children's education through transmission of innate ability, whereas DZ twins' had better educated mothers who raised their children's education by enhancing the "family learning environment." The effect of family background characteristics, i.e., parental education on returns to education is an important topic in the economics literature (Griliches, 1979). Part of this importance stems from the strong correlation between the educational attainment of parents and children, which may contribute to the transmission of socioeconomic status and inequality across generations. Parental education was found to positively and significantly affect the earnings of both MZ and DZ twins. Similarly, Anger and Schnitzlein (2013) concluded that family background variables play an important role in generating variation in the return to schooling. However, using twins data, Ashenfelter and Rouse (1998) are

of the view that the effects of family background (and ability) on returns are small. Altonji and Dunn (1996) also measured the effects of family background on the returns to schooling and found a positive though small effect of family background on returns in their preferred fixed effects specification.

Observed rates of returns to education may vary across individuals within the same educational group because of risk and unobserved heterogeneity. This study therefore added individual and family factors to the HLM to account for some of the variation in returns to schooling. The variance around the mean returns to schooling was decomposed into family heterogeneity, individual heterogeneity and risk. Significant ($p<0.05$) individual differences in the variance around the mean returns to schooling were observed for both MZ and DZ twins. However, in contrast to DZ twins, there were no significant unobservable family differences around the mean returns for MZ twins. This confirms the fact that MZ twins have similar ability and similar family background. This is consistent with findings by Ashenfelter and Krueger (1994) and Yew (2000) who did not find any statistically significant (i.e., $p<0.05$) sources of heterogeneity in the returns to schooling for MZ twins. These MZ twins' results suggest that individuals from higher ability families receive a lower marginal benefit from their human capital investment. On the contrary, significant family heterogeneity for DZ twins indicate that able individuals may attain more schooling because of higher marginal benefits to each additional year of education. Similarly, Bingley et al. (2005) exploited panel data using mixed model to show that there were significant variances to the returns to schooling estimates and found that individual variance in returns is smaller for MZ twins than for DZ twins. Correspondingly, Chen (2002) used US panel data (NLSY) to separate the variation in the returns to college into heterogeneity and risk components, and found that almost all the variation in returns is accounted for by the heterogeneity component.

Investing in education is always associated with some amount of risk (Hartog, 2011). This risk is the variation in the returns to education due to factors unobserved by the individual. The residual variance estimate which represents individual earnings risk was about 8% for MZ twins and 7% for DZ twins. These estimates are in line with some of the existing literature. Koop and Tobias (2004) apply the model to the NLSY and find a mean return of 12% with a dispersion of 7%. Chen (2002) also finds that the dispersion in returns to a US college education is 7%. Thus, the risk is quite large, even though we have allowed for differences by observable characteristics and it implies that a large number of the twins data set show very low returns to education. Interestingly, the residual variance for both MZ and DZ twins were statistically significant suggesting that the earnings risk associated with an additional year of schooling is important and therefore needs policy interventions. This earnings risk may result from lack of knowledge about individual ability and unanticipated changes in market conditions.

CONCLUSION

In this paper we have examined the restricted maximum likelihood (REML) estimation of a random coefficient model for earnings and its potential to provide unbiased returns to schooling regression coefficients. Results from our statistical and econometric analysis show that the mean return to schooling in the three selected cities in Ghana is between 7% and 9% which is comparable with worldwide estimates. Using the REML approach, the study also observed that there were significant variations around the mean returns to schooling across individuals which may partly be due to unobservable differences in individual ability and family background characteristics. The study further observed that family background characteristics (i.e. parental education) positively and significantly affect the earnings of both MZ and DZ twins. This is an indication that family background characteristics may play an important role in the relationship between earnings and schooling for genetically identical and similar twins. Consequently, the REML approach provides a robust alternative to the ordinary least squares method when returns to schooling vary across individuals and when data used is hierarchically structured. REML approach

to the estimation of the returns to schooling offers a measure of the true effect of schooling on earnings which has important implications for policy formulation and decision making within the education sector especially for developing countries.

ACKNOWLEDGEMENTS

We are grateful to the late Professor Louis Muyankazi of the Department of Mathematics and Statistics of the Kumasi Polytechnic and the academic staff of the Department of Mathematics and Statistics of the University of Cape-Coast, Ghana for their comments.

References

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

ALTONJI, J. G. AND DUNN, T. A. The Effects of Family Characteristics on the Return to Education. *Review of Economics and Statistics,* 1996, 78, pp. 692–704.

ANGER, S. AND SCHNITZLEIN, D. D. *Like brother, like sister? The importance of family background for cognitive and non-cognitive skills.* Annual Conference (Duesseldorf): Competition Policy and Regulation in a Global Economic Order, Verein für Socialpolitik/ German Economic Association, 2013.

ASHENFELTER, O. AND KRUEGER, A. Estimates of the economic return to schooling from a new sample of twins. *American Economic Review*, 1994, 84, pp. 1157–1173.

ASHENFELTER, O. AND ROUSE, C. Income, Schooling and Ability: Evidence from a New Sample of Identical Twins. *The Quarterly Journal of Economics*, 1998, 113, pp. 253–284.

BICKEL, R. *Multilevel analysis for applied research: It's just regression!* New York: Guilford Press, 2007.

BINGLEY, P., CHRISTENSEN, K., WALKER, I. *Twin-based estimates of the returns to education: evidence from the population of Danish twins*. Mimeo, 2005.

CARD, D. Estimating the Return to Schooling: Progress on Some Persistent Econometric Problems. *Econometrica*, 2001, 69, pp. 1127–1160.

CARD, D. The causal effect of education on earnings. *Handbook of Labor economics*, 1999, 3, pp. 1801–1863.

CARD, D. Earnings, schooling, and ability revisited. *Research in Labor Economics*, 1995, 14, pp. 23–48.

CHEN, S. *Is investing in college education risky?* University of Rochester and University of New York at Albany, Mimeo, 2002.

CONNEELY, K. AND UUSITALO, R. *Estimating Heterogeneous Treatment Effects in the Becker Schooling Model*. Princeton University, Mimeo, 1998.

GRILICHES, Z. Sibling models and data in economics: Beginning of a survey. *Journal of Political Economy*, 1979, 87, pp. 37–64.

HARTOG, J. Allocation and the earnings function. *Empirical Economics*, 1986, 11, pp. 97–110.

HECKMAN, J. J. AND VYTLACIL, E. J. Instrumental Variables Methods For the Correlated Random Coefficient Model: Estimating The Average Rate of Return to Schooling When the Return Is Correlated With Schooling. *Journal of Human Resources*, 1998, 33, pp. 974–1002.

HILDRETH, C. AND HOUCK, J. Some Estimators for a Linear Model with Random Coefficients. *Journal of the American Statistical Association*, 1968, 63, pp. 584–595.

KABOSKI, J. *Explaining schooling returns and output levels across countries*. Unpublished manuscript, 2007.

KOOP, G. AND TOBIAS, J. *Learning about Unobserved Heterogeneity in Returns to Schooling.* Dept. of Economics, University of Glasgow, Mimeo, 2002.

LAIRD, N. M. AND WARE, J. H. Random-effects models for longitudinal data. *Biometrics*, 1982, 38, pp. 963–974.

LINDSTROM, M. J. AND BATES, D. M. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association*, 1988, 83, pp. 1014–1022.

MAZUMDER, B. Sibling similarities and economic inequality in the US. *Journal of Population Economics*, 21, 2008, pp. 685–701.

NUNNALLY, J. AND BERNSTEIN, I. *Psychometric Theory*. New York: McGraw-Hill, 1994.

PATTERSON, H. D. AND THOMPSON, R. Recovery of interblock information when block sizes are unequal. *Biometrika*, 1971, 58, pp. 545–554.

PATRINOS, H. A., RIDAO-CANO, C. AND SAKELLARIOU, C. *Heterogeneity in Ability and Returns to Education: Multi-country Evidence from Latin America and East Asia*. World Bank Policy Research Working Papers, No. 4040, Washington, D.C., 2006.

PFEIFFER, F. AND POHLMEIER, W. *Causal returns to schooling and individual heterogeneity.* IZA Discussion Paper, No. 6588, 2012.

POLIT D. *Data Analysis and Statistics for Nursing Research.* Stamford, Connecticut: Appleton & Lange, 1996.
PSACHAROPOULOS, G. AND PATRINOS, H. Returns to investment in education: A further update. *Education Economics*, 2004, 12, pp. 111–134.
RAUDENBUSH, S. W. AND BRYK, A. S. *Hierarchical linear models: Applications and data analysis methods*, 2nd Ed. Newbury Park, CA: Sage Publications, 2002.
RAUDENBUSH, S., BRYK, A., CHEONG, Y. F. AND CONGDON, R. *HLM 5: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International, 2001.
SADEQ, TAREQ. *Formal-Informal Gap in Return to Schooling and Penalty to Education-Occupation Mismatch a Comparative Study for Egypt, Jordan, and Palestine*. Working paper series, 894, 2014.
VERBEKE, G. AND MOLENBERGS, G. *Linear Mixed Models for Longitudinal Data*. NY: Springer, 2000.
YEW LIANG LEE. Optimal Schooling Investments and Earnings: An Analysis Using Australian Twins Data. *The Economic Word*, 2000, 76, pp. 225–235.

Recent Publications and Events

New publications of the Czech Statistical Office

Czech Republic in International Comparison. Prague: CZSO, 2017.
Foreigners in the Czech Republic 2016. Prague: CZSO, 2016.
Indicators of Social and Economic Development of the Czech Republic 2000 – 3rd quarter 2016. Prague: CZSO, 2016.
Industrial Producer Price Indices 2016. Prague: CZSO, 2017.
Small Lexicon of Municipalities of the Czech Republic 2016. Prague: CZSO, 2016.

Other selected publications

BUCHER, S. *Konkurencieschopnosť a regionálne disparity v Európe.* Košice: Univerzita P. J. Šafárika, 2016.
EUROSTAT. *Energy, transport and environment indicators.* Luxembourg: Publication Office of the European Union, 2016.
EUROSTAT. *Sustainable development in the European Union.* Luxembourg: Publication Office of the European Union, 2016.
NEUBAUER, J., SEDLAČÍK, M., KŘÍŽ, O. *Základy statistiky. Aplikace v technických a ekonomických oborech.* 2nd Ed. Prague: Grada, 2016.
OECD Compendium of Productivity Indicators 2016. Paris: OECD, 2016.

Conferences

The ***61st ISI World Statistics Congress*** will take place in ***Marrakech,* Morocco from 16th to 21st July 2017**. More information available at: *http://www.isi2017.org*.

The ***20th International Scientific Conference Applications of Mathematics and Statistics in Economics (AMSE 2017)*** will this year be held in **Szklarska Poręba, Poland from 30th August to 3rd September 2017**. This scientific conference is organized each year by the Faculty of Informatics and Statistics of the University of Economics in Prague (the Czech Republic), Wrocław University of Economics (Poland) and the Faculty of Economics of Matej Bel University in Banská Bystrica (Slovakia). The conference aims to acquaint its participants with the latest mathematical and statistical methods that can be used in solving theoretical and practical problems and challenges of economics. The conference gives a unique chance to present research results achieved in the area where mathematics and statistics border with economics. More information available at: *http://amse.ue.wroc.pl.*

The ***19th Joint Czech-German-Slovak Conference on Mathematical Methods in Economy and Industry (MMEI)*** will take place in **Jindřichův Hradec, Czech Republic during 4–8 September 2017**. More information available at: *http://www.karlin.mff.cuni.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.

The *Discussion* section brings the opportunity to openly discuss the current or more general statistical or economic issues; in short, with what the authors would like to contribute to the scientific debate. Discussions shall have up to 6 000 words or up to 10 1.5-spaced pages.

The *Methodology* section gives space for the discussion on potential approaches to the statistical description of social, economic, and environmental phenomena, development of indicators, estimation issues, etc. Papers shall have up to 12 000 words or up to twenty (20) 1.5-spaced pages.

The *Book Review* section brings reviews of recent books in the fieled of the official statistics. Reviews shall have up to 600 words or one (1) 1.5-spaced page.

In the *Information* section we publish informative (descriptive) texts. The maximum range of information is 6 000 words or up to 10 1.5-spaced pages.

Language
The submission language is English only. Authors are expected to refer to a native language speaker in case they are not sure of language quality of their papers.

Recommended Paper Structure
Title (e.g. On Laconic and Informative Titles) — Authors and Contacts — Abstract (max. 160 words) — Keywords (max. 6 words / phrases) — JEL classification code — Introduction — … — Conclusion — Annex — Acknowledgments — References — Tables and Figures

Authors and Contacts
Rudolf Novak*, Institution Name, Street, City, Country
Jonathan Davis, Institution Name, Street, City, Country
* Corresponding author: e-mail: rudolf.novak@domainname.cz, phone: (+420) 111 222 333

Main Text Format
Times 12 (main text), 1.5 spacing between lines. Page numbers in the lower right-hand corner. *Italics* can be used in the text if necessary. *Do not* use **bold** or underline in the text. Paper parts numbering: 1, 1.1, 1.2, etc.

Headings
1 FIRST-LEVEL HEADING (Times New Roman 12, bold)
1.1 Second-level heading (Times New Roman 12, bold)
1.1.1 Third-level heading (Times New Roman 12, bold italic)

Footnotes
Footnotes should be used sparingly. Do not use endnotes. Do not use footnotes for citing references.

References in the Text
Place reference in the text enclosing authors' names and the year of the reference, e.g. "White (2009) points out that…", "… recent literature (Atkinson et Black, 2010a, 2010b, 2011, Chase et al., 2011, pp. 12–14) conclude…". Note the use of alphabetical order. Include page numbers if appropriate.

List of References
Arrange list of references alphabetically. Use the following reference styles: [for a book] HICKS, J. *Value and Capital: An inquiry into some fundamental principles of economic theory.* Oxford: Clarendon Press, 1939. [for chapter in an edited book] DASGUPTA, P. et al. Intergenerational Equity, Social Discount Rates and Global Warming. In PORTNEY, P., WEYANT, J., eds. *Discounting and Intergenerational Equity.* Washington, D.C.: Resources for the Future, 1999. [for a journal] HRONOVÁ, S., HINDLS, R., ČABLA, A. Conjunctural Evolution of the Czech Economy. *Statistika, Economy and Statistics Journal,* 2011, 3 (September), pp. 4–17. [for an online source] CZECH COAL. *Annual Report and Financial Statement 2007* [online]. Prague: Czech Coal, 2008. [cit. 20.9.2008]. <http://www.czechcoal.cz/cs/ur/zprava/ur2007cz.pdf>.

Tables
Provide each table on a separate page. Indicate position of the table by placing in the text "insert Table 1 about here". Number tables in the order of appearance Table 1, Table 2, etc. Each table should be titled (e.g. Table 1 Self-explanatory title). Refer to tables using their numbers (e.g. see Table 1, Table A1 in the Annex). Try to break one large table into several smaller tables, whenever possible. Separate thousands with a space (e.g. 1 528 000) and decimal points with a dot (e.g. 1.0). Specify the data source below the tables.

Figures
Figure is any graphical object other than table. Attach each figure as a separate file. Indicate position of the figure by placing in the text "insert Figure 1 about here". Number figures in the order of appearance Figure 1, Figure 2, etc. Each figure should be titled (e.g. Figure 1 Self-explanatory title). Refer to figures using their numbers (e.g. see Figure 1, Figure A1 in the Annex).

Figures should be accompanied by the *.xls, *.xlsx table with the source data. Please provide cartograms in the vector format. Other graphic objects should be provided in *.tif, *.jpg, *.eps formats. Do not supply low-resolution files optimized for the screen use. Specify the source below the figures.

Formulas
Formulas should be prepared in formula editor in the same text format (Times 12) as the main text.

Paper Submission
Please email your papers in *.doc, *.docx or *.pdf formats to statistika.journal@czso.cz. All papers are subject to double-blind peer review procedure. You will be informed by our managing editor about all necessary details and terms.

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

Managing Editor: Jiří Novotný
phone: (+420) 274 054 299
fax: (+420) 274 052 133
e-mail: statistika.journal@czso.cz
web: www.czso.cz/statistika_journal
address: Czech Statistical Office | Na padesátém 81 | 100 82 Prague 10 | Czech Republic

Subscription price (4 issues yearly)
CZK 372 (incl. postage) for the Czech Republic,
EUR 110 or USD 165 (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: Václav Adam
Print: Czech Statistical Office

97th year of the series of professional statistics and economy journals of the State Statistical Service in the Czech Republic: ***Statistika*** (since 1964), ***Statistika a kontrola*** (1962–1963), ***Statistický obzor*** (1931–1961) and ***Československý statistický věstník*** (1920–1930).

Published by the Czech Statistical Office
ISSN 1804-8765 (Online)
ISSN 0322-788X (Print)
Reg. MK CR E 4684