STATISTIKA

STATISTICS
AND ECONOMY
JOURNAL

VOL. **94** (1) 2014

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

ASSOCIATE EDITORS

MANAGING EDITOR

CONTENTS

ANALYSES

METHODOLOGY

BOOK REVIEW

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 a professional double-blind peer reviewed journal included (since 2008) in the List of Czech non-impact peer-reviewed periodicals (updated in 2013). 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 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

Formation of Aggregate Demand and Supply in the Czech Republic

Lukáš Kučera[1] | *Czech Statistical Office; University of Economics, Prague, Czech Republic*

Abstract

Great changes were happening in the economy of the Czech Republic during the last twenty years (1990–2011). One of the biggest was a huge increase of the foreign trade importance. The growing foreign trade then formed an aggregate demand as well as aggregate supply. If the economy in the beginning of the 90´s faced mostly the volatility of domestic demand, in 2011 the effect of domestic and foreign factor was comparable. As the economy in the 90´s was not capable to produce sufficient amount of products to satisfy domestic demand, at the end of last decade it had no problem to do so – the surplus of domestic supply over domestic demand was then situated to the foreign market.

Keywords

Aggregate demand, aggregate supply, domestic demand, domestic supply

JEL code

E20, E29, F41, O11

INTRODUCTION

Macroeconomic development can be analyzed from different perspectives. The most often used indicator is gross domestic product and further analysis of its expenditure items. Much less attention is paid to gross national income or real gross domestic income. Almost none is paid to aggregate demand and supply, their development in time or their mutual relations.

The aim of this paper is to analyze the way of formation of aggregate demand and aggregate supply with respect to their structure in the economy of the Czech Republic since 1990 till 2011.[2] Thus, the aspect of domestic and foreign demand/supply, their decomposition into individual items and their development in time – all with respect to factors standing behind, were subject to discussion.

1 METHODOLOGY

On the site of GDP use – the purposes of use of GDP generated in a certain time period are analyzed. Specifically, European system of accounts ESA 1995 (CZSO, 2000) distinguishes final consumption expenditure, gross capital formation and the difference between export and import. According to Hronová et al. (2009), we are able to describe the GDP use in the form of:

[1] Czech Statistical Office, Na padesátém 81, Prague, Czech Republic; University of Economics, W. Churchill Sq. 4, Prague, Czech Republic. E-mail: kucera-lukas@email.cz, phone: (+420)274052254.

[2] Data for 2012 were not available at the time of completion of this paper (export and import; annual national accounts).

$$GDP = FCE + GCF + EX - IM, \quad (1)$$

where:

FCE	final consumption expenditure,
GCF	gross capital formation,
EX	export,
IM	import.

Final consumption expenditure can be decomposed into final consumption expenditure of households, government and non-profit institutions serving to households (NPISH). Gross capital formation can be decomposed into gross fixed capital formation (investment), change in inventories and net acquisition of valuables. So, equation (1) can be rewritten in the form of:

$$GDP = FCE_H + FCE_G + FCE_{NPISH} + GFCF + CHII + NAoV + EX - IM, \quad (2)$$

where:

FCE_H	final consumption expenditure of households,
FCE_G	final consumption expenditure of government,
FCE_{NPISH}	final consumption expenditure of NPISH,
GFCF	gross fixed capital formation,
CHII	change in inventories,
NAoV	net acquisition of valuables.

Domestic demand is formed according to Spěváček (2006) by the sum of final consumption expenditure and gross capital formation. Mandel and Tomšík (2006) call this sum an absorption. Therefore, domestic demand equals to:

$$D = FCE_H + FCE_G + FCE_{NPISH} + GFCF + CHII + NAoV. \quad (3)$$

CZSO (2006) defines so-called domestic realized demand – it is domestic demand without change in inventories and net acquisition of valuables:

$$D = FCE_H + FCE_G + FCE_{NPISH} + GFCF. \quad (4)$$

Counterpart of domestic demand is domestic supply. Spěváček (2006) defines this supply as GDP. CZSO (2006) adjusts GDP by change in inventories and calls it domestic effective supply. When net acquisition of valuables is not included in domestic realized demand, we have to incorporate it into domestic effective supply. Therefore, domestic effective supply can be expressed as follows:

$$S = GDP - CHII - NAoV. \quad (5)$$

We identify domestic demand in this paper according to equation (4), domestic realized demand. Domestic supply is identified according to equation (5), domestic effective supply. We consider, therefore, that the negative change in inventories increases volume of offered value with respect to GDP and, therefore, it increases domestic supply. On the contrary, when change in inventories is positive, there is a decline of offered value with respect to GDP – created value is partly allocated into inventories.

So, foreign demand equals to export of goods and services, foreign supply to import of goods and services. *"The difference between domestic demand and domestic supply equals to the balance of foreign demand and foreign supply with the opposite sign"* (CZSO, 2006, pp. 15). If domestic demand is higher than domestic supply, foreign demand will be lower than foreign supply. And vice versa. If we sum domestic and foreign demand, we will get so called aggregate demand; if we sum domestic and foreign supply, we will get so called aggregate supply (CZSO, 2006).

Individual segments of aggregate demand and supply are analyzed between 1990 and 2011. We have two options of analysis – to use data at current prices or data at constant prices. Due to the intertemporal comparability we use constant prices (prices of year 2005).[3] However, we have to solve the problem of non-aditivity of equation (2) – this non-aditivity appears when chaining aggregates at current prices into aggregates at constant prices. Because every aggregate is chained by its relevant index (deflator), the equality at constant prices is not ensured according to equation (2) (see for example Fischer, 2005a or Široký, 2004).

For example Sixta et al. (2011) use data at constant prices to calculate an investment ratio – yet, according to them *"the information capability should not be negatively affected"* (Sixta et al., 2011, pp. 603). For this paper, there has been made a slight approximation of values at constant prices which eliminates the non-aditivity problem (similarly to Kučera, 2012).

At first, volume of change in inventories in every year is calculated using GCF, GFCF and NAoV. One can expect that using this approach, residuum originating from chaining of GCF, GFCF and NAoV is contained in CHII. Nevertheless, CHII value obtained in this way is more accurate, than in the case when it would be calculated as it is made at current prices – it partly balances GDP acquired by production approach and expenditure approach (Fischer, 2005b) – in this case, full residuum originating from chaining aggregates presented in equation (2) would be contained in CHII.

To establish aditivity in terms of equation (2), residuum between left side of the equation and right side of the equation in every particular year is calculated. Finally, this residuum is distributed into individual aggregates according to their weights. Obtained values using this approach are aditive according to equation (2) whereas mutual volume position of adjusted aggregates is not distorted.

2 RESULTS

Data of national accounts are used (CZSO, 2013b). Original volumes of GDP expenditure items and their balanced volumes are attached in the Annex (Table 1 and 2). Acquired data of aggregate demand and supply including domestic demand/supply and foreign demand/supply are attached in the Annex (Table 3). In follow-up analysis, exclusively balanced volumes are discussed, if not stated otherwise.

The trend of aggregate demand and supply was growing in 1990–2011. Average annual growth rate reached 3.8%. Volume of aggregate demand and supply increased from CZK 2,919 bil. in 1990 to CZK 6,341 bil. in 2011. Specific formation of aggregate demand (hereinafter only "AD") and aggregate supply (hereinafter only "AS") was influenced by domestic and foreign demand/supply development.

2.1 Long-term development

2.1.1 Aggregate demand

Domestic demand represented 79.3% of AD in 1990. The most significant part of AD was made up of the final consumption expenditure of households (42.2%); almost the same part was formed by final

[3] Singer (2013) proposes an alternative method, which is appropriate for an analysis of long-term performance of transformation and post-transformation economies. He suggests to switch an aggregate at current prices in domestic currency into an aggregate in currency of base economy (he suggests EUR) and than to adjust data for growth of prices in this area (thus, growth of prices in euro area). Singer (2013) uses this method to adjust GDP, calls it as "comparable real GDP" (Singer, 2013, pp. 9) and states, that performance of transformation or post-transformation economies is much higher using this "comparable real GDP" growth than in the case when one uses GDP growth at constant prices.

consumption expenditure of government (18.6%) and gross fixed capital formation (18.1%).[4] Remaining part of AD was formed by foreign demand (20.7%). It is obvious, that AD has been very vulnerable to changes of domestic demand in the beginning of the 90´s.

Domestic demand equaled to CZK 2,315 bil. in 1990 and CZK 3,333 bil. in 2011. This increase equaled to 44%. Foreign demand grew up from CZK 604 bil. in 1990 to CZK 3,008 bil. in 2011. This growth equaled to 397.9%. So, gradual increase in AD in the Czech Republic was determined by enormous growth of foreign demand due to the involvement of Czech producers-exporters on the foreign market.

In 2011, significant 47.4% of AD was formed by foreign demand while share of domestic demand fell to 52.6% only. Vulnerability of AD to changes of domestic demand rapidly decreased. On the other hand, the vulnerability of AD to external factors increased significantly. An example may be the financial recession in 2009, which broadly (and negativelly) affected AD through foreign demand.

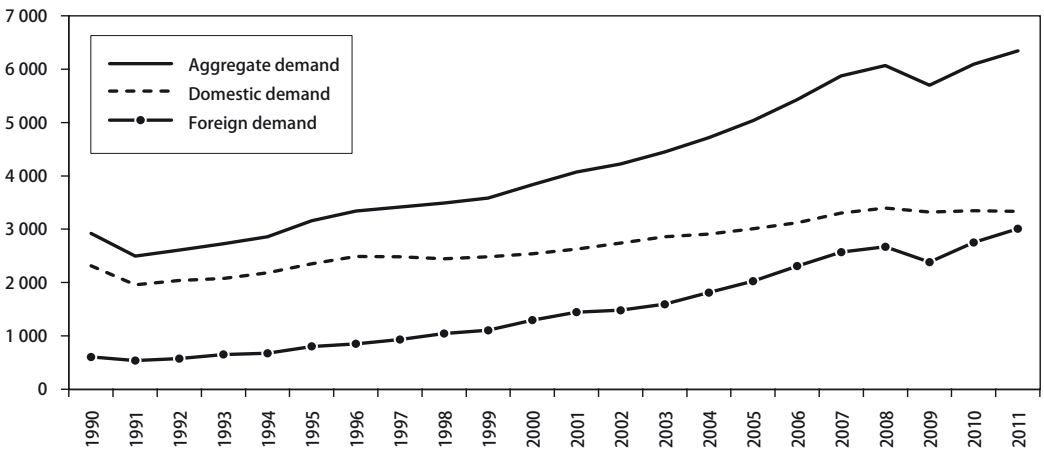
As we decompose domestic demand further, we find out, that share of final consumption expenditure of households and government dropped the most – to 27.1% and 10.8%. Share of gross fixed capital formation decreased as well, but only to 14.3%.

In 1990, final consumption expenditure of government formed greater part of AD than gross fixed capital formation. In 2011, however, it was vice versa. The break happened already in 1995, the change was affected by gradual transformation of Czech economy – due to the privatisation influencing growth of the investment activity.

With respect to the fact, that in this paper there are used balanced data at constant prices (due to the intertemporal comparability), the view of the AD structure development can be distorted. Thus, it is appropriate to mention, what was the AD structure development at current prices.

In 1990, AD was formed by 68.3% of domestic demand and by 31.7% of foreign demand. Therefore it is obvious, that foreign demand at current prices played more significant role in determining AD. In 2011, domestic demand formed 56.4% of AD, foreign demand formed the rest. In this year,

Figure 1 Aggregate demand and its components (constant prices, balanced, in bil. CZK)

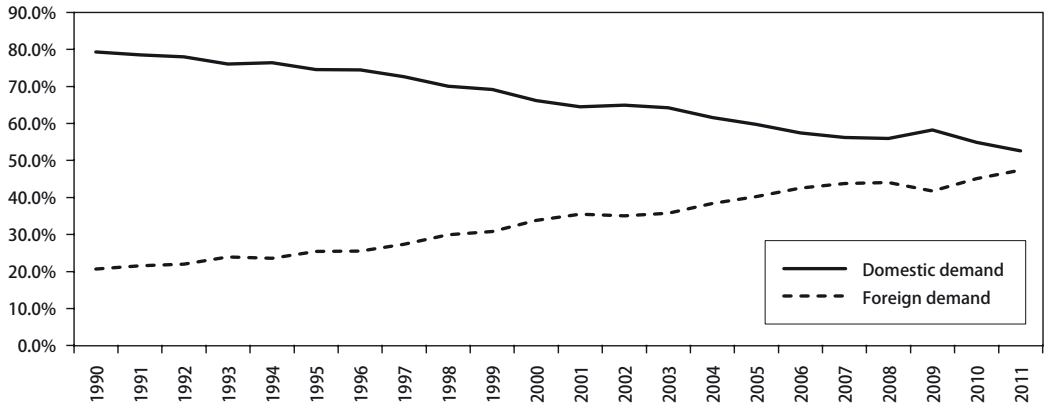


Source: CZSO (2013b), own calculations

[4] Final consumption expenditure of non-profit institutions serving to households formed since 1990 till 2011 only 0.4% of AD on average. So, this item is not considered in follow-up text.

on the contrary, balanced data at constant prices overestimate effect of foreign demand in forming AD. It is important, however, that in both approaches – balanced constant prices/current prices – there was apparent long-term increase of foreign demand significance at the expence of domestic demand.

Figure 2 Structure of aggregate demand (constant prices, balanced, in %)



Source: CZSO (2013b), own calculations

2.1.2 Aggregate supply

Domestic supply made 80.8% of AS in 1990 – 80.7% consisted of GDP produced, 0.1% of inventories decline.[5] Foreign supply formed 19.2% of AS. A short-term blip occured in 1991 – share of domestic supply increased to 84.7%, share of foreign supply dropped to 15.3%. However, the long-term development was the opposite.

Domestic supply was in 2011 only 53.7% higher compared to 1990, while foreign supply had grown in this time period by 383.8%. In 2011, domestic supply formed 57.1% of AS, foreign supply formed 42.9% of AS.

Due to this, we can say, that gradual growth of AS in the Czech Republic was driven primarily by foreign supply. The increase of foreign supply share was caused by higher demand for foreign products consumption and investment as well as by usage of imported products as inputs for production process of domestic producers-exporters. Regarding to AD, AS was getting more and more vulnerable to foreign changes as well.

Even in case of AS structure development, it is necessary to evaluate the structure development at balanced constant prices to the structure development at current prices.

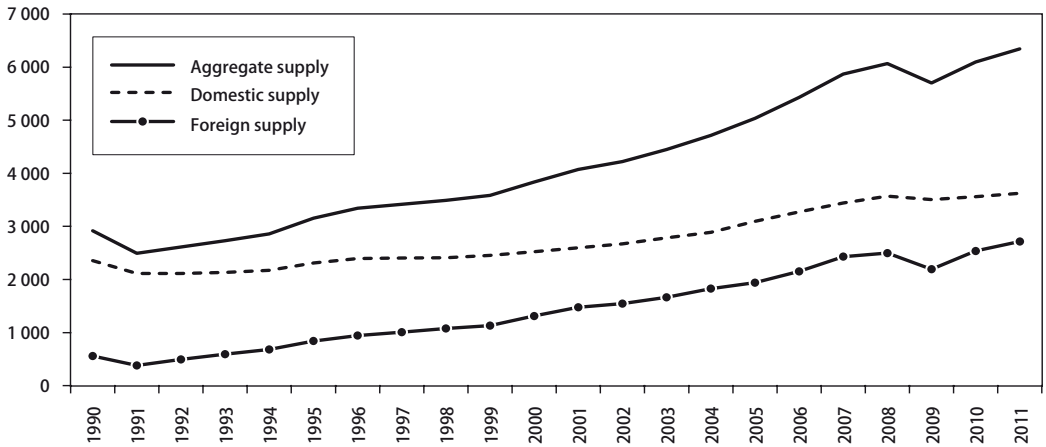
AS at current prices in 1990 was formed by 70% of domestic supply and by 30% of foreign supply. Even in this case applies, that foreign supply at current prices played more significant role in forming AS in this year. Till 2011, the significance of domestic supply had fell to 58.9%, the significance of foreign supply had increased to 41.1%. Thus, data at balanced constant prices slightly overestimate the effect of foreign supply in 2011.

The fact, that there was apparent long-term increase of foreign demand share on AD and long-term increase of foreign supply share on AS as well, was not accidental. Gradual increase of foreign demand was connected to foreign direct investment flow in the Czech economy which led to production of goods

[5] Net acquisition of valuables formed since 1990 till 2011 approximately –0.1% of AS on average. So, this item is not considered in follow-up text.

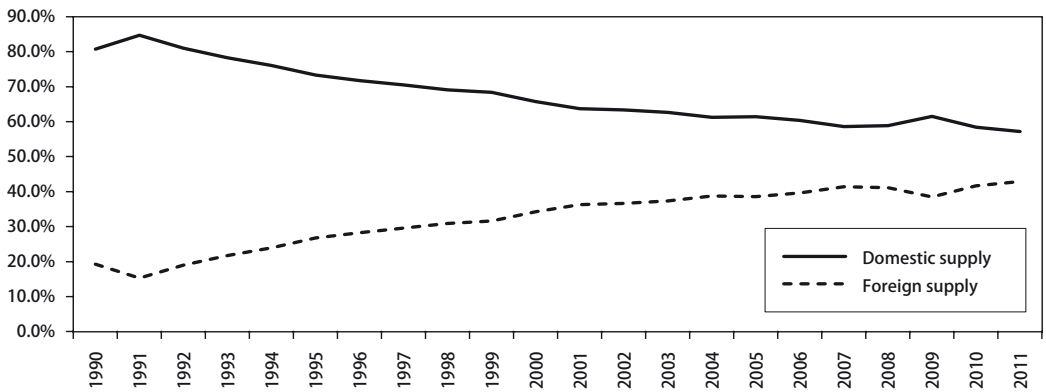
with higher value added – these goods were highly demanded from abroad. However, producers-exporters needed (and still need) inputs for production, which cannot have been fully satisfied by Czech economy – therefore, a growth of foreign supply and an increase of foreign supply share on AS were monitored. It can be summarized, that one of the main reason of long-term increase of foreign demand share on AD and foreign supply share on AS was strong orientation of Czech economy on product transformation.

Figure 3 Aggregate supply and its components (constant prices, balanced, in bil. CZK)



Source: CZSO (2013b), own calculations

Figure 4 Structure of aggregate supply (constant prices, balanced, in %)



Source: CZSO (2013b), own calculations

2.2 Structural changes in the economy from the perpective of aggregate demand and supply

2.2.1 Shock of the transition to a market economy

Significant drop of AD in 1991 was mostly affected by reduction of domestic demand. AD decreased by 14.5%. Domestic demand contributed by –12.2 percentage points, foreign demand by remaining –2.3 percentage points.

All parts of domestic demand contributed to decline of AD. The most significant items were final consumption expenditure of households (–7.1 percentage points) and gross fixed capital formation (–4.3 percentage points). Causes of this negative development may be found in devaluation of CZK in 1990 as well as in liberalization of prices[6] due to the transition to a market economy. This brought about very high annual inflation in the amount of 56.6% in 1991 (CZSO, 2013a).

The inflation caused devaluation of savings of households but also increase in the cost of living. Therefore, households were forced to reduce their consumption. Enormous growth of prices had an impact on the investment activity as well. The activity was tampered by devaluation of national savings – therefore by lack of domestic resources for investment. But there may be found another reason – substantial growth of prices made impossible to plan future costs, sales, own prices as well as price relations between individual products by producers – these difficulties reduced their investment too.

AS equals to AD – due to the balance equilibrium which is valid always. So, AS faced the same decline as AD.

The drop of AS was caused mainly by reduction of domestic supply. This contributed to decrease of AS by –8.3 percentage points. Very deep fall affected GDP – it decreased by CZK 264 bil. (–11.2%) and contributed to reduction of AS by –9 percentage points. Only a slight compensation was a decline of CHII – it contributed by 0.7 percentage points.

Source of domestic supply drop can be found in considerable reduction of domestic demand, which decreased by CZK 355 bil. Domestic producers lowered their production due to weaker domestic demand for their products. Nevertheless, domestic supply decrease was less pronounced – it was reduced by CZK 243 bil. only. Thus, Czech economy increased surplus of domestic supply over domestic demand and allocated it on the foreign market – even despite the negative development of foreign trade in the Czech Republic in this year.

The role of foreign factor in forming AD/AS in 1991 was stronger with respect to AS. Foreign supply decreased by 31.9%, foreign demand by 11.1%. Despite the fact that both foreign demand/supply dropped, decline of foreign demand was shallower. This was highly influenced by mentioned devaluation of CZK in 1990 which increased the price competitiveness of domestic exporters in the foreign market. Nevertheless, competitiveness of domestic producers was still considered as weak – firms still used old machinery and production technology which limited them in producing products with higher added value.

2.2.2 Gradual transformation of the economy (till 1996)

AD was growing till 1996 very fast – by 6% annualy on average. The highest growth-rate was achieved in 1995 (growth in the amount of 10.4%, the highest growth over the whole time period 1990–2011). Major role in rapid growth of AD till 1996 may be accounted for domestic demand (it contributed by 3.8 percentage points on average).

Households´ resources were growing so households may have increased their consumption. Growth of final consumption expenditure of households was driven by not saturated consumption due to unavailability of many products in previous period of centrally planned economy (see for example Dubská, 2013). Increase in domestic demand was attributed to a significant growth of investment activity as well, which was probably supported by the privatisation – private entrepreneurs were interested in renewal of old machinery and other production equipment – the highest growth of investment was observed in 1995 (by almost one quarter). Finally, foreign demand contributed to AD growth by 2.3 percentage points on average.

Growth of AS till 1996 was caused mainly by the increase in foreign supply – contribution of foreign supply was higher than contribution of domestic supply in every particular year. The highest con-

[6] Liberalization did not affect administered prices, electricity, rent, medical service etc. (Singer, 2007).

tribution of foreign supply was observed in 1995 (+5.5 percentage points) – moreover, the amount of this contribution was, except year 2010 (due to low comparative base in 2009), the highest contribution of this part of AS over the whole time period 1990–2011. It seems, that domestic producers were not able to satisfy consumption and investment requirements forming domestic demand yet (due to persistent underdeveloped production technology and low ability to produce products with high added value) – so, domestic requirements had to be satisfied mainly by foreign supply.

2.2.3 Monetary policy disturbances

Years 1997 and 1998 brought about significant changes in the economy. External imbalance (in terms of current account) was deepening and exchange rate of CZK was no longer sustainable. Czech National Bank was forced to abandon fixed rate regime and started to use managed floating. The inflation targeting was chosen as the transmission mechanism of monetary policy, which subsequently led to reduction of inflation rate.

However, realized changes adversely affected investment activity (gross fixed capital formation decreased in the amount of 6.5% in 1997) – entrepreneurs were probably not capable to assess future economic development and due to uncertainty they cut down their investment. Drop in investment activity reduced domestic demand by 0.4%. However, foreign demand continued in positive development and inflicted AD growth even in this year by +2.2%.

The fall in domestic demand in 1998 deepened and reached 1.4%. Decline in investment activity continued. However, in 1998, even consumption of households weakened – this fall was influenced by significant growth of rate of unemployment from 4.8% in 1997 to 6.5% in 1998 (CZSO, 2013c; methodology of Labour Force Survey) resulting in lower households´s revenue.[7] It is worth mentioning, that the decline of final consumption expenditure of households appeared only twice since 1991 till 2011 (in already mentioned 1991 – by dramatic 16.8%, and in 1998 by 1.2%).

Final consumption expenditure of government lowered in 1998 by 2.8%. It seems, that cut in planned government expenditure in 1997 (Páral, 2001) took effect a year later.

So, in 1998 all major items of domestic demand contributed to its decline. However, even in this year foreign demand growth was capable to compensate decline of domesic demand – AD increased the same rate as in 1997.

Alhough neither domestic/foreign supply in those years declined, growth-rates of both were much lower with respect to previous years – due to reduced domestic demand. GDP fell in both years, only decline of CHII slightly increased domestic supply. It seems, that large part of stored products were sold abroad.

If AS growth till 1996 was driven mainly by foreign supply growth, this fact was even deepened in 1997 and 1998 – foreign supply affected growth of AS by 86% and 94%.

2.2.4 Growth of foreign trade importance (till 2008)

Since 1999 (including), growth of AD was mainly a result of growth in foreign demand. Contributions of foreign demand outweighted contributions of domestic demand in every particular year except 2002 and 2003. In strong years 2004–2006 (after joining the European Union), contributions of foreign demand were several times higher.

Although it should be noted, that even domestic demand contributed to growth positively – households increased their final consumption expenditure annualy due to still not saturated consumption. Gradual flow of foreign direct investment in the Czech Republic initiated another investment growth. Even households contributed significantly to growth of investment in years 2005–2007 – they highly purchased own housing.

---

[7] In real terms – nominal values were deflated by CPI.

Growth of AS was still initiated primarily by increasing foreign supply. Only exceptions were years 2002, 2005 and 2008.

According to data of Czech National Bank (2013), there was a very high flow of foreign direct investment in the Czech Republic in years 1998 till 2002. This may have been partly accounted for the introduction of incentives in 1998. As Říman (2008) states, however, the effect of incentives was not that strong as one may expect – he argues, that foreign direct investment flow was mostly affected by political stability, good infrastructure, cheap and well-educated labor force and other; not by incentives themselves.

While Tomšík (2008) states that this inflow of foreign direct investment did not bring only positives – which we agree – it helped the Czech Republic in turning into a more open economy – due to these investment, Czech producers were able to produce products with higher added value which could be sold abroad. This has been increasing foreign demand. However, Czech producers were forced to produce these products using foreign inputs - this resulted in foreign supply growth.[8] As a whole, in the Czech Republic, the relationship between foreign demand/supply, which is a typical atribute for small open economies, had been depening.

2.2.5 Impact of financial crisis

AD faced second significant drop in 2009 (after decline in 1991). It fell by 6% compared to 2008. Despite the decline of AD in 1991, drop in 2009 was affected mostly by reduction of foreign demand (it contributed by –4.7 percentage points). Negative development of foreign demand was primarily influenced by lower performance of main foreign partners of the Czech Republic[9] due to financial crisis impacts. These countries among others reduced imports which negatively affected foreign demand for goods and services in the CR.

Domestic demand contributed to decline of AD by –1.3 percentage points. The reason may be found in decreasing investment activity. Domestic producers were afraid of future development of foreign demand (whether it will recover or not), households cut down expenditure for purchasing a housing (mainly due to a strong demand in previous years). Limited investment activity contributed to decline of AD by –1.8 percentage points. While final consumption expenditure of government tampered drop of AD by 0.5 percentage point (government acted countercyclically), final consumption expenditure of households nearly stagnated.

Drop of AS was primarily affected by foreign supply which contributed by –5 percentage points. As one can note, almost the same contribution was found out regarding the effect of foreign demand on AD (–4.7 percentage points). This was done by significant connection of foreign demand and supply, when domestic producers-exporters used imported products as inputs for another production. This phenomenon, however, could not be observed in the case of AD/AS drop in 1991 – at that time the Czech Republic formed a part of the international market only for a short time period with significant restrictions – full convertibility of CZK did not exist, economy almost did not possess of foreign direct investment from abroad which may have participated in higher production of products with high added value demanded from abroad (increasing demand for foreign inputs). Thus, connection between export and import in 1991 was not created yet.

Remaining share of AS decline in 2009 fell to domestic supply (–1 percentage point). Specifically, it was the decline of GDP (–2.8 percentage point). As in 1991, drop in CHII tampered reduction of AS (it contributed by 1.7 percentage points) – the volume the economy did not produce was partly provided from inventories again.

---

[8] Foreign supply was growing due to relatively strong domestic demand as well (products were imported for consumption and investment). However, we can deduce, that the weight of inputs for production in import volume gradually increased.

[9] Main foreign partners of the Czech Republic regarding export are Germany, Slovakia and Poland (CZSO, 2012).

2.2.6 Recovery of foreign trade in 2010 and 2011

Growth of AD after 2009 was influenced almost exclusively by foreign demand. AD increased in the amount of 6.9% in 2010 and by 4% in 2011. Contribution from the site of foreign demand reached 6.5 and 4.2 percentage points, respectively. Contribution from the site of domestic demand equaled only 0.4 percentage point in 2010, in 2011 it was even negative (–0.2 percentage point).

It is obvious, that domestic demand was very weak in these years. It was caused primarily by fiscal restrictions – according to Vintrová (2012), the government did not distinguish between current and capital expenditure. Therefore, fiscal restrictions did not reduce final consumption expenditure of government only, but it also tampered their investment activity. However, austerity measures negatively affected even final consumption expenditure of households – on the site of revenue, households faced decline.[10] On the site of expenditure, there was a decrease of savings, however, not deep enough to enable growth of consumption of households as before 2009. According to Zamrazilová (2012, pp. 11) *"weakening of consumption of households was a combination of worse revenue situation and more significant risk perception"*.

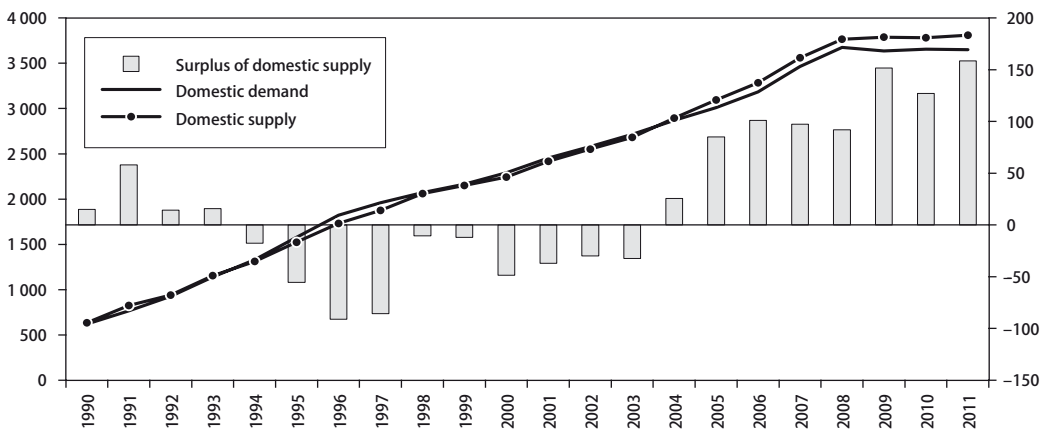
Growth of AS after 2009 was also mostly affected by foreign factor. Foreign supply contributed to growth of AS by 6 percentage points in 2010 and 3 points in 2011. So, even in these years one could have observed significant connection between export and import (foreign demand and supply).

Domestic supply contributed to AS by 0.9 percentage point only in 2010 and by 1.1 point in the next year. Czech producers adapted to weak domestic demand. As the main driver for a slight increase in domestic supply remained relatively strong foreign demand.

2.3 Domestic demand coverage

As mentioned before, AD equals AS and together make always the equilibrium. The difference, however, can be found between domestic demand and supply and foreign demand and supply. In this part, structure of AD and AS, but also difference between domestic supply and domestic demand, was discussed. Due to this fact, there are used data at current prices.

Figure 5 Domestic demand and domestic supply (current prices, in bil. CZK, left axis), surplus of domestic supply (current prices, in bil. CZK, right axis)



Source: CZSO (2013b), own calculations

[10] In real terms – nominal values were deflated by CPI.

Since 1990 till 1993 there was a surplus of domestic supply over domestic demand. The most significant difference was observed in 1991, when domestic demand was covered by domestic supply in the amount of 107.6%. Reasons may be found especially in the weak purchasing power of households and firms which did not have resources to buy foreign goods and services. Surplus of domestic supply was than allocated abroad.

Domestic demand coverage was decreasing with gradually enhancing economic situation of domestic subjects in later years. It dropped below 100% value in 1994 and reached 98.7%. Till 2003 (including), it did not exceed 100% threshold. Domestic supply was not able to satisfy the needs of Czech economy. Hunt for consumption and investment oppurtunities exceeded domestic supply capacity.

In 2004, domestic supply exceeded domestic demand in the amount of 0.9%. Domestic demand coverage has dropped below 100% threshold never again since this year. On the contrary, there was a gradual increase in the difference between domestic supply and domestic demand. In 2011, coverage reached 104.3%. Positive difference between domestic supply and domestic demand iniciated surplus of foreign demand over foreign supply – positive balance of goods and services.

However, the roots of surplus of domestic supply over domestic demand in these years where much different compared to the beginning of the 90´s. Households and firms could have purchased variety of goods and services for consumption or purchased investment on the foreign market for global prices – they had resoures to do so. This was the first difference in these years compared to the beginning of the 90´s. On the other hand, the economy passed during last twenty years considerable development and had a lot to offer on the foreign market. This was the second difference compared to the beginning of the 90´s. What is important, this factor prevailed and domestic supply had been exceeding domestic demand. The economy is producing surpluses.

CONCLUSION

Aggregate demand and supply describe macroeconomic development from a different perspective. While the method of gross domestic product use analyzes for what purposes GDP generated in a certain time period is used, theorem of aggregate demand and supply discusses a mutual interaction of forces of domestic and foreign environment in shaping the aggregate equilibrium.

This equilibrium was changing a lot in the Czech Republic. While the aggregate demand/supply was formed in the beginning of the 90´s (at balanced constant prices) by 79.3%/80.8% of domestic demand/supply, in 2011 it was only by 52.6%/57.1%. Vulnerability to the impact of foreign shocks of the economy increased significantly (more in the case of aggregate demand).

While aggregate demand equals aggregate supply always, domestic demand and supply does not. Their mutual position is determined by the dynamics of their components. In the beginning of the 90´s, domestic demand was fully covered by domestic supply (at current prices). This relation was gradually weakening and between 1994 and 2003 there was observed a surplus of domestic demand over supply. This was caused by a high growth of overall consumption and investment combined with an insufficient increase in domestic economy performance. This trend was interrupted in 2004 – domestic demand was fully covered by domestic supply since this year again. Performance of the economy surpassed domestic demand and created surpluses could have been situated on the foreign market.

References

CZECH NATIONAL BANK. *ARAD systém časových řad* (ARAD Data Time Series System) [online]. Prague: CNB, 2013. [cit. 1.7.2013]. <www.cnb.cz/docs/ARADY/HTML/index.htm>.

CZECH STATISTICAL OFFICE. *Česká republika od roku 1989 v číslech* (Czech Republic in Numbers since Year 1989) [online]. Prague: CZSO, 2013a. [cit. 1.7.2013]. <www.czso.cz/csu/redakce.nsf/i/cr_od_roku_1989#03>.

CZECH STATISTICAL OFFICE. *Databáze ročních národních účtů* (Database of Annual National Accounts) [online]. Prague: CZSO, 2013b. [cit. 19.6.2013]. <apl.czso.cz/pll/rocenka/rocenka.indexnu>.

CZECH STATISTICAL OFFICE. *Evropský systém účtů ESA 1995* (European System of Accounts ESA 1995) [online]. Prague: CZSO, 2000. [cit. 21.11.2013]. <apl.czso.cz/nufile/ESA95_cz.pdf>.

CZECH STATISTICAL OFFICE. *Zahraniční obchod České republiky 2011* (External Trade of the Czech Republic in 2011) [online]. Prague: CZSO, 2012. [cit. 19.6.2013]. <www.czso.cz/csu/2012edicniplan.nsf/p/6008-12>.

CZECH STATISTICAL OFFICE. *Zaměstnanost, nezaměstnanost – časové řady* (Employment, Unemployment – Time Series) [online]. Prague: CZSO, 2013c. [cit. 17.7.2013]. <www.czso.cz/csu/redakce.nsf/i/zam_cr>.

CZECH STATISTICAL OFFICE. *Zdroje HDP a jejich užití v letech 1995 až 2005* (Resources of GDP and Their Use between 1995 and 2005) [online]. Prague: CZSO, 2006. [cit. 27.6.2013]. <www.czso.cz/csu/2005edicniplan.nsf/p/1124-05>.

DUBSKÁ, D. *Domácnosti v ČR: příjmy, spotřeba, úspory a dluhy v letech 1993–2012* (Households in CR: Revenues, Consumption, Savings and Debts in years 1993–2012) [online]. Prague: CZSO, 2013. [cit. 7.7.2013]. <www.czso.cz/csu/2011edicniplan.nsf/publ/1159-11-n_2011>.

FISCHER, J. Ke čtvrtletním odhadům vývoje HDP (To the Quarterly Estimates of GDP Development). In: *Měříme správně HDP?* (sborník textů) (Do We Measure GDP Correctly? Texts collection), 2005b, 39, pp. 113–122.

FISCHER, J. Problémy měření HDP (Problems of GDP Measurement). In: *Měříme správně HDP?* (sborník textů) (Do We Measure GDP Correctly? Texts collection), 2005a, 39, pp. 11–20.

HRONOVÁ, S., FISCHER, J., HINDLS, R., SIXTA, J. *Národní účetnictví. Nástroj popisu globální ekonomiky* (National Accounts. A Tool for Describing of the Global Economy). Prague: C.H.Beck, 2009.

KUČERA, L. Změna stavu zásob a hospodářský cyklus (Change in Inventories and the Business Cycle). In: *The 13th Annual Doctoral Conference of the Faculty of Finance and Accounting, University of Economics, Prague* (sborník textů) (Texts collection), 2012, 13, pp. 435–446.

MANDEL, M., TOMŠÍK, V. Přímé zahraniční investice a vnější rovnováha v tranzitivní ekonomice: Aplikace teorie životního cyklu (Foreign Direct Investment and The External Balance in a Transition Economy: The Application of Live Cycle Theory). *Politická ekonomie*, 2006, 6, pp. 723–741.

PÁRAL, P. *Poslední muž v sedle* (The Last Man in the Saddle) [online]. Prague, 2001. [cit. 27.6.2013]. <euro.e15.cz/posledni-muz-v-sedle-817126>.

ŘÍMAN, M. Zahraniční investice ano, pobídky ne (Yes to Foreign Investment, No to Incentives). In: *Zahraniční investice. Cíl hospodářské politiky?* (sborník textů) (Foreign Investment. The Goal of Economic Policy? Texts collection), 2008, 65, pp. 11–16.

SINGER, M. *Hodnocení úspěšnosti transformace* (Evaluation of the Transformation Success) [online]. Prague: CNB, 2007. [cit. 27.6.2013]. <www.cnb.cz/miranda2/export/sites/www.cnb.cz/cs/verejnost/pro_media/konference_projevy/vystoupeni_projevy/download/singer_20070313_plzen_univerzita.pdf>.

SINGER, M. A Comparison of the Rates of Growth of Post-Transformation Economies: What Can(not) Be Expected from GDP? *Prague Economic Papers*, 2013, 1, pp. 3–27.

SIXTA, J., VLTAVSKÁ, K., ZBRANEK, J. Souhrnná produktivita faktorů založená na službách práce a kapitálu (Total Factor Productivity Measurement Based on Labour and Capital Services). *Politická ekonomie*, 2011, 5, pp. 599–617.

SPĚVÁČEK, V. Národohospodářská poptávka a makroekonomická rovnováha (Demand of the National Economy and the Macroeconomic Equilibrium). *Working Paper CES VŠEM*, 2006, 4.

ŠIROKÝ, M. *Řetězení stálých cen v národních účtech* (Constant Prices Chaining in the National Accounts) [online]. Prague: CZSO, 2004. [cit. 27.6.2013]. <www.mfcr.cz/cs/o-ministerstvu/vzdelavani/seminare/2004/rok-2004-podzimni-seminar-6690>.

TOMŠÍK, V. Zajistí přímé zahraniční investice prosperitu ČR? (Will Foreign Direct Investment Ensure the Prosperity of the CR?). In: *Zahraniční investice. Cíl hospodářské politiky?* (sborník textů) (Foreign Investment. The Goal of Economic Policy? Texts collection), 2008, 65, pp. 17–29.

VINTROVÁ, R. *Význam domácí poptávky v české ekonomice* (The Role of the Domestic Demand in the Czech Economy) [online]. Prague: Association of Regions of the Czech Republic, 2012. [cit. 27.6.2013]. <www.asociacekraju.cz/files/files/temata/Vyznam-domaci-poptavky26_3-RUZENA-VINTROVA.pdf >.

ZAMRAZILOVÁ, E. *Proč slábne spotřeba domácností? Pesimistický spotřebitel: světlo na konci tunelu* (Why the Consumption of Households weakens? Pessimistic Consumer: The Light at the End of a Tunnel) [online]. Prague: CNB, 2012. [cit. 27.6.2013]. <www.cnb.cz/cs/o_cnb/bankovni_rada/clenove_bankovni_rady/zamrazilova_projevy.html>.

ANNEX – TABLES

Table 1 Volumes at prices of 2005 (in millions of CZK)

Name	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
FCE (households)	1 215 809	998 970	1 043 921	1 057 426	1 096 216	1 146 311	1 230 786	1 250 154	1 233 845	1 261 629	1 272 973
FCE (government)	534 965	506 839	530 185	535 844	557 381	552 346	545 978	563 799	547 227	570 909	570 940
FCE (NPISH)	14 577	11 730	11 494	8 625	14 154	15 585	16 990	17 737	17 862	17 321	17 621
GFCF	521 629	393 324	434 066	465 255	519 922	641 227	699 570	654 121	647 511	634 198	675 357
CHII	–1 529	–22 579	–30 263	–17 164	17 153	12 921	38 150	8 790	–4 265	–5 843	24 496
NAoV	1 214	908	1 066	999	1 302	3 060	3 191	2 910	2 748	2 659	2 869
EX	596 871	523 702	569 158	649 508	674 382	801 809	853 765	936 442	1 045 373	1 104 835	1 295 868
IM	568 558	392 370	499 978	595 040	683 933	842 788	944 139	1 006 549	1 077 867	1 131 291	1 312 767
GDP	2 386 105	2 147 574	2 104 232	2 129 715	2 191 675	2 328 028	2 433 713	2 412 965	2 407 271	2 447 696	2 550 148
Residuum	**71 127**	**127 050**	**44 583**	**24 262**	**–4 902**	**–2 443**	**–10 578**	**–14 439**	**–5 163**	**–6 721**	**2 791**
Residuum/GDP	**3.0%**	**5.9%**	**2.1%**	**1.1%**	**–0.2%**	**–0.1%**	**–0.4%**	**–0.6%**	**–0.2%**	**–0.3%**	**0.1%**

Name	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
FCE (households)	1 315 419	1 356 289	1 427 673	1 472 666	1 515 680	1 581 021	1 645 963	1 695 423	1 698 399	1 714 924	1 723 883
FCE (government)	593 169	640 199	678 826	656 695	667 479	663 694	666 464	674 161	701 214	702 869	683 912
FCE (NPISH)	15 790	16 196	16 901	18 738	21 908	24 093	26 871	24 838	25 194	25 012	24 974
GFCF	705 428	732 560	736 940	758 808	804 594	851 276	963 948	1 003 509	892 622	901 714	904 928
CHII	29 772	12 059	–362	30 013	18 371	55 587	83 516	64 096	–42 000	–4 391	–455
NAoV	2 204	3 669	5 145	3 088	2 891	2 958	3 573	3 782	3 943	3 639	3 812
EX	1 446 283	1 481 239	1 594 027	1 811 639	2 025 872	2 309 507	2 570 441	2 671 441	2 381 014	2 751 106	3 012 290
IM	1 476 634	1 545 775	1 660 899	1 828 822	1 940 739	2 153 321	2 431 263	2 496 163	2 196 034	2 536 703	2 713 588
GDP	2 629 135	2 685 643	2 786 789	2 918 955	3 116 056	3 334 815	3 526 071	3 635 344	3 471 494	3 557 216	3 621 908
Residuum	**–2 296**	**–10 793**	**–11 462**	**–3 870**	**0**	**0**	**–3 442**	**–5 743**	**7 142**	**–954**	**–17 848**
Residuum/GDP	**–0.1%**	**–0.4%**	**–0.4%**	**–0.1%**	**0.0%**	**0.0%**	**–0.1%**	**–0.2%**	**0.2%**	**0.0%**	**–0.5%**

Source: CZSO (2013b), own calculations

Table 2 Balanced volumes at prices of 2005 (in millions of CZK)

Name	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
FCE (households)	1 230 613	1 024 364	1 052 829	1 062 125	1 095 282	1 145 870	1 228 862	1 247 520	1 232 933	1 260 447	1 273 433
FCE (government)	541 479	519 723	534 709	538 225	556 906	552 133	545 124	562 611	546 822	570 374	571 146
FCE (NPISH)	14 754	12 028	11 592	8 663	14 142	15 579	16 963	17 700	17 849	17 305	17 627
GFCF	527 981	403 322	437 770	467 323	519 479	640 980	698 476	652 743	647 032	633 604	675 601
CHII	–1 510	–22 005	–30 005	–17 088	17 138	12 916	38 090	8 771	–4 268	–5 848	24 505
NAoV	1 229	931	1 075	1 003	1 301	3 059	3 186	2 904	2 746	2 657	2 870
EX	604 139	537 015	574 015	652 394	673 808	801 500	852 430	934 469	1 044 600	1 103 800	1 296 336
IM	561 635	382 396	495 711	592 396	684 515	843 113	945 615	1 008 670	1 078 664	1 132 351	1 312 293
GDP	2 357 050	2 092 982	2 086 275	2 120 251	2 193 541	2 328 924	2 437 518	2 418 049	2 409 051	2 449 988	2 549 226

Name	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
FCE (households)	1 315 051	1 354 561	1 425 836	1 472 066	1 515 680	1 581 021	1 645 488	1 694 629	1 699 462	1 714 790	1 721 458
FCE (government)	593 003	639 384	677 953	656 427	667 479	663 694	666 272	673 845	701 653	702 814	682 950
FCE (NPISH)	15 786	16 175	16 879	18 730	21 908	24 093	26 863	24 826	25 210	25 010	24 939
GFCF	705 231	731 627	735 992	758 499	804 594	851 276	963 670	1 003 039	893 181	901 643	903 655
CHII	29 764	12 044	–362	30 001	18 371	55 587	83 492	64 066	–41 974	–4 391	–456
NAoV	2 203	3 664	5 138	3 087	2 891	2 958	3 572	3 780	3 945	3 639	3 807
EX	1 445 879	1 479 352	1 591 976	1 810 901	2 025 872	2 309 507	2 569 699	2 670 190	2 382 504	2 750 891	3 008 053
IM	1 477 047	1 547 744	1 663 036	1 829 567	1 940 739	2 153 321	2 431 965	2 497 331	2 194 660	2 536 901	2 717 405
GDP	2 629 870	2 689 064	2 790 375	2 920 144	3 116 056	3 334 815	3 527 089	3 637 046	3 469 321	3 557 494	3 627 002

Source: CZSO (2013b), own calculations

Table 3 Aggregate demand and supply at prices of 2005 (in billions of CZK, balanced)

Name	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Aggregate demand in bil. CZK	2 919	2 496	2 611	2 729	2 860	3 156	3 342	3 415	3 489	3 586	3 834	4 075	4 221	4 449	4 717	5 036	5 430	5 872	6 067	5 702	6 095	6 341
y/y change in bil. CZK	N/A	–423	114	118	131	296	186	73	74	96	249	241	146	228	268	319	394	442	195	-365	393	246
y/y change in %	N/A	–14.5%	4.6%	4.5%	4.8%	10.4%	5.9%	2.2%	2.2%	2.8%	6.9%	6.3%	3.6%	5.4%	6.0%	6.8%	7.8%	8.1%	3.3%	–6.0%	6.9%	4.0%
Domestic demand in bil. CZK	2 315	1 959	2 037	2 076	2 186	2 355	2 489	2 481	2 445	2 482	2 538	2 629	2 742	2 857	2 906	3 010	3 120	3 302	3 396	3 320	3 344	3 333
y/y change in bil. CZK	N/A	–355	77	39	109	169	135	–9	–36	37	56	91	113	115	49	104	110	182	94	–77	25	–11
y/y change in %	N/A	–15.4%	4.0%	1.9%	5.3%	7.7%	5.7%	–0.4%	1.4%	1.5%	2.3%	3.6%	4.3%	4.2%	1.7%	3.6%	3.7%	5.8%	2.8%	–2.3%	0.7%	–0.3%
Foreign demand in bil. CZK	604	537	574	652	674	802	852	934	1 045	1 104	1 296	1 446	1 479	1 592	1 811	2 026	2 310	2 570	2 670	2383	2 751	3 008
y/y change in bil. CZK	N/A	–67	37	78	21	128	51	82	110	59	193	150	33	113	219	215	284	260	100	–288	368	257
y/y change in %	N/A	–11.1%	6.9%	13.7%	3.3%	19.0%	6.4%	9.6%	11.8%	5.7%	17.4%	11.5%	2.3%	7.6%	13.8%	11.9%	14.0%	11.3%	3.9%	–10.8%	15.5%	9.3%

Source: CZSO (2013b), own calculations

Table 3 Aggregate demand and supply at prices of 2005 (in billions of CZK, balanced) – continuation

Name	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
Aggregate supply in bil. CZK	2 919	2 496	2 611	2 729	2 860	3 156	3 342	3 415	3 489	3 586	3 834	4 075	4 221	4 449	4 717	5 036	5 430	5 872	6 067	5 702	6 095	6 341
y/y change in bil. CZK	N/A	–423	114	118	131	296	186	73	74	96	249	241	146	228	268	319	394	442	195	-365	393	246
y/y change in %	N/A	–14.5%	4.6%	4.5%	4.8%	10.4%	5.9%	2.2%	2.2%	2.8%	6.9%	6.3%	3.6%	5.4%	6.0%	6.8%	7.8%	8.1%	3.3%	–6.0%	6.9%	4.0%
Domestic supply in bil. CZK	2 357	2 114	2 115	2 136	2 175	2 313	2 396	2 406	2 411	2 453	2 522	2 598	2 673	2 786	2 887	3 095	3 276	3 440	3 569	3 507	3 558	3 624
y/y change in bil. CZK	N/A	–243	1	21	39	138	83	10	4	43	69	76	75	112	101	208	181	164	129	–62	51	65
y/y change in %	N/A	–10.3%	0.1%	1.0%	1.8%	6.3%	3.6%	0.4%	0.2%	1.8%	2.8%	3.0%	2.9%	4.2%	3.6%	7.2%	5.9%	5.0%	3.8%	–1.7%	1.5%	1.8%
Foreign supply in bil. CZK	562	382	496	592	685	843	946	1009	1 079	1 132	1 312	1 477	1 548	1 663	1 830	1 941	2 153	2 432	2 497	2195	2 537	2717
y/y change in bil. CZK	N/A	–179	113	97	92	159	103	63	70	54	180	165	71	115	167	111	213	279	65	-303	342	181
y/y change in %	N/A	–31.9%	29.6%	19.5%	15.6%	23.2%	12.2%	6.7%	6.9%	5.0%	15.9%	12.6%	4.8%	7.4%	10.0%	6.1%	11.0%	12.9%	2.7%	–12.1%	15.6%	7.1%

Source: CZSO (2013b), own calculations

Corporate Sector in the Czech Republic: What is the Role of Ownership in the Dissolution of Company?[1]

Drahomíra Dubská[2] | *University of Economics; Czech Statistical Office, Prague, Czech Republic*

Abstract

Empiric analysis of the sector of non-financial corporations in the Czech Republic shows history of their terminations during the period 1995–2011. The purpose was to observe, what differences in the evolution of terminations (bankruptcies) of companies are connected with the predominant form of their ownership (only legal persons were analysed). Therefore, selected methods of the companies´ terminations were monitored according to a group of legal persons with prevailing domestic private capital and also group with predominance of the public capital. Also, the firms under foreign control were analysed. The companies with prevailing private domestic capital have dominated due to their largest number. However, the foreign-controlled companies have showed faster pace of the terminations during the analysed period and also their greater sensitivity to the economic cycle has been demonstrated. Liquidation was the most common method of the companies´ termination (roughly two-thirds of the total number of terminations in 2011). Dissolution without liquidation with successors represented one tenth of the terminations in the segment with predominance of private national capital and 15% in the segment of foreign-controlled companies.

Keywords

Non-financial corporations, ownership, liquidations, bankruptcy

JEL code

M2, O12, L16, L26

INTRODUCTION

Besides the view through branches of the economy it is possible to observe the sector of non-financial corporations also from point of view of their ownership. This is not very often used preview but it can provide especially in the long-term analyses of transitive economies very interesting information

[1] This paper is designed as one of the outputs of the research project "The development of the transaction costs of Czech businesses in bankruptcy proceedings, the possibility of their reducing on the level common in the EU by improving legislation, the possibility of improving statistics insolvency proceedings and creating a model of financial fragility of households" which is registered with the Technology Agency of the Czech Republic under the registration number TD 010093. The international scientific team is coordinated by the University of Economics in Prague.

[2] University of Economics, Nám. W. Churchilla 4, 130 67 Prague, Czech Republic; Czech Statistical Office, Na padesátém 81, 100 82 Prague, Czech Republic. E-mail: dubd00@isis.vse.cz, drahomira.dubska@czso.cz.

including the lifetime of companies by type of ownership. This analysis regards the institutional sector of non-financial corporations which is devided intothe groups of the firms with the prevailing ownership of the capital of public sector of the Czech Republic and of so-called private national corporations and also private foreign-controlled companies. The choice is primarily focused on legal persons because of their weight in producing gross value added in the Czech economy.

As regards the shares of institutional sectors in the Czech economy, in 2011 the firms under foreign control accounted for 68% of total turnover of the non-financial corporations and in the manufacturing even 83%. The private national corporations participated just only from one fifth in the mentioned total turnover and the rest (ie. 12%) represented the firms with the prevailing ownership of the capital of public sector. If we watch on the proportion on employment, almost three fifths (59%) of employed persons of the non-financial firms were working in the foreign-controlled segment while just only one fifth in the private national segment. The firms with the prevailing ownership of the public sector capital of the Czech Republic employed 17% of the total number of employed persons of the non-financial sector. Finally, the firms with the prevailing ownership of the capital of public sector of the Czech Republic were providing the job to 17% of the total number of employees of the non-financial sector. The proportions between the achieved turnover and employment, of course, reflects the level of labor productivity in the analyzed segments.

Therefore, the results of the analysis must be viewed through the prism of the position of each institutional sector in the Czech economy.

METHODOLOGY

Data regarding numbers of the corporations which ceased their acitivites (ie. deaths, terminations of firms, their downfall, extinction, the companies which were on the decline or ended their activities) during 1995–2011 were available by selection from the register of economic entities, ie. Business Register (BR) which is created and managed by the Czech Statistical Office (CZSO).[3] However, the CZSO is not primarily the place which should show the state of the company in the phase of insolvency and even in the foreseeable future it will not be possible to concentrate such information there. Therefore, the company at the stage of insolvency can be included into the Business Register of this official national statistics only when any form of death of such company is notified. Then, the statistics is available with the code of list defining the form of deaths. The code of list in the CZSO defines twelve variants of possible deaths of legal (artificial) persons and household-trades (i.e. enterprising individuals, natural persons).[4]

There is no ambition of this article to analyse the forms of death of the number of registered economic subjects in the Czech Republic in total due to considerable robustness of the topic and the size of the database. Therefore, for the purpose of this analysis non-financial corporations by the type of ownership were chosen only and the attention was concentrated on legal persons. As well only some forms of deaths of legal persons were selected of the institucional sector of non-financial corporations which are devided by the type of ownership in the system of the ESA 95 (European System of Accounts). They are three types of entities in this system: Public non-financial corporations, National private non-financial corporations

[3] Thus, synonymous with defunct in this analysis is extinction, dissolution, termination, death, annuelment. In this sense a defunct company is the company which had ceased its activity and was deleted from the Commercial Register.

[4] The code of list defines these forms of the death of legal persons and household-trades (enterprising individuals): 00 Not identified, 01 Annuelment of legal entities by liquidation, 02 Cancellation of legal entities without liquidation with one successor, 03 Cancellation of legal entities without liquidation with more successors, 04 Cancellation of legal persons without liquidation without successors, 05 Notification of the household-trade about termination, 06 The decision to withdraw authorization for a household-trade, 07 The death of household-trade , 08 Moving out of the district, 09 Decision because of the non re-registration, 10 Unauthorized extradited Identification Numbers Organisation (IC), 11 Ending of Registration of the duplicate Identification Number (IC).

and Foreign-controlled non-financial corporations (in the next text and also in the chart labels, the titles of these entities are presented in abbreviated form as "public", national private" and "foreign-controlled"). Only these forms of deaths of companies are considered in this article:

- Annuelment of legal persons by liquidation,
- Annuelment of legal persons without liquidation,
- Annuelment of legal persons without liquidation with one successor,
- Annuelmentof legal persons without liquidation with more successors.

In order to capture the longest time horizon and to track changes in selected sector of the firms, the interval from 1995 to 2011 was chosen. Relations in a broader sense enabled through data within the Business Register (BR) are analysed of period 2005–2010.

1 PROPORTION OF THE INSTITUTIONAL SECTORS IN THE BUSINESS REGISTER OF THE CZECH REPUBLIC

In 2008, the number of the economic subjects registered in the Czech Republic[5] for the first time exceeded the level of 2.5 million entities. At the end of 2010 a total 2 637 thousand entities were registered, according to official data available. Regarding the numbers, household-trades (ie. individuals or in other words entrepreneurs with Identification Number) dominate the file of registered subjects. Their number exceeded the level of two milion entities at the end of 2010. Regarding others, there was a total of 2 164 companies from the financial sector[6] and 17 956 from the government sector which includes also the municipalities, cities and counties. Economic entities of institutional sector named Non-profit institutions serving household (NPISHs) is also a relatively high number (118 848 at the end of 2010) while non-residents registered as an economic entities in the Czech Republic were only hundreds and reached even only 0.1 ‰ of the total number of entities listed in the BR.

The institutional sector Non-financial corporations (in the text of this post as "corporates") participated in the total number of economic entities in the Czech Republic less than one fifth (18.9%) at the end of 2010. However, its share has increased over the time to 15% in 2005. At the end of 2010 almost half a milion (497.8) of non-financial corporations existed in the Czech Republic.

In 2010, full three-quarters of registered economic entities in the Czech Republic, ie. 75.9%, were entrepreneurs in the household sector, ie. household-trades. Their proportion, on the contrary, decreased during the second half of last decade because in 2005 it represented 78.6% of total registered economic entities. This may be obviously explained by development of the economic cycle at this time which passed through reversals. The other institutional sectors of the Czech economy made only a very little part in total number of economic entities – in total, according to data from the end of 2011 it was only 5.3%.

In spite of the mentioned decrease of the share of enterpreneurs in household sector (household-trades) during period 2006–2010, their absolute number has increased (+6.6%). However, the figures regarding registered non-financial companies has increased much more rapidly (+38.7%). As for the other institutional sectors of the Czech economy, the number of financial institutions grew only slightly (2%) and the number of institutions in the government sector and the non-profit sector on the contrary declined (–2.8% and –9.7%, respectively).

---

[5] The Business Register is not a perfect database as regards the accuracy and timeliness of the classification of the economic entities according various criteria (incl. problem with the principal activity which may change over time and the companies often do not give the important information to the database administrators note immediately). However, it is the only source of official statistics of this magnitude.

[6] Institutional sector of financial institutions consists of the central bank, other monetary-financial institutions, ie. above all commercial banks and savings construction banks, credit unions and other financial intermediators as the leasing and factoring companies and the hire purchase companies). In addition to these entities, this sector includes so called financial auxiliaries such as brokers and also the insurance companies.

1.1 The structure of non-financial corporations in the BR by form of ownership

Looking at the structure of non-financial companies registered in the BR by form of ownership we can see long-term growth of number of private national corporations, but especially of foreign-controlled companies. During 2005–2010 in both segments their numbers increased by 31.2%, and by 52.7%, respectively. This trend reflects the slowly lingering period when the companies with the prevailing public ownership have been long-term retreating due to privatisations. In spite of it also during the second half of last decade their number decreased by more than one fifth (–21.5%) though the main stream of privatisation had passed significantly before.

With the rapid growth in the number of the foreign-controlled companies registered in the BR, their weight reached two-fifths (40.1%) of the total number of registered non-financial corporations in 2010 when in 2005 it was 36.4%. The increase in their number during the robust boom of the Czech economy in 2005–2008 was likely erased on a larger scale in 2009, the year of the economic crisis. The post-crisis year 2010 showed the decreasing share of number of foreign-controlled companies (by 0.6 pp) but it had been caused above all by development of their numbers in the segment households-trades. Probably it was related to the deterioration in the labour market. The employment is the macroeconomic indicator for which it is characteristic the lagging of pace of development (in both directions) against the development of real economy according changes in GDP. Therefore, it is possible to suppose that the unsatisfactory labor market had made a lot of non-residents working in the CR in 2010 on the Trade Licencing to abolish their entrepreunership.

Private national corporations still dominate the number of non-financial companies in the CR in stratification according type of ownership. Nevertheless, their share weakens and in 2009, for the first time, fell below the level of three-fifths (58.9%). In 2010 the share of the private national non-financial companies approached again to the three-fifths (+0.7 pp) due to weakened dynamics of the numbers of registered foreign controlled firms. The number of companies controlled by the public sector in terms of ownership has become practically negligible in the Czech Republic. In 2009 and 2010 it represented only 0.4% of the total number of registered non-financial corporations (in 2005 the share was still 0.7%).

2 THE OWNERSHIP STRUCTURE OF THE DEFUNCT LEGAL ENTITIES IN THE CORPORATE SECTOR IN THE CZECH REPUBLIC

Firms owned by the public sector are overwhelmingly legal entities in terms of types of companies in the sector of non-financial corporations. Also, the private national corporations has taken predominantly form of legal entities (Table 1). For private firms under foreign control, this effect is not noticeable, but it should be borne in mind that while domestic households-trades are recorded in the statistics of BR in the household sector, in the case of non-residents´ household-trades (ie. those which are foreign-controlled), these entities are considered as a part of non-financial sector (not as a part of household sector). Logically, the number of such companies (under foreign-controlled enterprising individuals) then increases the number of entities in this sector. Consequently, a change in the numbers of these non-residents working in the Czech Republic on the Trade Licensing obviously affects the figures relating to the numbers of the non-financial entities.

Table 1 Representation of legal persons in the total number of registered businesses in the non-financial sector in 2009

	Legal persons	Entities in total	Share
Public sector ownership	1 944	1 995	97.4%
Private national corporations	268 069	277 862	96.5%
Foreign-controlled companies	79 417	191 829	41.4%

Source: CZSO, withdrawals from database BR

2.1 Comparison of dissolution (termination) of legal persons in the sector of non-financial corporation

Graphical analysis of the dissolution of legal persons in the Czech Republic in the years 1995-2011 shows (Figure 1), that the largest numbers of terminations occurs in the segment of private national companies. This is not surprising given that this segment has the greatest weight in the structure. Numbers of dissolutions of legal persons with a predominance of private national capital were below the trend line in period 1997–2002, according to the equalizing line of regression analysis with a relatively high reliability coefficient. So it was in a period of relatively strong wave of privatization in the Czech Republic. Also in 2005 and 2006 the number of dissolved companies in the Czech Republic was lower, i.e. below the trend line. It was the period when the Czech economy grew at its fastest rate in its history. On the contrary, higher numbers of dissolutions of the legal persons with the predominance of private national capital compared to the trend of total period are seen in years around the currency crisis at the turn of the 90th years (Figure 1). The same development, although somewhat weaker but consistent, can be observed in the pre-crisis period and in subsequent years at the end of the last decade and the beginning of this decade.

For terminations of legal persons under foreign control a similar trend is monitored like in the analyzed segment of firms with prevailing domestic capital. But regression analysis (again, with a relatively high reliability coefficient) also shows higher numbers of dissolutions from 2007 to 2011. Nevertheless, during relatively long period 1999–2006, the numbers of legal persons´ terminations in the segment of firms under foreign control were slightly lower than would correspond linear equalizing.

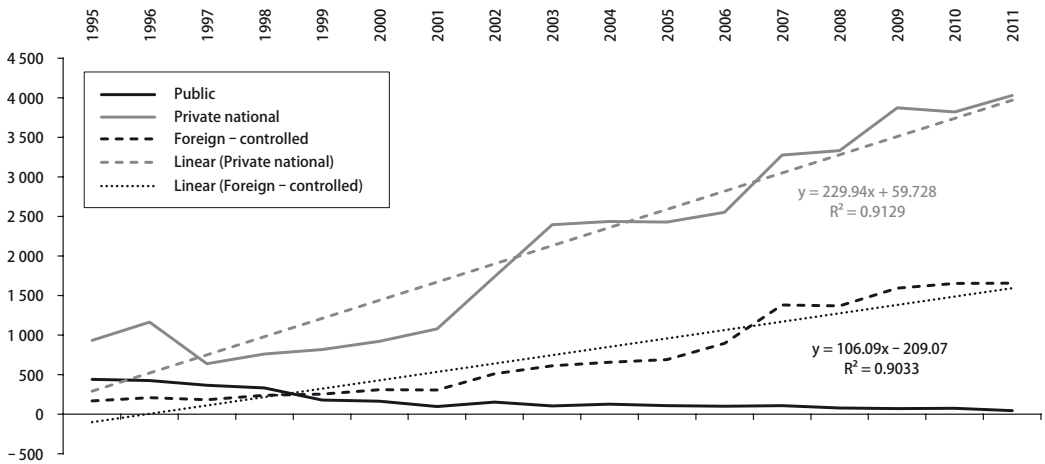
2.1.1 Dynamics of extinction and feedback on the performance of the economy

If we will use the annual changes in numbers of defunct firms for comparison legal persons´ terminations in various forms of ownership, ie. how many of them were more or less than a year ago, the results are different compared to absolute data which we observed in Figure 1 – the legal persons under foreign control were disappearing more quickly. During the period 1995–2011 their numbers were going down by on average 17% annually while in case of the private national firms by 12%.

On the contrary, the deaths of the legal persons with dominancy of the capital of the public sector were decreasing by 9.9%, on average, against previous year. However, partly it is attributed to the significant fading of this type of companies from the map of economic subjects registered in the Czech Republic. Another part of this explanation may lie in certain stability of these companies arising from their position in the industry with a predominance of public ownership. Namely they were able in spite of the crisis to overcome relatively very well the adverse conditions in demand during 2009. The output of electricity sector grew due to the results of company Ceske energeticke zavody (CEZ) and a performance of forestry increased also with the company Lesy CR. In health care gross added value rose, year-on-year, by about one tenth in 2009. The increase was recorded also in land transport and transport via pipeline as well as in the performance of the Ceska posta (Czech Post).

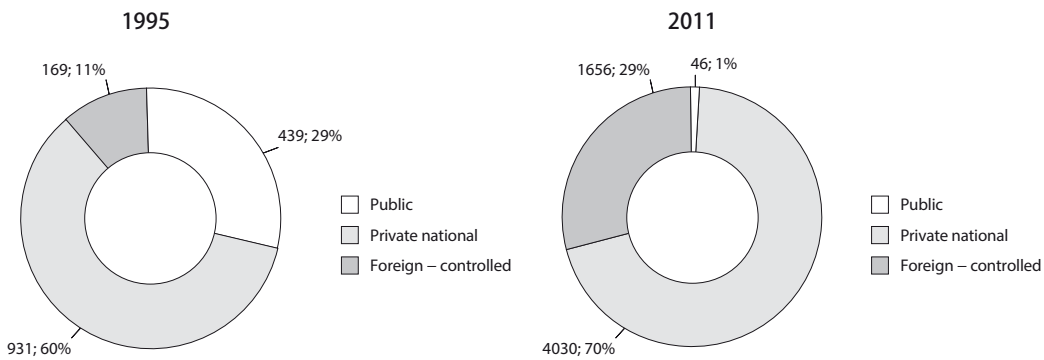
The correlation between the pace of economic growth on one hand and the intensity of the dissolving of legal persons on the other hand – according to the level of correlation coefficient – confirms with a relatively high probability the hypothesis that the faster an economy grows, the lower should be the activity in the deaths of the companies. This applies to a greater degree to legal entities under foreign control (with correlation coefficient –0.65), but relatively significantly also to legal persons in corporate sector which are controlled by private national capital (with correlation coefficient –0.61). The specificity of the results of the correlation analysis for the legal persons owned by the public sector consists in the fact that the number of deaths has been reduced significantly due to the decrease in the number of these entities objectively (of privatization).

Figure 1 The numbers of deaths of legal entities by type of ownership (1995–2011, end of the year)



Source: CZSO, withdrawals from database BR, author's own calculations

Figure 2 Share of dissolution of the legal persons in the corporate sector by type of ownership (absolute data; % of total number of dissolution of legal persons in corporate sector according three analyzed forms of ownership)



Source: CZSO, withdrawals from database BR, author's own calculations

3 MAIN FORMS OF DISSOLUTION OF LEGAL PERSONS IN THE CORPORATE SECTOR OF THE CZECH REPUBLIC

Now, we focus on the main ways of the deaths of the companies according the Nomenclature of the BR concerning legal persons (more on the Methodology citations, footnote 4): ie. terminations of legal persons through liquidation or without liquidation, namely with one successor or with multiple successors. The selection was also done in the "Not identified" where the numbers of dissolutions were very strong in 1995–1996 (about 1 500 extinction), although since 1999 strongly reduced to a few dozen per year. The rest of total numbers of dissolution of legal persons, taking into account the above-defined forms of extinction, was the dissolution of the legal entity without liquidation without successors.

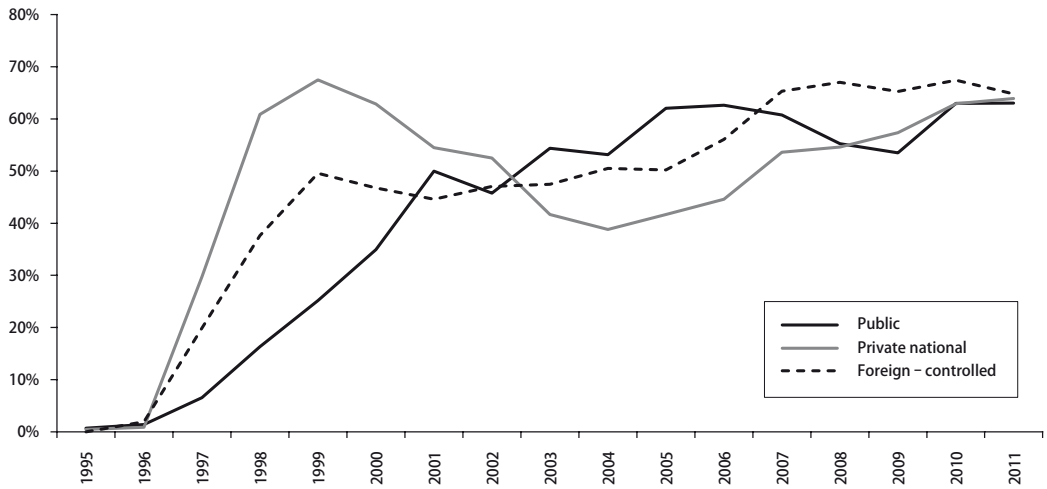
3.1. Dissolution of legal person through liquidation

In the Czech economy the termination of the legal persons through liquidation has gradually become a predominant mode of termination of legal entities in each of the forms of their ownership. Figure 3

depicts the massive increase in the share of extinction of private national companies through liquidation during period 1997–1999. At that time the Czech Republic has undergone a shallow recession after the currency crisis with a decline in GDP in real terms by 0.9% in 1997 and 0.2% in 1998. But, undoubtedly, also the important fact played a role and namely that in this period the number of undetected ways of death declined significantly. It naturally resulted in an increase in recorded cases of termination of liquidation, as well as dissolution without liquidation. Also, annual dynamics of terminations through liquidation seemed to be very strong during this period. In 1996 only 10 private national legal persons ceased to exist through liquidation according data in the BR. But in 1997 already it was 189 and in 1998 even 464 dissolutions through liquidations.

At the end of the analyzed period, in 2011, a total 2 575 of legal persons in private ownership ended through liquidations. It was almost two-third of total number 4 030 dissolutions of legal persons of this form of ownership that disappeared in the Czech Republic. In the case of legal persons under foreign control was a similar proportion of subjects in the numbers 1 073 against 1 656 defunct legal persons. In the segment of public ownership of legal entities, then it was 29 against 46 defunct legal persons. In all cases, liquidations in 2011 accounted for almost two-thirds of dissolutions of legal persons in each of the analyzed forms of ownership compared with about one-half at the beginning of the last decade.

Figure 3 Termination of legal persons through liquidation (the share on the total deaths of legal persons according respective forms of ownership, in %)



Source: CZSO, withdrawals from database BR, author's own calculations

3.2 Dissolution of legal persons without liquidation

In 2011, total 357 legal entities with the control of private national capital were dissolved without liquidation with one successor or without liquidation with more successors. In the segment of legal persons under foreign control it was 225 subjects and six legal entities with a prevailing of public ownership. At the beginning of the last decade, ie. in 2011, analogous numbers were 142, 39 and 19. So, the numbers of dissolutions with successors showed the fastest growth in the segment of the foreign-controlled.

In 2011, the distinctive retreat of the number of dissolutions without liquidation in favor of the dissolutions with liquidation was seen. In the segment of legal entities owned by the private national capital it was 357 against 2 575, ie. exceeding more than 7 times. In the segment of legal

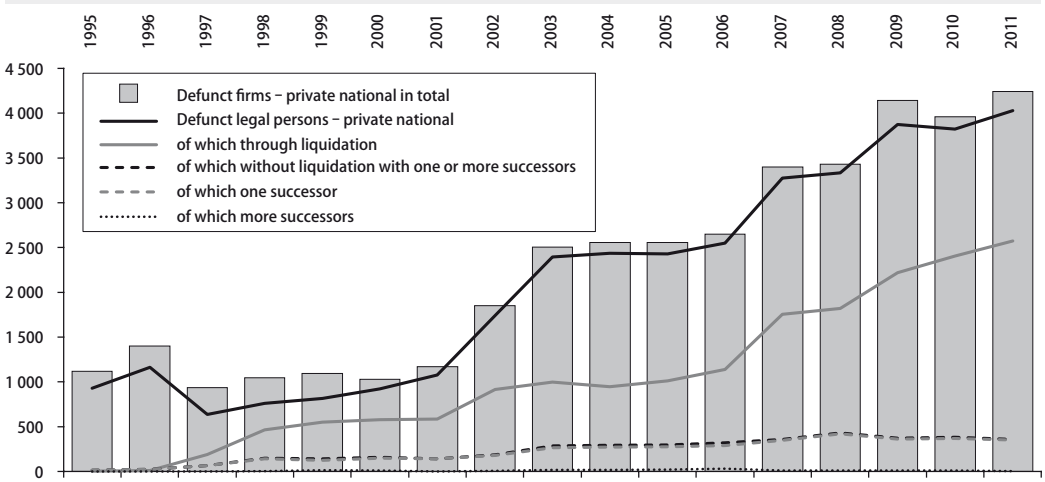
persons under foreign control the comparable data (225 against 1 073) meant the increase in the number by approximately 5 times.

Cases when the dissolving company (legal person) was taken over by only one successor were far more common than with more successors. Cases with more successors were only of the order of units among the firms under foreign control with exception of the year 2000 (13 cases). For companies which were controlled by the public capital these cases were more frequent during period 1997–2000. It was an average of 19 per year but for example 31 in 1997. During 2009–2011 it was not a single case. Also the numbers of dissolutions without liquidation in this period were negligible in this segment (11 cases in 2009, 8 in 2010 and 6 cases in 2011). Firms controlled by private domestic capital which were ceasing to exist without liquidation with more successors showed in the numbers of cases per year since 2006 a gradual decrease (from 32 to 5 cases in 2011).

3.3 Overview of deaths of legal entities in the sector of non-financial corporations by ownership and development during period 1995–2011

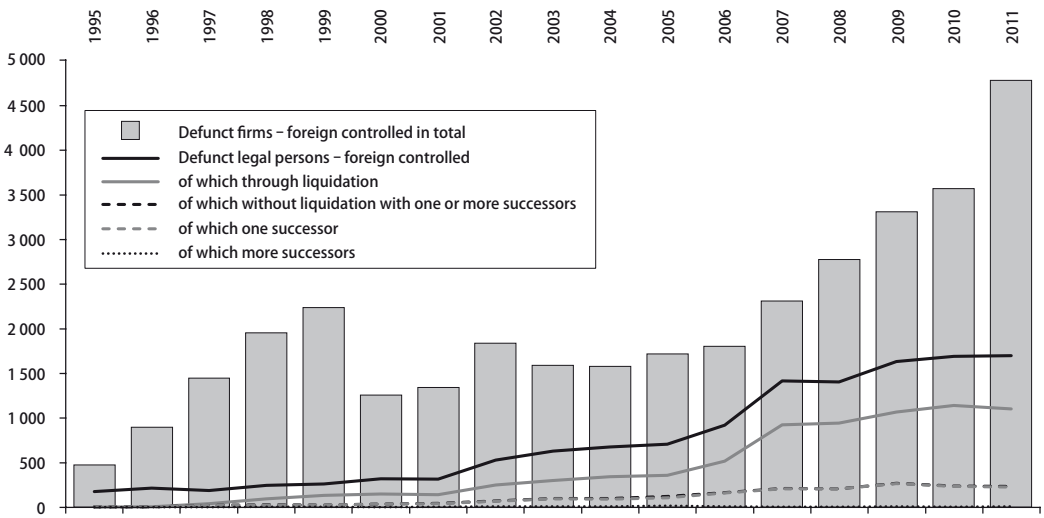
A comprehensive view on the development in the period 1995–2011 according to the analysed types of the dissolution of legal persons according to their form of ownership is provided by a graphical analysis (Figures 4, 5, and 6). The curves in these graphs illustrate the above-mentioned conclusions with the important role of foreign household-trades (non-residents working in the Czech Republic on the Trade Licensing) in the total number of dissolution of companies under foreign control (Figure 5). Roughly, since 2001, the trend in development of the dissolution of legal entities under the control of a private national capital through liquidation is practically identical to the development of these total deaths in this segment (ie. legal persons plus household-trades with dominance of private national capital) as shown on Figure 4. Figure 6 shows a decrease in the numbers of termination of the companies which were controlled by public capital. This is primarily due to the decreasing numbers of subjects with this form of ownership.

Figure 4 Development of dissolution of legal persons with a predominance of private national capital and their selected forms (numbers at the end of year)



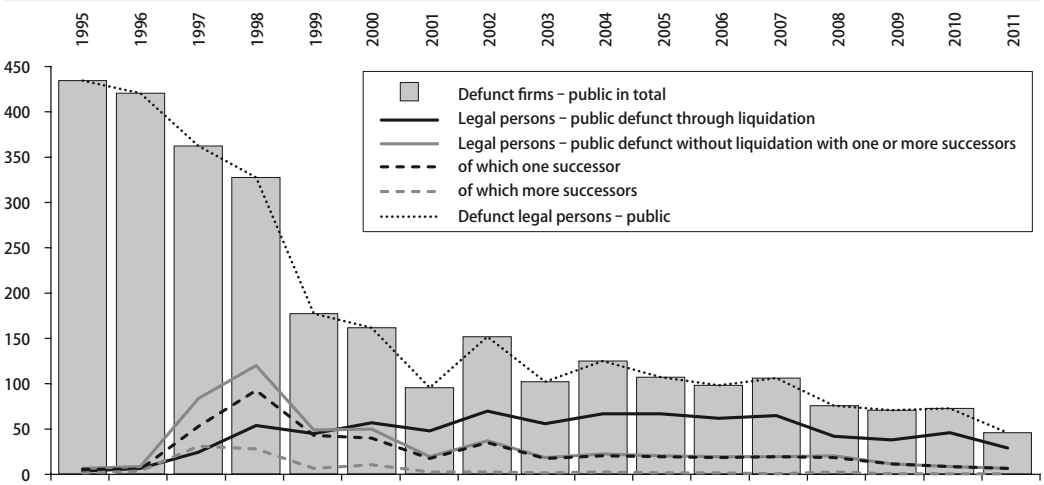
Source: CZSO, withdrawals from database BR, author's own calculations

Figure 5 Development of dissolution of legal persons under foreign control and their selected forms (numbers at the end of year)



Source: CZSO, withdrawals from database BR

Figure 6 Development of dissolution of legal persons with predominance of public capital (numbers at the end of year)



Source: CZSO, withdrawals from database BR

CONCLUSION

In the period 1995–2011 the organizational structure of the corporate sector in the Czech Republic (non-financial corporations) passed through significant changes in the ownership of companies. These were caused both by the privatization into domestic hands, and, secondly through a strong inflows

of foreign direct investment. Due to this a very strong segment of the firms under foreign control has appeared in the Czech Republic.

Most numerous group of legal entities under the control of domestic private capital showed logically the highest numbers of dissolution. However, according to annual dynamics it is possible to observe that faster, ie. an average of 17% per year, numbers of termination of foreign-controlled companies increased compared to the private national (+12%).

Termination through liquidation has become the predominant mode of termination of legal entities in each of forms of non-financial companies´ ownership in the Czech Republic. In 2011, less of companies under foreign control disappeared through liquidation than in 2010 (this was the first annual decline since 2002). For private domestic firms, the number of dissolution through liquidations grew in 2011. But this growth was the slowest since 2009 in which 400 private national legal persons ceased to exist. In 2011, roughly two-thirds of all cases of termination of companies in each of three analyzed forms of ownership ceased through liquidation.

Dissolution without liquidation with one or more successors is not so frequent in the Czech Republic, especially with more successors. During the period 2000–2011, the numbers regarding dissolution of legal persons without liquidation with the successors represented annual average of less than 12% of the defunct legal persons with domestic private capital and in case of legal persons under foreign control roughly 15%.

Correlation of deaths of companies with the development of business cycle in the Czech Republic is relatively significant.

References

BALLEISEN, E. *Navigating Failure: Bankruptcy and Commercial Society in Antebellum America.* Chapel Hill: University of North Carolina Press, 2001, p. 322. ISBN 0-8078-2600-6.

CZSO. *Annual national accounts* (database) [online]. Prague: Czech Statistical Office. <http://apl.czso.cz/pll/rocenka/rocenka.indexnu>.

CZSO. *Institutional sector accounts* (database) [online]. Prague: Czech Statistical Office. <http://apl.czso.cz/pll/rocenka/rocenkavyber.sek>.

CZSO. *Business Register* [online]. Prague: Czech Statistical Office. <http://www.czso.cz/csu/redakce.nsf/i/organizacni_statistika>.

Demografie podniků v ČR – výsledky za roky 2000 až 2005 (Demography of enterprises in the Czech Republic – results for the years 2000–2005). Prague: Czech Statistical Office, 2008.

DUBSKÁ, D. Impact of the Economic Crisis on the Institutional Sectors of the Czech Economy. *Statistika: Statistics and Economy Journal*, No. 4/2011. pp. 4–21.

DUBSKÁ, D. *Krize: kdo z ní vyšel nejlépe?* Analýza institucionálních sektorů české ekonomiky. Vystoupení na semináři (Analysis of the institutional sectors of the Czech economy. Presentation at the seminar). Prague: Czech Statistical Office, February 2011.

KISLINGEROVÁ, E. *Podnik v časech krize* (The company in times of crisis). Prague: Grada, 2009, p. 208. ISBN 978-80-247-3136-0.

KORÁB, V., MIHALISKO, M., VAŠKOVIČOVÁ, J. *Založení a řízení podniků* (Establishment and management od enterprises). Brno: CERM, 2008, pp. 3–155. ISBN: 978-80-214-3792- 0.

NEČADOVÁ, M., BREŇOVÁ, L. Změna ve výkonnosti firem v podmínkách krize a jejich adaptace na změněné tržní podmínky (Changes in the performance of companies in times of crisis and their adaptation to the changed market conditions). *Ekonomika a management*, 4/2010, pp. 2–17.

Statistics of Remittances in the Czech Republic

Martina Šimková[1] | *University of Economics; Czech Statistical Office, Prague, Czech Republic*
Jaroslav Sixta[2] | *University of Economics; Czech Statistical Office, Prague, Czech Republic*

Abstract

Remittances represent transfers of earned money of foreigners to home country and they belong to often discussed topics in the connection with migration issues. The reason is the recent increase of the number of migrants and the amount of sent remittances. Since the Czech Republic became the immigration country, remittances have gained the importance. Many foreigners come to the Czech Republic because of work or study. However, many of them send earned money back to their country of origin, to their families and relatives. The information about the foreigners is crucial for describing their behaviour (incomes, consumption expenditures) and recording the transactions in national accounts and balance of payments. Permanent lack of data sources causes problems with such statistics. The aim of the paper is the description and interpretation of the issue of remittances. Procedures that are used for these estimates are briefly described with respect to the users' needs. The description is mainly focused on the estimates of numbers and structures of foreigners, covering the length of stay, economic activity, their behaviour and estimation of sent remittances.

Keywords

Migration, remittances, national accounts

JEL code

F22, F24, J61

INTRODUCTION

Remittances are defined as migrants' transfers of money to their country of origin (home country) resulting from their temporary or permanent incomes. The increasing attention is going to be put on remittances; currently many migration and development studies deal with cross-border money transfers of migrants. The main reason of these studies is the fact, that migration and remittances are generally considered as efficient instrument for development of developing countries but it has also impact on the labour market. A host developed country (usually with worsening demographic structure) is getting a work force (temporary or permanent) that will help to satisfy the needs of the pensioners (Šimková, Sixta, 2013). These issues are also important for construction of national accounts and balance of payments. This has become an important issue in the EU and it covers new member states, as well. The number of immigrants and the amount of sent remittances have been still increasing since 1990s.

[1] University of Economics, Faculty of Informatics and Statistics, nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: martina.simkova@vse.cz. Author is working also at the Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic.

[2] University of Economics, Faculty of Informatics and Statistics, nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: sixta@vse.cz. Author is working also at the Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic.

On the other hand, remittances are causing the outflow of funds from economically developed countries and this outflow became noticeable.

At the national level, the importance of remittances is naturally connected with their position in the economy. Remittances influence social and economic situation, not only in the country of origin, but also in the host country and therefore migration and remittances are reflected in the national accounts as well as in the balance of payments of the countries. Their impact on social condition of households is evident in the current world. We can find very persuading examples not only for African or Asia refugees but also for Ukraine, former Yugoslavian republics, Caucasus, etc. In the Czech Republic, the remittances gained its importance primarily because of the positive migration balance (CZSO, 2013a). It is obvious that labour migration plays an important role in maintaining stable levels of the workforce; especially in construction or in low-level paid services (e.g. cleaning in hospitals). Since mid-90s, these jobs have been very often occupied by people from former Soviet bloc. The structure of immigrants also influences the human capital of the country in terms of their education and profession (Mazouch, Fischer, 2011). Currently this breakdown is still far from our possibilities and we concentrate on country of origin only. Obviously, a part of created values outflows abroad because this was in fact (at least for some group of workers) one of the most important motivation factors for arrival.

1 DATA AND METHODOLOGY

This area is still being developed as the migration issues going to be more and more important. In fact, the definition of remittances is not uniform and recently has been still changing. Currently, the Czech Statistical Office (CZSO) uses the definition given by the International Monetary Fund (IMF). The IMF (2009) derives remittances from two items of the balance of payments:

- Revenues earned by workers in economies without permanent resident (eventually from employers seated abroad).
- Transfers from residents of one economy (the residents - see below) to residents of other economies.

For quantification of the total amount of remittances is necessary to know the amount of the income of migrants. The estimation of incomes and remittances is based on the behaviour and structure of foreigners. We suppose that the amount of foreigner's income depends on the type of their activity, whether they are employees or employers (entrepreneurs), whether they work legally or illegally. Similar assumption is applied on the consumption behaviour; the consumption expenditures depend on the income, length and purpose of their stay.

The foreigners (according to current national account standard ESA 1995) are distinguishing primarily by the length of stay:

- Residents – economic entities, living and working in the Czech Republic for at least one year or longer.
- Non-residents – economic entities, living and working in the Czech Republic less than one year. This category also includes the cross-border workers, seasonal workers, foreign students studying in the Czech Republic, foreigners working in the Czech embassies abroad.

The breakdown of foreigners between residents and non-residents is crucial for correct capturing and quantifying all of accompanying flows. Subsequently, the nature of their economic activity has to be taken into account and foreigners are broken down by following categories:

- economically active persons:
  - employees (legal x illegal),
  - employers (entrepreneurs),
- economically non-active persons:
  - students,
  - others.

According to current rules, entrepreneurs are always considered as residents only. The last breakdown focuses on the country of origin of foreigners. We suppose that people from poorer countries have slightly different behaviour even within each group (e.g. a difference between students).

1.1 Data sources on numbers of foreigners

Unfortunately, the number and structure of foreigners living in the Czech Republic is not recorded uniformly. There is no central database collecting all information about foreigners. There are three independent resources for statistics of foreigners in the Czech Republic:

1) The Ministry of the Interior (MOI) – Alien police inspectorate records the total number of foreigners by length of stay and country of origin.
2) The Ministry of Labour and Social Affairs (MLSA) – employment bureau issue work permits.
3) The Ministry of Industry and Trade (MIT) – monitoring of trades licenses.

The estimate of the number of foreigner employees and entrepreneurs is done by combination of all available data sources. Known total number of foreigners broken down by the length of stay from MOI makes possible the division the group into foreign residents and non-residents. We estimate the number of economically non-active residents by difference between the total number of foreign residents, the number of legal employed residents (from MLSA) and entrepreneurs (from MIT).

A similar procedure is applied for calculation of economically non-active non-residents. At first, it has to be estimated the number of illegal employment. Estimates of illegally employed foreigners are based on the results of inspection performed by controls at employers employing foreigners in the co-operation among Employment Bureau, Alien Policy Inspectorate and Customs Offices (MLSA, 2011). Illegal employees most often recruit from the citizens of third countries asking for international protection and waiting for resolution. According the report from controls, if foreigners wait longer than 3 to 6 months, they usually start working illegally. According to the employment law, the foreigner may obtain permission for Czech labour market for a period of one year from the date of making an application.

The economically non-active non-residents are represented by foreign students and others non-active persons (e.g housewives). The numbers of foreign students studying in the Czech Republic is obtained from the Institute for Information on Education (IIE).[3] The number of other economically non-active non-residents is calculated as a difference between the numbers of the total foreign non-residents and economically active non-residents (legal + illegal employed) and foreign students.

1.2 Estimation of remittances

Usually, there are no or very limited direct information about remittances. These international flows are very difficult to capture regardless transferred via cash or via bank accounts. At first, central banks lost lots of their power in recent years when the law regulating banking statistics was changed in 2005 (Kudlák, Písaříková, 2008). At second, these transactions were often tried to hide. Actually, the issue of migration is very often difficult to measure especially with respect to national accounts requirements.[4] Therefore model approach is the option how to provide at least rough estimates of these flows. The model is based on the number of people, country of origin, activity, their consumption habits etc. Such model requires some survey or qualified judgement to put into practice. Therefore the CZSO conducted the research project with the help of the Institute of Sociology of the Academy of Sciences of the Czech Republic (ISAS). The ISAS prepared survey conducted in 2010 and it was partial financed by the CZSO (Leontiyeva, Tollarová, 2011 a, b). The project was focused on labour migration, incomes,

[3] The Institute for Information on Education was closed to 31st December 2011. From 1st January 2012 agenda IIE associated with the collection and processing of data takes over the Ministry of Education, Youth and Sports.

[4] The methodology of national accounts is very complex. The description of the account of the rest of the world can be found in Hronová et al. (2009).

consumption expenditures, savings and remittances of several nationalities in the Czech Republic: Ukraine, Vietnam, Russia, Moldova and the countries of the former Yugoslavia. The results of this project were used for estimation of monetary remittances and in kind remittances of foreigners of five countries mentioned above. It was possible to survey only five countries within this the project. It means that behaviour of migrants from other countries has to be expertly estimated on the basis of similarity with selected countries. On this basis, groups of countries were created representing clusters with similar behaviour in terms of incomes, consumption and remittances. When estimating non-resident's remittances, the CZSO supposes the equality between their savings and remittances. It means that all earned funds remaining after deducting consumption expenditures are sent back to their home country. It is supposed that non-residents do not create savings in the Czech Republic or their savings are only temporary and will be spent (sooner or later) in their home country.

The basis for calculation of savings and remittances is the net wage. The net wage has to be broken down by type of activity, whether the foreigner is employed legally or illegally, or if he is entrepreneur. The CZSO estimates also gross wage (wages, salaries plus taxes and insurance) based on database maintained by private company Trexima (ISPV, 2011). There are available the data on average wages of selected employees' citizenship working in the Czech Republic. Unfortunately the survey does not differentiate the type, length of stay and all of citizenship. Selected obtained average wages are used for approximation of average wages in all other countries. The net wage (legal non-resident employees and resident employees) is equal to wages and salaries less social contribution, health insurance and income taxes, see formula (1).

Net wage = Wages and salaries (D.11) – social contribution of employees (D.6112) – income taxes (D.51). (1)

Formula (1) is applied for legal employees only; in the case of illegal employees the net wage is equal to gross wage, because illegal workers do not pay any contributions or taxes. However, such employees probably receive lower wages and the CZSO assumes that such workers receive 25% lower wage than legal employee.

Something like a net wage of entrepreneurs is recorded in national account as net lending/borrowing (B.9) of the entrepreneurs' sub-sector. They have such income at their disposal for personal expenses and for the possible transfer of money to their country of origin (in case of foreign entrepreneurs), see formula (2).

*Net lending/borrowing (B.9) = The average net lending(+)/borrowing(–) of entrepreneurs * The number of entrepreneurs * k,* (2)

where k is the coefficient of ratio of the earnings of Czech entrepreneurs to earnings of foreign entrepreneurs. This coefficient k is expertly estimated and the level differs according to the type of country. It usually applies that the net lending/borrowings of foreign entrepreneurs are slightly higher than those of Czech entrepreneurs.

Economically non-active persons have no income from employment, but they have other revenues, e.g. scholarships, retirement pensions, disability pensions, etc. The CZSO also estimates scholarships of foreign students and revenues from retirement pensions of foreign pensioners.

Savings of foreigners are obtained by the deduction of final consumption expenditures from net wages. These expenditures are estimated according to the Czech households' expenditure consumption broken down by CZ-COICOP and subsequently the figures are adjusted individually for each group of countries:

*Consumption expenditures (P.31) = Σ consumption expenditures in particular sections of CZ-COICOP * r,* (3)

where *r* is the coefficient of ratio of the expenditures of Czech consumers to expenditures of foreign consumers. This coefficient r is expertly estimated for each group of countries. It provides consistent estimates with Czech households, on the contrary its weaknesses is subjectivity of adjustments. Besides, the CZSO uses the research project from the ISAS (mentioned above); where it can be determined consumption habits of some groups of foreigners.

The estimation of the total of remittances is also based on the research project from the ISAS. There was determined the average amount of remittances in CZK, the percent of their income and the percent of sent money abroad for last 3 years. The estimates of total remittances of all foreign residents are shown in Table 1, row 7. Also the information about value of sent gifts (valuables) was determined within this research project. In the same way as a percent of monetary remittances, it can be computed the percent of gifts[5] from income (see Table 1, row 14).

Table 1 Calculation of the percent of remittances and gifts from net income

	Number	Formula	Item	Ukraine	Vietnam	Russia	Moldova	Yugoslavia[5)]
REMITTANCES	1		**Average remittance in CZK** [1)]	34 228	33 375	34 813	31 667	41 698
	2		**% Remittances from income** [1)]	26.63	22.33	22.33	27.55	18.49
	3		**Donor YES**	61.00	47.00	44.00	43.00	40.00
	4		**Donor NO**	39.00	49.00	54.00	55.00	60.00
	5		**Donor without answer**	0.00	4.00	2.00	3.00	0.00
	6	3/(3+4)	**% Donor of remittances from foreigners** [2)]	61.00	48.96	44.90	43,88	40.00
	7	2*6	**% Remittances from income** [3)]	**16.24**	**10.93**	**10.03**	**12.09**	**7.40**
GIFTS	8		**Average Gifts in CZK** [1)]	17 668	17 771	16 857	24 159	22 642
	9	8/1	**Share remittance/gifts**	0.52	0.53	0.48	0.76	0.54
	10	2*9	**% Gifts from income** [1)]	13.75	11.89	10.81	21.02	10.04
	11		**N Gifts** [4)]	136	23	31	6	12
	12		**N Income** [4)]	550	238	125	46	39
	13	11/12	**% Donor of gifts from foreigners** [2)]	24.73	9.66	24.80	13.04	30.77
	14	10*13	**% Gifts from income** [3)]	**3.40**	**1.15**	**2.68**	**2.74**	**3.09**

[1)] only donors
[2)] only donors, which answer
[3)] all foreigners
[4)] number of respondents
[5)] countries of former Yugoslavia except Slovenia
Source: Leontiyeva, Tollarová (2011b), calculations of the CZSO

The remittances of resident employees and entrepreneurs are estimated as the percent from net wage or percent from net lending/borrowing.

Remittances of resident employees are equal to:

[5] Valuable items like electronic goods, expensive drugs, clothes, jewellery and even automobiles.

*Remittances (D.75) = Net wages * (% remittances + % gifts).* (4)

Remittances of entrepreneurs are equal to:

*Remittances (D.75) = Net lending/borrowing (B.9) * (% remittances + % gifts).* (5)

Remittances of entrepreneurs are equal to:

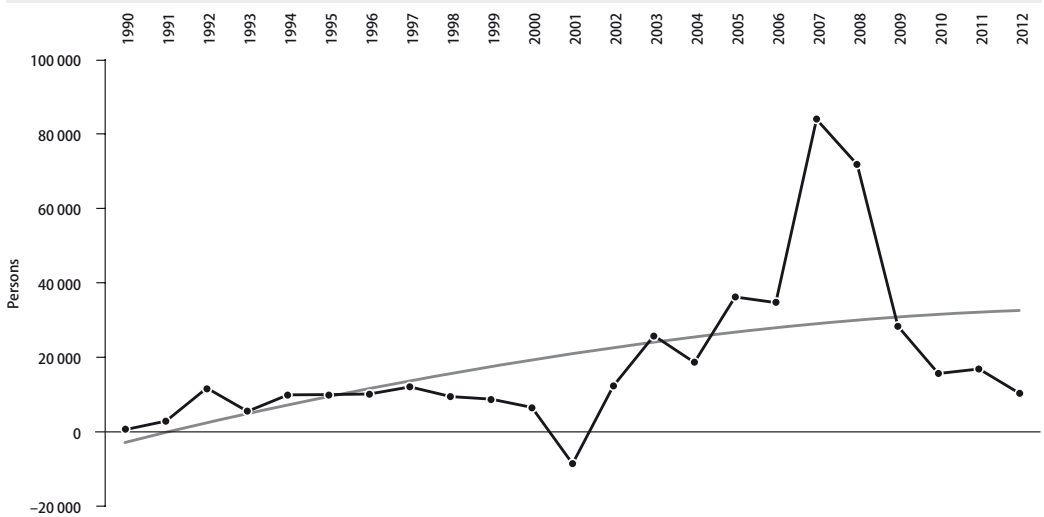
Remittances (D.75) = Savings (B.8) = Net wages – Consumption expenditures (P.31). (6)

The CZSO estimates remittances with respect to the country of origin of foreigners. That is why that the key determinant is the citizenship of migrants. For that model, it is not important where money is actually sent. Moreover, such information is not available.

2 REMITANCES FROM THE CZECH REPUBLIC IN 2011

Migration has become very important issue since 1990 (see Figure 1) and Czech society is still getting used to it. The problems connected with crime, smuggling and wild black labour market in 1990s mainly passed away and current migrants have different structure and different aims. Moreover, implementation of European law and joining to Schengen Agreement lead to major changes in Czech society. Nowadays we can identify more rich people from former Soviet bloc moving permanently to the Czech Republic. Therefore the breakdown of foreigners between residents and non-residents is more important in recent years. During the 90s, the net migration was around 10 000 persons a year, in the years 2007–2008 it rose to as much as 70–80 000 persons per year but then it dropped again to less than 30 000. In 2012, it was only just over 10 000 persons in the year. The cumulative impact on the economy is clear. Some of them became Czech citizens and the rest who are living more than one year in the Czech Republic are regarded as foreign residents. More than 530 thousand foreigners (nearly 5% of total population) living in the Czech Republic in 2011 represent significant economic power.

Figure 1 Development of net migration to the Czech Republic, annual net increase



Source: The black line represent simple moving average for rough determination of trend.
Source: CZSO (2013a)

According to the various administrative data resources, they were approximately 537 000 foreigners (see Table 2) in the Czech Republic in 2011. They represent about 5% from total population. Working foreigners count about 4% (non-resident employees, resident employees and entrepreneurs). Most foreigners come from Slovakia, Ukraine, Vietnam, Russia and Poland, see Table 3. The income according to formula (1) is estimated only for employees. Entrepreneur's earnings are expressed by net lending/borrowing (formula (2)). The revenues of economically non-active persons represent scholarships and retirement pension. In 2011, non-resident employees' net wages represented about 17 billion CZK (see Table 2), resident employees earned about 62 billion CZK and the earnings of entrepreneurs were 33 billion CZK.

Table 2 The number of foreigners in the Czech Republic, their incomes, expenses, savings and remittances

2011 (million CZK)	Total Foreigners	Non-residents			Residents			
		Total	Employees	Economically nonactive	Total	Entrepreneurs	Employees	Economically nonactive
The number of foreigners	536 815	118 484	82 358	36 126	418 331	129 513	222 946	65 872
Wages and salaries (D.11)	107 001	24 068	19 163	9 474	82 933	x	82 933	8 331
Social contribution of employers (D.12)	49 944	3 958	6 466	0	45 986	x	28 166	0
Compensation of employees (D.1)	139 145	28 029	23 124	9 474	111 116	x	111 116	8 331
Social contributions of employees (D.6112)	20 232	747	2 203	0	19 485	x	9 088	0
Income taxes (D.51)	12 716	1 333	1 333	0	11 383	x	11 383	0
Net wage	84 433	21 988	17 083	9 474	62 445	x	62 445	8 331
Consumption expenditures (P.31)	85 579	14 634	9 528	5 106	70 945	21 463	44 496	8 243
Net lending/ borrowing (B.9)	x	x	x	x	x	33 238	x	x
Net savings for export (B.8n)	41 735	11 923	7 555	4 368	29 812	11 775	17 949	88
Remittances (D.75)	14 995	x	x	x	14 995	4 933	10 062	0
Gifts (D.75)	3 191	x	x	x	3 191	1 045	2 146	0
Savings in CZ	23 549	11 923	7 555	4 368	11 626	5 797	5 741	88

Source: Calculations of the CZSO

The Czech Republic has also emigrants working abroad, who bring earned money back, of course. The CZSO estimates the Czechs abroad, but we deals only with foreigners and remittances in the Czech Republic, as outflow of funds abroad.

In 2011, the resident remittances (in cash or in kind) represented almost 18 billion CZK. Cash remittances constituted 82% of total remittances, gifts (or kind remittances) only 18%. Foreigners living and working in the Czech Republic less than one year saved and transferred away over 11 billion CZK. The total amount of transferred funds from the Czech Republic achieved almost 30 billion CZK. Total savings of foreigners for the Czech Republic was 23.5 billion CZK (see Table 2).

Labour migration is getting an important role in maintaining of a stable level of the labour force. From the perspective of economically developing countries, remittances are considered as an effective

tool for their development. However, the Czech Republic does not belong to countries for which sending remittances represents a huge problem up to now because remittances constitute approximately 0.7% of GDP, (CZSO, 2013b). For comparison we mention Luxembourg, which is in first place on the world-wide table as far as sent remittances are concerned and where in 2011 remittances amounted to 19.3% of the GDP (The World Bank, 2013).

The following Table 3 describes the number of foreigners and their remittances to countries of the five most common nationalities in the Czech Republic in 2011. The number of immigrants from Slovakia and Ukraine exceeded 100 000 persons.

Table 3 Calculation of remittances from the Czech Republic

2011	Foreigners			Remittances (million CZK)		
	Total	Non-residents	Residents	Total	Non-residents	Residents
Total	536 815	118 484	418 331	30 109	11 923	18 186
Poland	23 602	1 275	22 327	884	140	744
Russia	33 939	9 584	24 355	3 190	2 474	716
Slovakia	148 224	30 366	117 858	6 103	1 022	5 081
Ukraine	136 044	43 176	92 868	7 710	4 318	3 392
Vietnam	59 281	3 610	55 671	1 687	225	1 462

Note: Savings of non-residents will be transferred out of the Czech Republic.
Source: Calculations of the CZSO

The greatest amount of remittances was transferred to Slovakia, Ukraine and Russia in 2011. Total amount of remittances transferring to countries of five most frequent nationalities of foreigners in the Czech Republic took 65% of total remittances; 26% (7.7 billion CZK) went to Ukraine, 20% (6.1 billion CZK) to Slovakia and 11% (3.2 billion CZK) to Russia from total remittances.

3 PROBLEMS WITH STATISTICS OF MIGRANTS AND REMITTANCES

Generally, the statistics of remittances is difficult to compile due to lack of data sources. Some estimates were done but it should be noted that these estimation usually included only formal remittances in most cases. The results may be distorted by the amount of informal (unregistered) remittances (Schiopu, Siegfried, 2006). Many immigrants send remittances through informal channels, such as carrying in cash, sending via friends, drivers or passengers on public transport. It is estimated that remittances sent informally can represent up to 85% of total remittances in some countries (Rejšková, Stojanov, Šolcová, Tollarová, 2009). There are lots of reasons ranging from hiding money from financial institutions to mafia, it is present especially persons with lower incomes.

The key advantage of our model is that we do not have to study money channels, because the amount of remittances is based on a percentage from income, not from reporting banks and financial institutions. However, it occurs plenty of other problems related to the statistics of the number of migrants and the amount of remittances.

The long-term problem represents recording of the number of illegally employed foreigners. Recently the number of controls of employers decreased and thereby the estimation of illegal employment is more complicated.

Statistics of foreigners' employment currently faces a problem of the lack of data. Employment bureaus stopped registering and publishing numbers of work permits in January 2012. Unfortunately

the CZSO is still obliged to estimate the number of employed foreigners and therefore it is based on the recent development.

Another problem is the lack of data sources on average wages, consumption expenditures and remittances of foreigners by all of citizenship. Only the information about five countries is available from the research project of the ISAS. Other citizenships must be expertly estimated.

Besides, the general problem with statistics of remittances is the comparability of data on remittances between different institutions. The Table 3 shows that the amount of remittances in 2011 was 30 billion CZK according to the CZSO estimates. However, according to international statistics from the World Bank the amount of remittances outflow from the Czech Republic achieved in 2011 approximately 41 billion CZK (The World Bank, 2013). The reason for discrepancies is obviously different approach to remittances. The CZSO methodology considers remittances (in case of non-resident employees) only as savings, i.e. the difference between net wage and consumption expenditures. In case of resident employees and entrepreneurs the remittances are estimated as the percent from net wage or percent from net lending/borrowing. However, some institutions including the World Bank define the non-resident employees' remittances as total compensation of employees, regardless of consumption in foreign country. Consequently, the amount of remittances is higher using this methodology. We consider the approach of the World Bank as incorrect because it is not in line with other macroeconomic information from national accounts.

CONCLUSION

The role of the statistics of migration significantly increased in the last years. Information about migration and remittances are still gaining the importance as the Czech Republic is facing increasing migration. It is clear that these issues have to be reflected in balance of payments as the outflows and inflows of money. They are reflected in gross national income since the outflow of money of foreign residents decreases sources for domestic consumption expenditures.

Therefore it is important to record the foreigners living in the Czech Republic as well as Czech citizens living abroad. Moreover, the group of foreigners has to be split between residents and non-residents. Foreign non-residents buy goods and services that is recorded as exports and Czech residents abroad affect import of goods and services. Detailed data on the numbers of foreigners in the Czech Republic and Czech citizens abroad are difficult to obtain in appropriate detail. It is due to the unknown behaviour of foreigners, their incomes and expenses. Nowadays, the Czech Republic is country of immigrants and although we have emigrants working abroad and bringing earned money back, remittances are more important in terms of outflow of funds abroad. Therefore we focused on the foreigners in the Czech Republic, although the CZSO estimates a Czech citizens working and studying abroad, as well.

Even though actual estimates based solely on the physical numbers of people have some limitations, they provide useful information. Since the key problem lying in the lack of data sources is remaining, model approach is probably the only way how to satisfy statistical needs. Lots of assumptions have to be used to obtain required results. The improvement of this statistics represents a great challenge and the CZSO is hardly working on it. Definitely a significant progress can be made again in connection with some researches dealing with migration and social issues. We expect that this issue is still going to be more popular in connection with planning some migration strategy or promotion of migration of high skilled workers. Our labour market was very attractive mainly for the people from the east but current long-term economic crisis in the EU may bring change. These issues are connected not only with looking for people for low-qualified jobs but also for people settling in the country and creating values that remain in the Czech Republic.

References

CZSO (a). *Pohyb obyvatelstva v Českých zemích 1785–2012, absolutní údaje* (Population of the Czech Lands 1785–2012, absolute data) [online]. Prague: Czech Statistical Office, 2013. [cit. 4.9.2013]. <http://www.czso.cz/csu/csu.nsf/i/tab1_obyrcr/$File/c-4001-13.xls>.

CZSO (b). *HDP Výrobní metoda* (GDP Production Method) [online]. Prague: Czech Statistical Office, 2013. [cit. 2.9.2013]. <http://apl.czso.cz/pll/rocenka/rocenkavyber.makroek_prod>

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

HRONOVÁ, S., FISCHER, J., HINDLS, R., SIXTA, J. *Národní účetnictví. Nástroj popisu globální ekonomiky* (National Accounting. Tool for Describtion of the Global Economy). Prague: C.H.Beck, 2009.

IMF. *International transactions in remittances: guide for compilers and users*. Washington D.C.: International Monetary Fund, 2009.

ISPV. *Informační systém o průměrném výdělku – Mzdová sféra – rok 2011* (Information System on Average Earnings – Wage Sphere – 2011) [online]. Prague: Trexima, 2011. [cit. 4.9.2013]. <http://www.ispv.cz/getattachment/c9c1bcda-7b4a-44d7-bb1b-806a2197b2c0/CR_114_MZS-pdf.aspx?disposition=attachment>.

KUDLÁK, K., PÍSAŘÍKOVÁ, Š. Statistická šetření vývozu a dovozu služeb (Statistical surveys of exports and imports of services). *Statistika*, 2008, 5, pp. 441–448.

LEONTIYEVA Y., TOLLAROVÁ B. (a). *Šetření cizinců o jejich příjmech, výdajích a remitencích. Závěrečná zpráva z výzkumu* (Investigation of foreigners on their incomes, expenditures and remittances. Final report). Prague: Sociologický ústav AV ČR, 2011.

LEONTIYEVA Y., TOLLAROVÁ B. (b). *Tabulková příloha k závěrečné zprávě z výzkumu Šetření cizinců o jejich příjmech, výdajích a remitencích* (Table annex to the final report about the Investigation of foreigners on their incomes, expenditures and remittances). Prague: Sociologický ústav AV ČR, 2011.

MAZOUCH, P., FISCHER, J. *Lidský kapitál – měření, souvislosti, prognózy* (Human capital – measurement, context, forecasts). Prague: C.H.Beck, 2011.

MLSA. *Souhrnná informace za rok 2011 o aktivitách realizovaných příslušnými resorty v oblasti potírání nelegálního zaměstnávání cizinců* (Summary information for 2011 on the activities carried out by relevant departments in combating illegal employment of foreigners) [online]. Prague: MLSA, 2011. [cit. 2.9.2013]. <http://www.mpsv.cz/files/clanky/13355/potirani_cerne_prace_2011.pdf>.

ONDRUŠ, V. National Accounts and economic migration – Remittances in the Czech Republic [online]. *17th International Input-output Conference*. Sao Paulo, Brazil 2009. [cit. 2.9.2013]. <http://www.iioa.org/conferences/17th/papers/1068331043_090529_132336_PAPER_ONDRUS.PDF>.

REJŠKOVÁ, T., STOJANOV, R., ŠOLCOVÁ, P., TOLLAROVÁ, B. *Studie: Remitence zasílané z České republiky a jejich rozvojový dopad* (Study: Remittances sent from the Czech Republic and their development impact) [online]. Migraceonline.cz, 2009. [cit. 2.9.2013]. <http://www.migraceonline.cz/cz/e-knihovna/studie-remitence-zasilane-z-ceske-republiky-a-jejich-rozvojovy-dopad>.

SCHIOPU, I., SIEGFRIED, N. *Determinants of workers' remittances. Evidence from the European neighbouring region* [online]. Frankfurt: European Central Bank, 2006. [cit. 2.9.2013]. <http://www.suomenpankki.fi/pdf/128539.pdf>.

ŠIMKOVÁ, M., SIXTA, J. Vývoj životní úrovně osob v důchodovém věku (The development of the standard of living of the people of retirement age). *Acta Oeconomica Pragensia*, 2013/3, Vol. 21, pp. 14–31.

THE WORLD BANK. *Migration and Remittances*. Annual remittances data [online]. Washington D.C.: The World Bank group, 2013. [cit. 2.9.2013]. <http://econ.worldbank.org/WBSITE/EXTERNAL/EXTDEC/EXTDECPROSPECTS/0,,contentMDK:22759429~pagePK:64165401~piPK:64165026~theSitePK:476883,00.html>.

Predictive Estimation of Finite Population Mean Using Exponential Estimators

Housila P. Singh[1] | *Vikram University, M. P., India*
Ramkrishna S. Solanki[2] | *Vikram University, M. P., India*
Alok K. Singh[3] | *Vikram University, M. P., India*

Abstract

This paper suggested the ratio-type and product-type exponential estimators of the population mean of a study variable through predictive approach using Bahl and Tuteja (1991) ratio-type and product-type exponential estimators as a predictor of the mean of the unobserved units of the population. Properties of the suggested estimators are studied up to first order of approximation in simple random sampling using information on an auxiliary variable. The theoretical conditions under which the suggested estimators are less biased and more efficient than the usual unbiased, ratio, product estimators and estimators due to Srivastava (1983) and Bahl and Tuteja (1991) have been obtained. In support of the theoretical study numerical illustration is also given and determined that the suggested estimators showed also an improvement over the classical estimators empirically.

Keywords

Predictive approach, auxiliary information, population mean, exponential estimators, bias, mean squared error

JEL code

C13, C83

INTRODUCTION

Sample surveys are widely used as a cost effective apparatus of data collection and for making valid inference about population parameters. Since in sample surveys the sample is only a part of the whole, extrapolation inevitably leads to errors. The main aim of survey statisticians is to reduce the errors either by devising suitable sampling schemes or by formulating efficient estimators of the parameters, see Singh and Solanki (2012, 2013) or both. To detract the errors various researchers have attempted to use additional information, which is correlated to the information under the study and about which the information is available before start of the survey known as auxiliary information. The literature on survey sampling describes a great variety of techniques/approaches including design based and model based methods for using auxiliary information to obtain more efficient estimators.

[1] Professor, S. S. in Statistics, Vikram University, Ujjain-456010, M. P., India.
[2] Research Scholar, S. S. in Statistics, Vikram University, Ujjain-456010, M. P., India. Corresponding author: e-mail: ramkssolanki@gmail.com.
[3] Research Scholar, S. S. in Statistics, Vikram University, Ujjain-456010, M. P., India.

In the predictive approach a model is specified for the population values and is used to predict the non-sampled values. Prediction theory for sample surveys (or model-based theory) can be con¬sidered as a general framework for statistical inferences on the character of finite population. Well known estimators of population parameters encountered in the classical theory, as expansion, ratio, regression, another estimators can be predictors in the general prediction theory under some special model. Several authors have applied the predictive approach either to form new predictive estimators or to examine the existing estimators from the predictive viewpoint. It is observed that the use of usual unbiased, ratio and regression estimators as a predictor for the mean of the unobserved units of the population result in the corresponding customary (usual) estimators of the mean of whole population.

Srivastava (1983) has shown that if the usual product estimator is used as a predictor for the mean of the unobserved units of the population, the resulting estimator of the mean of the whole population is different from the customary (usual) product estimator. Biradar and Singh (1998), Agrawal and Roy (1999) and Nayak and Sahoo (2012) provided some predictive estimators for finite population variance. Sahoo and Panda (1999) developed the regression type estimator for two stage sampling procedure. Sahoo and Sahoo (2001) and Sahoo et al. (2009) introduced a class of estimators for the finite population mean availing information on two auxiliary variables in two stage sampling. Ahmed (2004) proposed some estimators for finite population mean in two stage sampling using multivariate auxiliary information. Saini (2013) proposed a class of predictive estimators for two stage design consisting especially of two estimators namely ratio and regression.

In the preset study we attempt to examine the existing Bahl and Tuteja (1991) exponential estimators as predictor of the mean of the unobserved units of the population using the information of observed units in sample. Remaining part of the paper is organized as follows: Section 1 defines some notations and discusses some existing estimators of population mean. We suggest estimators with their properties in Section 2. We perform the theoretical comparison among different estimators in the Sections 3 and 4. In Section 5, the real data sets are used to observe the performance of various estimators numerically. Finally, last section provides some concluding remarks.

1 THE NOTATIONS

Much literature has been produced on sampling from finite populations to address the issue of the efficient estimation of the mean (or total) of a survey variable when auxiliary variables are available. Our analysis refers to simple random sampling without replacement (SRSWOR) and considers, for brevity, the case when only a single auxiliary variable is used.

Consider a finite population $U = (U_1, U_2, ..., U_N)$ of N units on which the study (survey) variable y and auxiliary variable x are defined, which take values $y1$ and $y2$ respectively for the unit U_1 of U $(1 \leq i \leq N)$. We are interested in estimating the population mean:

$$\overline{Y} = \frac{\sum_{i=1}^{N} y_i}{N},$$

of the study variable y on the basis of observed values of y on the units of a sample taken from finite population U. Any ordered subset of U is called a sample from U. Let S denote the collection of all possible samples from U. For any given $s \in S$, let $\vartheta(s)$ denote its effective sample size (the number of distinct units in (s) and $\bar{s}$ denote the set of all those units of U which are not in s. We designate:

$$\overline{Y}_s = \frac{1}{\vartheta(s)} \sum_{i \in s} y_i ,$$

$$\overline{Y}_{\bar{s}} = \frac{1}{(N - \vartheta(s))} \sum_{i \in s} y_i .$$

For any given $s \in S$, we can write:

$$\overline{Y} = \left[\frac{\vartheta(s)}{N} \overline{Y}_s + \frac{(N - \vartheta(s))}{N} \overline{Y}_s \right]. \quad (1)$$

In the representation of population mean $\overline{Y}$ at (1), the sample mean $\overline{Y}$ is known because it is based on the units of the sample *s* whose *y* values have been observed. Therefore, the statisticians should attempt a prediction of the mean $\overline{Y}_s$ of the unobserved units of the population *U* on the basis of observed units in *s*. While admitting that a decision-theorist might object to making the choice of estimator after looking at the data, Basu (1971) nevertheless considered such an approach to represent the "heart of the matter" in estimating the finite population mean [see Cessal et al. (1977, p. 110)].

For a simple random sampling procedure with sample size n (*i.e.* $\vartheta(s) = n$) and the sample mean:

$$\bar{y} = \frac{1}{n} \sum_{i \in s} y_i, \text{ (i.e. } \overline{Y}_s = \bar{y}).$$

We can write (1) as:

$$\overline{Y} = \left[\frac{n}{N} \overline{Y}_s + \frac{(N - n)}{N} \overline{Y}_s \right]. \quad (2)$$

From (2), an estimator of population mean $\overline{Y}$ can be written as:

$$t = \left[\frac{n}{N} \bar{y} + \frac{(N - n)}{N} T \right],$$

where *T* is considered as a predictor of $\overline{Y}_s$.

Srivastava (1983) has shown that if we adopt the prediction approach described earlier, use of:

$$\bar{y} = \frac{1}{n} \sum_{i \in s} y_i \quad \text{(mean per unit estimator)},$$

$$\bar{y}_{\bar{lr}} = [\bar{y} + b(\overline{X}_s - \bar{x})] \quad \text{(the regression estimator)},$$

$$\bar{y}_{\bar{R}} = \overline{X}_s \left(\frac{\bar{y}}{\bar{x}} \right) \quad \text{(the ratio estimator)},$$

for predicting the mean $\overline{Y}_s$ of the unobserved units of the population result in the corresponding customary:

$$\bar{y} = \frac{1}{n} \sum_{i \in s} y_i \quad \text{(mean per unit estimator)},$$

$$\bar{y}_{lr} = [\bar{y} + b(\overline{X} - \bar{x})] \quad \text{(the regression estimator)},$$

$$\bar{y}_R = \bar{X}\left(\frac{\bar{y}}{\bar{x}}\right) \qquad \text{(the ratio estimator)},$$

of the population mean $\bar{Y}$, [*i.e.* if $T = \bar{y}$, $t = \bar{y}$; $T = \bar{y}_{\bar{R}}$, $t = \bar{y}_R$; $T = \bar{y}_{\bar{lr}}$, $t = \bar{y}_{lr}$], where b is the regression coefficient estimated from the sample s and:

$$\bar{x} = \frac{1}{n}\sum\nolimits_{i \in s} x_i,$$

$$\bar{X} = \frac{1}{n}\sum\nolimits_{i=1}^{N} x_i,$$

$$\bar{X}_{\bar{s}} = \frac{1}{(N-n)}\sum\nolimits_{i \in \bar{s}} x_i = \frac{(N\bar{X} - n\bar{x})}{(N-n)}.$$

However, if the product estimator:

$$\bar{y}_{\bar{P}} = \bar{y}\left(\frac{\bar{x}}{\bar{X}_{\bar{s}}}\right)$$

is used with such an approach, the resulting estimator of population mean $\bar{Y}$ is not the customary product estimator:

$$\bar{y}_P = \bar{y}\left(\frac{\bar{x}}{\bar{X}}\right),$$

i.e. if:

$$T = \bar{y}_{\bar{P}},\ t = \frac{n\bar{X} + (N-2n)\bar{x}}{(N\bar{X} - n\bar{x})} = t_P.$$

To the first degree of approximation the biases and mean squared errors (*MSE*s) of the estimators $\bar{y}_R$, $\bar{y}_P$ and t_p are respectively given by:

$$Bias(\bar{y}_R) = \theta\bar{Y}C_x^2(1 - C), \tag{3}$$

$$Bias(\bar{y}_p) = \theta\bar{Y}C^2\ C, \tag{4}$$

$$Bias(t_p) = \theta\bar{Y}C_x^2[C + f(1-f)^{-1}], \tag{5}$$

$$MSE(\bar{y}_R) = \theta\bar{Y}C^2[C_y^2 + C_x^2(1 - 2C)], \tag{6}$$

$$MSE(\bar{y}_p) = MSE(t_p) = \theta\bar{Y}^2[C_y^2 + C_x^2(1 + 2C)], \tag{7}$$

where:

$\theta = (1-f)^{-1}, f = (n/N), C_y^2 = (S_y^2/\overline{Y}^2), C_x^2 = (S_x^2/\overline{X}^2), S_y^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \overline{Y})^2$,

$S_x^2 = (N-1)^{-1}\sum_{i=1}^{N}(x_i - \overline{X})^2$, $C = \rho(C_y/C_x)$, $\rho = S_{yx}/(S_y S_x)$) and

$S_{yx} = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \overline{Y})(x_i - \overline{X})$.

Bahl and Tuteja (1991) suggested the ratio-type and product- type exponential estimators of the population mean $\overline{Y}$ respectively as:

$$\bar{y}_{Re} = \bar{y}\, exp\left(\frac{\overline{X} - \bar{x}}{\overline{X} + \bar{x}}\right),$$
$$\bar{y}_{Pe} = \bar{y}\, exp\left(\frac{\bar{x} - X}{\bar{x} + \overline{X}}\right).$$

To first degree of approximation, the biases and mean squared errors of $\bar{y}_{Re}$ and $\bar{y}_{Pe}$ are respectively given by:

$$Bias(\bar{y}_{Re}) = \frac{\theta}{8}\overline{Y}C_x^2(3-4C), \tag{8}$$

$$Bias(\bar{y}_{Pe}) = \frac{\theta}{8}\overline{Y}C_x^2(4C-1), \tag{9}$$

$$MSE(\bar{y}_{Re}) = \theta\overline{Y}^2\left[C_y^2 + \frac{C_x^2}{4}(1-4C)\right], \tag{10}$$

$$MSE(\bar{y}_{Pe}) = \theta\overline{Y}^2\left[C_y^2 + \frac{C_x^2}{4}(1+4C)\right]. \tag{11}$$

In the following Sec. 2 we have suggested alternative ratio-type and product-type exponential estimators of population mean $\overline{Y}$ by using $\bar{y}_{Re}$ and $\bar{y}_{Pe}$ as a predictor T of $\overline{Y}_{\bar{s}}$ of the unobserved units of the population U on the basis of observed units in s. The biases and mean squared errors of suggested ratio-type and product-type exponential estimators up to first order approximation have obtained.

2 THE SUGGESTED PREDECTION APPROACH

In case, information on an auxiliary variable x positively correlated with the study variable y is available and one intends to use this in the form of Bahl and Tuteja (1991) ratio-type exponential estimator $\bar{y}_{Re}$, an obvious choice for T is:

$$\bar{y}_{\overline{Re}} = \bar{y}\, exp\left(\frac{\overline{X}_s - \bar{x}}{\overline{X}_s + \bar{x}}\right).$$

For this choice of T:

$$t = t_{Re} = \left[\frac{n}{N}\bar{y} + \left(\frac{N-n}{N}\right)\bar{y}\, exp\left(\frac{\overline{X}_{\bar{s}} - \bar{x}}{\overline{X}_{\bar{s}} + \bar{x}}\right)\right] = \left[\frac{n}{N}\bar{y} + \left(\frac{N-n}{N}\right)\bar{y}\, exp\left(\frac{N(\overline{X} - \bar{x})}{N(\overline{X} - \bar{x}) - 2n\bar{x}}\right)\right], \tag{12}$$

which is not the Bahl and Tuteja (1991) ratio-type exponential estimator $\bar{y}_{Re}$.

If the auxiliary variable x is negatively correlated with the study variable y and one wants to use this in the form of Bahl and Tuteja (1991) product-type exponential estimator $\bar{y}_{Pe}$ an obvious choice for T is:

$$\bar{y}_{\overline{Pe}} = \bar{y}\, exp\left(\frac{\bar{x} - \overline{X}_s}{\bar{x} + \overline{X}_s}\right).$$

For this choice of T:

$$t = t_{Pe} = \left[\frac{n}{N}\bar{y} + \left(\frac{N-n}{N}\right)\bar{y}\, exp\left(\frac{\bar{x} - \overline{X}_s}{\bar{x} + \overline{X}_s}\right)\right] = \left[\frac{n}{N}\bar{y} + \left(\frac{N-n}{N}\right)\bar{y}\, exp\left(\frac{N(\bar{x} - X)}{N\overline{X} + (N-2n)\bar{x}}\right)\right], \tag{13}$$

which is not the Bahl and Tuteja (1991) product-type exponential estimator $\bar{y}_{Pe}$.

To obtain the biases and MSEs of t_{Re} and t_{Pe}, we define:

$$e_0 = \left(\frac{\bar{y} - \overline{Y}}{\overline{Y}}\right) \text{and } e_1 = \left(\frac{\bar{x} - \overline{X}}{\overline{X}}\right),$$

such that:

$$E(e_0) = E(e_1) = 0\,,$$

and up to first degree of approximation:

$$E(e_0^2) = \theta C_y^2\,,$$

$$E(e_1^2) = \theta C_x^2\,,$$

$$E(e_0 e_1) = \theta C C_x^2\,,$$

Expressing (12) in terms of *e*'s, we have:

$$\begin{aligned} t_{Re} &= \overline{Y}(1 + e_0)\left[\frac{n}{N} + \left(\frac{N-n}{N}\right) exp\left(\frac{Ne_1}{2(N-n) + (N-2n)e_1}\right)\right] \\ &= \overline{Y}(1 + e_0)\left[f + (1-f) exp\left(-\frac{e_1}{2(1-f) + (1-2f)e_1}\right)\right] \\ &= \overline{Y}(1 + e_0)\left[f + (1-f) exp\left\{-\frac{e_1}{2(1-f)}\left(1 + \frac{(1-2f)}{2(1-f)}e_1\right)^{-1}\right\}\right]. \end{aligned} \tag{14}$$

Expanding the right hand side of (14), multiplying out and neglecting terms of *e*'s having power greater than two we have:

$$t_{Re} \approx \overline{Y}\left[1 + e_0 - \frac{e_1}{2} - \frac{e_0 e_1}{2} + \frac{e_1^2}{8}(3 - 4f)\right],$$

or:
$$(t_{Re} - \overline{Y}) \approx \overline{Y}\left[e_0 - \frac{e_1}{2} - \frac{e_0 e_1}{2} + \frac{e_1^2}{8}(3 - 4f)\right], \tag{15}$$

Taking expectation of both sides of (15), we get the bias of t_{Re} to the first degree of approximation as:

$$Bias(t_{Re}) = \frac{\theta}{8}\overline{Y}C_x^2[3 - 4(C + f)]. \tag{16}$$

Squaring both sides of (15) and neglecting terms of *e*'s having power greater than two, we have:

$$(t_{Re} - \overline{Y})^2 \approx \overline{Y}^2\left(e_0^2 + \frac{e_1^2}{4} - e_0 e_1\right). \tag{17}$$

Taking expectation of both sides of (17) we get the MSE of t_{Re} to the first degree of approximation as:

$$MSE(t_{Re}) = \theta\overline{Y}^2\left[C_y^2 + \frac{C_x^2}{4}(1 - 4C)\right], \tag{18}$$

which equals to the *MSE* of Bahl and Tuteja (1991) ratio-type exponential estimator $\bar{y}_{Re}$.

Now expressing t_{Pe} in terms of e's, we have:

$$t_{Pe} = \overline{Y}(1 + e_0)\left[\frac{n}{N} + \left(\frac{N-n}{N}\right)exp\left(\frac{Ne_1}{2(N-n) + Ne_1}\right)\right]$$

$$= \overline{Y}(1 + e_0)\left[f + (1-f)exp\left(-\frac{e_1}{2(1-f) + e_1}\right)\right]$$

$$= \overline{Y}(1 + e_0)\left[f + (1-f)exp\left(\frac{e_1}{2(1-f)}\left(1 + \frac{e_1}{2(1-f)}\right)^{-1}\right)\right]. \tag{19}$$

Expanding the right hand side of (19), multiplying out and neglecting terms of *e*'s having power greater than two we have:

$$t_{Pe} \approx \overline{Y}\left[1 + e_0 + \frac{e_1}{2} + \frac{e_0 e_1}{2} - \frac{e_1^2}{8(1-f)}\right]$$

or: $$(t_{Pe} - \overline{Y}) \approx \overline{Y}\left[e_0 + \frac{e_1}{2} + \frac{e_0 e_1}{2} - \frac{e_1^2}{8(1-f)}\right]. \tag{20}$$

Taking expectation of both sides of (20), we get the bias of t_{Pe} to the first degree of approximation as:

$$Bias(t_{Pe}) = \frac{\theta}{8}\overline{Y}C_x^2\left(4C - \frac{1}{(1-f)}\right). \tag{21}$$

Squaring both sides of (20) and neglecting terms of *e*'s having power greater than two we have:

$$(t_{Pe} - \overline{Y})^2 \approx \overline{Y}^2\left(e_0^2 + \frac{e_1^2}{4} + e_0 e_1\right). \tag{22}$$

Taking expectation of both sides of (22), we get the MSE of t_{Pe} to the first degree of approximation as:

$$MSE(t_{Pe}) = \theta\overline{Y}^2\left[C_y^2 + \frac{C_x^2}{4}(1 + 4C)\right], \tag{23}$$

which equals to the MSE of Bahl and Tuteja (1991) product-type exponential estimator $\bar{y}_{Pe}$.

3 BIAS COMPARISON

In this Section we have compared the absolute biases of the different estimators of the population mean $\bar{Y}$. The relevant conditions are given in which the proposed ratio-type exponential estimator $\bar{y}_{Re}$ (product-type exponential estimator t_{Pe}) is less biased to usual ratio estimator $\bar{y}_R$ (product estimator $\bar{y}_p$) and Bahl and Tuteja (1991) ratio-type exponential estimator $\bar{y}_{Re}$ (product-type exponential estimator $\bar{y}_{Pe}$).

3.1 Bias comparison of ratio type estimators

From (3), (8) and (16), we have:

(i) $|Bias(t_{Re})| < |Bias(\bar{y}_R)|$,if:

$$\frac{1}{8}|3-4(C+f)| < |1-C|,$$

i.e. if:

$$[48C^2 - 104C - 16f^2 + 24f - 32Cf + 55] > 0. \tag{24}$$

(ii) $|Bias(t_{Re})| < |Bias(\bar{y}_{Re})|$,

if:

$$3-4(C+f)| < |3-4C|,$$

i.e. if:

$$C < \frac{1}{4}(3-2f). \tag{25}$$

If the conditions (24) and (25) are satisfied the proposed alternative ratio-type exponential estimator t_{Re} is less biased respectively to customary ratio estimator $\bar{y}_R$ and Bahl and Tuteja (1991) ratio-type exponential estimator $\bar{y}_{Re}$.

3.2 Bias comparison of product type estimators

From (4), (5), (9) and (21), we have:

(i) $|Bias(t_{Pe})| < |Bias(\bar{y}_p)|$,

if:

$$\frac{1}{8}\left|4C - \frac{1}{(1-f)}\right| < |C|,$$

i.e. if:

$$\left[48C^2 + \frac{8C}{(1-f)} - \frac{1}{(1-f)^2}\right] > 0. \tag{26}$$

(ii) $|Bias(t_{Pe})| < |Bias(t_p)|$

If:

$$\frac{1}{8}\left|4C-\frac{1}{(1-f)}\right|<\left|C+\frac{f}{(1-f)}\right|,$$

i.e. if:

$$\left[48C^2+\frac{(8f-1)(8f+1)}{(1-f)^2}-\frac{2C(4+f)}{(1-f)}\right]>0. \tag{27}$$

(iii) $|Bias(t_{Pe})|<|Bias(\bar{y}_{Pe})|$

If:

$$\left|4C-\frac{1}{(1-f)}\right|<|4C-1|,$$

i.e. if:

$$C>\frac{(2-f)}{8(1-f)}. \tag{28}$$

If the conditions (26), (27) and (28) are satisfied the proposed alternative product-type exponential estimator t_{Pe} is less biased respectively to customary product estimator $\bar{y}_p$, Srivastava (1983) product estimator t_P and Bahl and Tuteja (1991) product-type exponential estimator $\bar{y}_{Pe}$.

4 EFFICIENCY COMPARISON

In this Section we have obtained the conditions under which the proposed ratio-type exponential estimator t_{Re} and Bahl and Tuteja (1991) ratio-type exponential estimator $\bar{y}_{Re}$ [product-type exponential estimator t_{Pe} and Bahl and Tuteja (1991) product-type exponential estimator $\bar{y}_{Pe}$] are better than the usual unbiased estimator $\bar{y}$ and ratio estimator $\bar{y}_R$ (product estimator $\bar{y}_p$).

It is very well known under simple random sampling without replacement that the:

$$MSE(y)=Var(y)=\theta Y^2C_y^2. \tag{29}$$

4.1 Efficiency comparison of ratio type estimators

From (6), (10), (18) and (29), we have:

(i) $MSE(\bar{y}_R)<MSE(\bar{y})$ if $C>\frac{1}{2}$. (30)

(ii) $[MSE(t_{Re})=MSE(\bar{y}_{Re})]<MSE(\bar{y})$ if $C>\frac{1}{4}$. (31)

(iii) $[MSE(t_{Re})=MSE(\bar{y}_{Re})]<MSE(\bar{y}_R)$ if $C<\frac{3}{4}$. (32)

Thus from (31) and (32) it follows that if the condition:

$$\left(\frac{1}{4}<C<\frac{3}{4}\right), \tag{33}$$

is satisfied the proposed estimator t_{Re} and Bahl and Tuteja (1991) estimator $\bar{y}_{Re}$ are better than the usual unbiased estimator $\bar{y}$ and usual ratio estimator $\bar{y}_R$.

4.2 Efficiency comparison of product type estimators

From (7), (11), (23) and (29), we have:

$$\text{(i)} \quad [MSE(\bar{y}_P) = MSE(t_P)] < MSE(\bar{y}) \text{ if } C < -\frac{1}{2}. \tag{34}$$

$$\text{(ii)} \quad [MSE(t_{Pe}) = MSE(\bar{y}_{Pe})] < MSE(\bar{y}) \text{ if } C < -\frac{1}{4}. \tag{35}$$

$$\text{(iii)} \quad [MSE(t_{Pe}) = MSE(\bar{y}_{Pe})] < [MSE(\bar{y}_P) = MSE(t_P)] \text{ if } C > -\frac{3}{4}. \tag{36}$$

Thus from (35) and (36) it follows that the condition:

$$\left(-\frac{3}{4} < C < -\frac{1}{4}\right), \tag{37}$$

is sufficient for the proposed estimator t_{Pe} and Bahl and Tuteja (1991) estimator $\bar{y}_{Pe}$ are better than the usual unbiased estimator $\bar{y}$ and usual product estimator $\bar{y}_P$ and Srivastava (1983) estimator t_P.

5 EMPIRICAL STUDY

To judge the merits of the suggested estimators t_{Re} and t_{Pe} over the estimators $\bar{y}$, $\bar{y}_R$, $\bar{y}_P$, $\bar{y}_{Re}$, $\bar{y}_{Pe}$ and t_P we have considered four natural population data sets. The description of the population data sets are given in Table 1.

Table 1 The population data sets

Population	N	n	C_x	C_y	ρ	C
I: Steel and Torrie (1960, p. 282) *y*: Log of leaf burn in sec. *x*: Chlorine percentages	30	6	0.7493	0.7000	0.4996	0.4667
II: Murthy (1967, p. 228) *y*: Output *x*: Fixed capital	80	20	0.7507	0.3542	0.9413	0.4441
III: Das (1988) *y*: The number of agricultural laborers for 1961 *x*: The number of agricultural laborers for 1971	278	60	1.6198	1.4451	0.7213	0.6435
IV: Cochran (1977) *y*: The number of persons per block *x*: The numbers of rooms per block	20	8	0.1281	0.1445	0.6500	0.7332

Source: Own construction

To examine the biasedness of various estimators of population mean $\overline{Y}$ we have computed the following quantities:

$$Q_{R1} = \left|\frac{Bias\,(\bar{y}_R)}{\theta\overline{Y}C_x^2}\right| = |1 - C|,\ Q_{R2} = \left|\frac{Bias\,(\bar{y}_{Re})}{\theta\overline{Y}C_x^2}\right| = \frac{1}{8}|3 - 4C|,\ Q_{R3} = \left|\frac{Bias\,(t_{Re})}{\theta\overline{Y}C_x^2}\right| = \frac{1}{8}|3 - 4(C + f)|,$$

$$Q_{P1} = \left|\frac{Bias\,(\bar{y}_P)}{\theta\overline{Y}C_x^2}\right| = |C|,\ Q_{P2} = \left|\frac{Bias\,(t_P)}{\theta\overline{Y}C_x^2}\right| = \left|C + \frac{f}{(1-f)}\right|,\ Q_{R3} = \left|\frac{Bias\,(\bar{y}_{Pe})}{\theta\overline{Y}C_x^2}\right| = \frac{1}{8}|4C-1)|,$$

$$Q_{P4} = \left|\frac{Bias\,(t_{Pe})}{\theta\overline{Y}C_x^2}\right| = \frac{1}{8}\left|4C - \frac{1}{(1-f)}\right|,$$

and findings are shown in Table 2.

Table 2 Values of the quantities Q_{Ri} ($i = 1, 2, 3$) and ($j = 1, 2, 3, 4$)

	Quantities						
Population	Q_{R1}	Q_{R2}	Q_{R3}	Q_{P1}	Q_{P2}	Q_{P3}	Q_{P4}
I	0.5333	0.1416	**0.0416**	0.4667	0.7167	0.1084	0.0771
II	0.5559	0.1529	**0.0279**	0.4441	0.7775	0.0971	0.0554
III	0.3565	**0.0532**	0.0547	0.6435	0.9174	0.1967	0.1625
IV	0.2668	**0.0084**	0.1916	0.7332	1.3998	0.2416	0.1583

Note: Bold numbers indicate the least biased value in the relevant data set.
Source: Own construction

It is observed from Table 2 that:

(i) The proposed ratio-type exponential estimator t_{Re} is less biased than the customary ratio estimator $\bar{y}_R$ (i.e. $Q_{R3} < Q_{R1}$) for all population data sets I–IV because the condition (24) is satisfied for all the data sets.

(ii) The proposed ratio-type exponential estimator t_{Re} is less biased than the Bahl and Tuteja (1991) ratio-type exponential estimator $\bar{y}_R$ (i.e. $Q_{R3} < Q_{R2}$) only for data sets I and II because the condition (25) is not fulfill in data sets III and IV.

(iii) The proposed product-type exponential estimator t_{Pe} is less biased than the customary product estimator $\bar{y}_P$, Srivastava (1983) product estimator t_p and Bahl and Tuteja (1991) product-type exponential estimator $\bar{y}_{Pe}$ (i.e. $Q_{P4} < Q_{Pj}$, $j = 1, 2, 3$) for all the population data sets I-IV because the conditions (26), (27) and (28) are satisfied in all of the data sets.

To see the relative performances of different estimators of the population mean $\overline{Y}$ we have computed the percent relative efficiencies (*PRE*s) of the estimators with respect to the usual unbiased estimator $\bar{y}$ by following formulae:

$$PRE(\bar{y}_R, \bar{y}) = \frac{MSE(\bar{y})}{MSE(\bar{y}_R)} = \frac{C_y^2}{[C_y^2 + C_x^2(1-2C)]} \times 100,$$

$$PRE(\bar{y}_P, \bar{y}) = \frac{MSE(\bar{y})}{MSE(\bar{y}_p)} = \frac{C_y^2}{[C_y^2 + C_x^2(1+2C)]} \times 100 = PRE(t_P, \bar{y}),$$

$$PRE(\bar{y}_P, \bar{y}) = \frac{MSE(\bar{y})}{MSE(\bar{y}_p)} = \frac{C_y^2}{[C_y^2 + C_x^2(1+2C)]} \times 100 = PRE(t_P, \bar{y}),$$

$$PRE(\bar{y}_{Pe}, \bar{y}) = \frac{MSE(\bar{y})}{MSE(\bar{y}_{Pe})} = \frac{C_y^2}{\left[C_y^2 + \frac{C_x^2}{4}(1+4C)\right]} \times 100 = PRE(t_{Pe}, \bar{y}),$$

and finding are shown in Table 3.

Table 3 The *PREs* of different estimators with respect to $\bar{y}$

Population	$PRE(\bar{y}_R, \bar{y})$	$PRE(\bar{y}_{Re}, \bar{y}) = PRE(t_{Re}, \bar{y})$	$PRE(\bar{y}_P, \bar{y}) = PRE(t_P, \bar{y})$	$PRE(\bar{y}_{Pe}, \bar{y}) = PRE(t_{Pe}, y)$
I	92.9156	**133.0374**	31.1004	54.9076
II	66.5810	**781.3982**	10.5463	24.2836
III	156.3967	**197.7846**	25.8171	47.1121
IV	157.8695	**161.2267**	34.0327	56.4111

Note: Bold numbers indicate the largest *PRE* in relevant data set.
Source: Own construction

It is observed from Table 3 that:

(i) The proposed ratio-type estimator t_{Re} and Bahl and Tuteja (1991) ratio estimator $\bar{y}_R$ both have the largest percent relative efficiency than the usual unbiased estimator and usual ratio estimator $\bar{y}_R$ in all the population data sets I-IV because the condition (33) has been satisfied by all data sets.

(ii) The suggested product-type estimator t_{Pe} and Bahl and Tuteja (1991) product estimator $\bar{y}_{Pe}$ are superior to the usual product estimator $\bar{y}_P$ and Srivastava (1983) estimator t_p because the condition (36) is satisfied by all data set I–IV but inferior to the usual unbiased estimator $\bar{y}$ because of dissatisfied condition (35).

(iii) The suggested ratio-type exponential estimator t_{Re} has maximum percent relative efficiency (= **781.3982**) in population II as well as least bias (= **0.0279**). Therefore, the proposed ratio-type exponential estimator t_{Re} appears to be the best in the sense of having largest percent relative efficiency as well as least bias in Population II.

CONCLUSION

We have utilized Bahl and Tuteja (1991) ratio-type and product-type exponential estimators as a predictor of the mean of the unobserved units of the population and observed that the resulting ratio-type and product-type exponential estimators of the mean of the whole population are different from the customary Bahl and Tuteja (1991) ratio and product estimators. The biases and mean squared errors of suggested ratio-type and product-type exponential estimators, up to first order approximation are obtained and observed that the mean squared errors of suggested ratio-type and product-type exponential estimators are equal to the Bahl and Tuteja (1991) ratio-type and product-type exponential estimators respectively. The theoretical conditions under which the proposed estimators are less biased and more efficient than the usual unbiased, ratio, product estimators and estimators due to Srivastava (1983) and Bahl and Tuteja (1991) have been obtained. It has been also found empirically that the suggested estimators are less biased and more efficient than other existing estimators if the theoretical conditions under which the proposed estimators are less biased and more efficient are satisfied. Thus we recommend the use of the proposed estimators in practice. However this conclusion cannot be extrapolated due to limited empirical study.

ACKNOWLEDGMENT

Authors are thankful to the Managing Editor of the Statistika journal and two anonymous referees for their valuable, constructive and positive suggestions which led to an improved version of this article.

References

AGRAWAL, M. C., ROY, D. C. Efficient estimators of population variance with regression-type and ratio-type predictor-inputs. *Metron*, 1999, 57(3), pp. 4–13.

AHMED, M. S. Some Estimators for a finite population mean under two-stage sampling using multi-auxiliary information. *Applied Mathematitics and Computation*, 2004, 153(2), pp. 505–511.

BAHL, S., TUTEJA, R. K. Ratio and product type exponential estimators. *Journal of Information and Optimization Sciences*, 1991, 12(1), pp. 159–164.

BASU, D. *An Essay on the Logical Foundation of Survey Sampling*, Part I. Foundations of Statistical Inference, eds. GODAMBE, V. P., SPROTT, D. A. Toronto: Holt, Rinehart and Winston, 1971, pp. 203–242.

CASSEL, C. M., SARNDAL, C. E., WRETMAN, J. H. *Foundation of Inference in Survey Sampling*. New York, USA: Wiley, 1977.

COCHRAN, W. G. *Sampling Techniques*. 3rd ed. New York, USA: John Wiley and Sons, 1977.

DAS, A. K. *Contributions to The Theory of Sampling Strategies Based on Auxiliary Information*. Ph.D. thesis submitted to B. C. K. V. Mohanpur, Nadia, West Bengal, India, 1988.

MURTHY, M. N. *Sampling: Theory and Methods*. Statistical Publishing Society, Calcutta, India, 1967.

NAYAK, R., SAHOO, L. Some alternative predictive estimators of population variance. *Revista Colombiana de Estadística*, 2012, 35(3), pp. 509–521.

SAHOO, L. N., DAS, B. C., SAHOO, J. A Class of Predictive Estimators in Two stage sampling. *Journal of the Indian Society of Agricultural Statistics*, 2009, 63(2), pp. 175–180.

SAHOO, L. N., PANDA P. A predictive regression-type estimator in two-stage sampling. *Journal of the Indian Society of Agricultural Statistics*, 1999, 52(3), pp. 303–308.

SAHOO, L. N., SAHOO, R. K. Predictive estimation of finite population mean in two phase sampling using two auxiliary variables. *Journal of the Indian Society of Agricultural Statistics*, 2001, 54(2), pp. 258–264.

SAINI, M. A class of predictive estimators in two-stage sampling when auxiliary character is estimated at SSU level. *International Journal of Pure and Applied Mathematics*, 2013, 85(2), pp. 285–295.

SINGH, H. P., SOLANKI, R. S. A new procedure for variance estimation in simple random sampling using auxiliary information. *Statistical Papers*, 2013, 54(2), pp. 479–497.

SINGH, H. P., SOLANKI, R. S. Improved estimation of population mean in simple random sampling using information on auxiliary attribute. *Applied Mathematics and Computation*, 2012, 218(15), pp. 7798–7812.

SRIVASTAVA, S. K. Predictive estimation of finite population mean using product estimator. *Metrika*, 1983, 30, pp. 93–99.

STEEL, R. G. D., TORRIE, J. H. *Principles and Procedures of Statistics*. New York, USA: McGraw, 1960.

Aviation Demand and Economic Growth in the Czech Republic: Cointegration Estimation and Causality Analysis

Bilal Mehmood[1] | *Government College University, Lahore, Pakistan*
Amna Shahid[2] | *Government College University, Lahore, Pakistan*

Abstract

The main purpose of the paper is to empirically examine the aviation-led growth hypothesis for the Czech Republic by testing causality between aviation and economic growth. We resort to econometric tests such as unit root tests and test of cointegration purposed by Johansen (1988). Fully Modified OLS, Dynamic OLS and Conical Cointegration Regression are used to estimate the cointegration equation for time span of 42 years from 1970 to 2012. Empirical results reveal the existence of cointegration between aviation demand and economic growth. Graphic methods such as Cholesky impulse response function (both accumulated and non-accumulated) and variance decomposition have also been applied to render the analysis rigorous. The positive contribution of aviation demand to economic growth is similar in all three estimation techniques of cointegration equation. Finally, Granger causality test is also applied to find the direction of causal relationship. Findings help in lime-lighting the importance of aviation industry in economic growth for a developing country like the Czech Republic.

Keywords

Aviation, economic growth, Unit Root Tests, Fully Modified Ordinary Least Square (FMOLS), Dynamic Ordinary Least Square (DOLS), Conical Cointegration Regression (CCR), Aviation Multiplier

JEL code

L93, O40, C22

INTRODUCTION

Role of transportation has been pivotal in transporting of human beings (services) and goods since historic times. Economic activities, both from production (supply) and consumption (demand) side depend on transportation. This paper analyses 'aviation/air transportation' as covariate in association with economic growth. Recent work dealing with this issue has shown positive effects of aviation on economic growth of a country. Nearly no heed has been paid to the empirical analysis of the relationship between economic growth and aviation of the Czech Republic. This is a justification of this research. The aim

[1] Department of Economics, Government College University, Lahore, Pakistan. E-mail: digital.economist@gmail.com.
[2] Department of Economics, Government College University, Lahore, Pakistan. E-mail: samna10@hotmail.com.

of this research is to explore the causal relationship between aviation and economic growth in the Czech Republic. To measure aviation, we used 'passengers carried by air transport' (PAX). While for incorporating economic growth, GDP in constant local currency unit is used. For statistical analysis, this paper resorts to econometric tests such as unit root tests (ADF and Phillips Perron) and test of cointegration purposed by Johansen (1988). The time span covered by the study is the period from 1970 to 2012. This paper scrutinizes the relationship between aviation and economic growth by applying the Johansen cointegration approach for the long-run and the standard error correction method (ECM) for the short-run. This paper contributes to the existing methodology in Marazzo et al. (2010) by using FMOLS, DOLS and CCR to estimate cointegrating equations. Estimation of cointegration equations is becoming a popular practice. For recent application of FMOLS, see Mehmood, et al. (2012).

1 LITERATURE REVIEW

Empirical work on aviation-led economic growth is still in its infancy. A few existing examples of it are reviewed as follows. Beneš, et al. (2008) discuss the development of the transport sector of the Czech Republic. Before 1989, there was a planned economy while after the advent of market economy, main focus was placed on the market of developed European countries and especially on the transport sector covering individual transport systems, transport preferences and transported commodities. They faced many difficulties because although most changes were favorable for meeting transport demands in domestic and global magnitudes but there were many additional problems, too. Thus, authors have recommended to focus on the efficiency of transport systems, with a special emphasis on quality, infrastructure development, lower energy demands, environmental protection and, most importantly, on investment in this sector as this contributes to the GDP of the country very well.

Pioneering research on aviation-growth nexus is conducted by Marazzo et al. (2010). They empirically tested the relationship between aviation demand and GDP for Brazil. They used passenger-kilometer as a proxy of aviation demand and found a long-run equilibrium between the two variables using bi-variate Vector Autoregressive Model. Their findings reveal strong positive causality between GDP and aviation demand, and relatively weaker causality the other way round. Robustness tests were applied through Hodrick and Prescott filter to capture the cyclical components of the series and the results withstood these robustness tests. Their interpretation of positive causality indicates the existence of multiplier effect. Oxford Economic Forecasting (2009) performed some quantifications affirming that Czech aviation sector generates economic benefits for its customers and international economy. Analysis of economic indicators shows that 0.7% of the Czech GDP and 31 400 jobs or 0.6% of the Czech labor force is attributed to the Czech aviation sector. Including the contribution of tourism sector, GDP upsurges to 0.9% and job creation increases to 42 900 jobs (or 0.9% of the labor force). Czech-based carriers were responsible for 67% of passengers carried and 39% of freight. All the income and revenues by these air companies have generated aviation multiplier effects on the Czech economy. Macroeconomic significance of Czech aviation are highlighted in this work.

Mehmood & Kiani (2013) examine the aviation-led growth hypothesis for Pakistan by testing Granger causality between aviation and economic growth using unit root tests and cointegration tests. Using the data from 1973 to 2012, they innovated the work of Marazzo et al. (2010) by applying Fully Modified OLS and Dynamic OLS for the estimation of cointegration equation. Estimations reveal that positive contribution of aviation demand to economy is more prominent as compared to that of economic growth to aviation demand. They found out that positive contribution of aviation demand to economic growth is similar in both FMOLS and DOLS. To our knowledge no further instances of research on the Czech aviation exist. To significantly add to empirical literature, this paper aims at analyzing the aviation-growth nexus for the Czech Republic. Specific testable proposition is as follows:

P_A: *There exists a (Granger) causal relationship between Aviation Demand and Economic Growth in the Czech Republic.*

For scrutinizing the above set proposition, data dimensions and sources are explained below. Moreover, detailed explanation of the analysis methodology is provided:

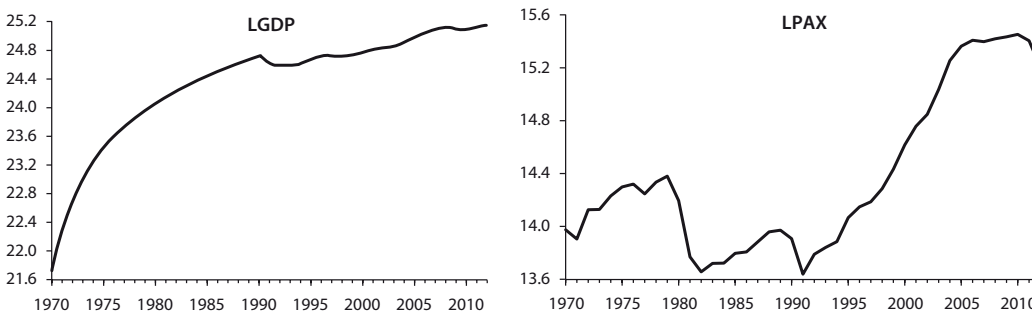
2 DATA AND METHODOLOGY

Borrowing from Marazzo et al. (2010), the demand for aviation is represented by 'air transport, passengers carried' and economic growth by GDP is used in local currency (at constant terms). Data for these variables is taken from World Development Indicators (WDI). For the Czech Republic data on passengers carried and GDP is available from 1970 to 2012. The time span allows us to use 43 observations for our time series analysis. EViews 8 is used for all estimations. Before conducting the inferential analysis, line chart and descriptive analysis is furnished.

3 DESCRIPTIVE STATISTICS

Economic growth is proxied by GDP (current LCU), while demand for aviation is proxied by 'passengers carried by air transport' (PAX). The line charts of GDP (current LCU) and passengers carried by air transport are plotted against time in years. Both of these shows trend and intercepts. This information will be helpful in conducting the stationarity tests.

Figure 1 Chart of GDP and PAX (Natural logged forms for GDP and PAX)



Note: Line charts of GDP and PAX are plotted that show intercept (constant) and trend (slope) in both the variables.
Source: World Development Indicators, own construction

4 INFERENTIAL ANALYSIS

4.1 Stationarity Tests

Both stationarity tests, Augmented Dickey Fuller (ADF) and Phillip Peron (PP), are applied with the assumptions that GDP and PAX in their logarithmic form reveal intercept and trend. Both variables are stationary at first level using ADF and PP tests. So GDP and PAX are stationary at first difference i.e. I(1). Such is tabulated in Table 1.

4.2 Augmented Dickey Fuller Test

For scrutinizing non-stationarity in a time series Augmented Dickey–Fuller test (ADF) test was purposed by Dickey and Fuller (1979). In order to check if the series carry one unit root, the ADF test presents the following specification:

$$\Delta Y_t = \alpha + \beta T + \varphi Y_{t-1} + \sum_{i=1}^{p} \Delta Y_{t-i} + \varepsilon_t, \quad (1)$$

where Y_t and ΔY_t are respectively the level and the first difference of the series, T is the time trend variable, and α, β, φ, ψ are parameters to be estimated. The p *lagged* difference terms are added in order to remove serial correlation in the residuals.

The null hypothesis is H0: $\varphi \neq 0$ and the alternative hypothesis is H_1: $\varphi \neq 0$. ε_t is the error term presenting zero mean and constant variance. First order integrated series can present stationary linear combinations (I(0)). In these cases, we say variables are cointegrated. It means there is a long-run equilibrium linking the series, generating a kind of coordinated movement over time. In order to assess the existence of cointegration between I(1) series, Engle and Granger (1987) proposed a regression between two non-stationary variables (Y_t, X_t) to check the error term integration order. If the error term is stationary one can assume the existence of cointegration.[3] Thus:

$$Y_t = \alpha + \beta X_t + \mu_t \quad (2)$$

is an equation of cointegration if μt is stationary. This condition can be evaluated through the ADF test. A more recent approach is provided by Johansen and Juselius (1990). They suggested an alternative method which has been applied under the following specification:

$$\Delta Y_t = \Pi Y_{t-1} + \sum_{i=1}^{p-1} \Gamma_i \Delta Y_{t-1} + \beta X_t + \varepsilon_t, \quad (3)$$

where Y_t is a vector of 'k' non-stationary variables, X_t is a vector of d deterministic variables and ε_t is a vector of random terms (zero mean and finite variance). The number of cointegration relations is represented by the rank of Π coefficient matrix. The Johansen method relies on estimating the P matrix in an unrestricted form and testing whether it is possible to reject the imposed restrictions when reducing the rank of Π. The maximum likelihood test, which checks the hypothesis of a maximum number of r cointegration vectors, is called the trace test. It should be highlighted that variables under cointegration analysis should present the same integration order. If one concludes that cointegration exists in (3), then there is at least one stationary variable that may be included in the model. This representation is known as Error Correction Model (ECM), specified as follows:

$$\Delta Y_t = \lambda + \sum_{i=1}^{m} \alpha_i \Delta Y_{t-i} + \sum_{j=1}^{n} \beta_j \Delta X_{t-j} + \phi Z_{t-1} + \varepsilon_t, \quad (4)$$

where λ is the constant term, α, β, φ are coefficients, m and n are the required number of lags to make the error term ε_t a white noise and Z_{t-1} is the cointegration vector ($Z_{t-1} = Y_{t-1} - \delta X_{t-1}$), where δ is a parameter to be estimated). In this case, Z_{t-1} works as an error correction term (ECT). The ECT provides valuable information about the short run dynamics between Y and X. In Eq. (4), all the terms are I(0).

4.3 Phillip Perron Test

Phillips and Perron (1988) propose an alternative (nonparametric) method of controlling for serial correlation when testing for a unit root. The PP method estimates the non-augmented DF test equation $[\Delta y_t = \alpha y_{t-1} + x_t'\delta + \epsilon_t]$ and modifies the t-ratio of the α coefficient so that serial correlation does not affect the asymptotic distribution of the test statistic. The PP test is based on the statistic:

[3] For more see Bouzid (2012).

$$\bar{t}_{\alpha} = t_{\alpha}\left(\frac{\gamma_0}{f_0}\right)^{1/2} - \frac{T(f_0-\gamma_0)\left(se(\hat{\alpha})\right)}{2f_0^{1/2}s}, \tag{5}$$

where $\hat{\alpha}$ is the estimate, and t_{α} the t-ratio of α, $se(\hat{\alpha})$ is coefficient standard error, and s is the standard error of the test regression. It is a consistent estimate of the error variance in equation (1) (calculated as $(T - k)s^2/T$, where k is the number of regressors). The remaining term, f_0, is an estimator of the residual spectrum at frequency zero.

Table 1 ADF and PP Tests

Using constant and trend	Stationarity	Variables	t-Statistic	Prob. value
I	II	III	IV	V
Augmented Dickey Fuller (ADF)	At level	GDP	–2.3707	0.3886
		PAX	–1.5658	0.7892
	At first difference	ΔGDP	–10.9086	0.0000
		ΔPAX	–3.7524	0.0298
Phillips & Perron (PP)	At level	GDP	–3.3445	0.0731
		PAX	–1.4104	0.8434
	At first difference	ΔGDP	–13.1762	0.0000
		ΔPAX	–3.7524	0.0298

Note: (i) t-statistics estimates listed in column IV. (ii) ADF and PP tests of GDP show stationarity at 1st difference with significance at all levels (1%, 5% & 10%) while of PAX show stationarity at 1st difference with significance at 5% & 10%.
Source: World Development Indicators, own construction

Johansen cointegration test is applied on the variables of concern and mathematically this is expressed in equation (6) and (7):

$$\Delta PAX_t = \alpha_1 + \sum_i \alpha_{11}(i)\Delta PAX_{t-i} + \sum_i \alpha_{12}(i)\Delta GDP_{t-i} + \beta_1 Z_{t-1} + e_{1t}, \tag{6}$$

$$\Delta GDP_t = \alpha_1 + \sum_i \alpha_{21}(i)\Delta PAX_{t-i} + \sum_i \alpha_{22}(i)\Delta GDP_{t-i} + \beta_2 Z_{t-1} + e_{2t}. \tag{7}$$

Here ΔPAX_{t-i} and ΔGDP_{t-i} are the lagged differences which seize the short term disturbances; e_{1t} and e_{2t} are the serially uncorrelated error terms and Z_{t-1} is the error correction (EC) term, which is obtained from the cointegration relation identified and measures the magnitude of past disequilibrium.

Table 2 Johansen-Juselius Likelihood Cointegration Tests

Null	Alternative	Statistic (GDP & PAX)	Critical Value (95%)
I	II	III	IV
Maximal eigenvalue test			
$\gamma = 0$	$\gamma = 1$	15.7949	17.1477
Trace test			
$\gamma = 0$	$\gamma \geq 1$	15.7949	17.1477

Note: (i) Values of Maximal eigenvalue test and Trace tests. (ii) Optimum lag length is '2' in this case which is selected using the SIC and AIC.
Source: World Development Indicators, own construction

Maximal eigenvalue test and Trace tests reveal the existence of one cointegrating vector. Cointegration is evidenced, using which estimation of cointegrating equations is conducted in the next step.

4.4 Vector Error Correction Model

The model is a first order VEC (Vector Error Correction) model as shown in equation (6) & (7). The lag length was found to be '2' which is established on the basis of SI and AI criteria. Based on column 1 of Table 2, the cointegration vector confirms the expected positive relationship between aviation demand and economic growth (1 PAX = 0.9681 GDP).

Succeeding in uncovering of the cointegration between GDP and PAX, an Error Correction Model (ECM) is estimated for scrutinizing short and long-run causality. In the ECM, the first difference of each endogenous variable (GDP or PAX) was regressed on a one period lag of the cointegrating equation and lagged first differences of all the endogenous variables in the system.

Table 3 shows the results of causality test. We have performed several tests for Granger causality: (1) short-run causality — the significance of the sum of lagged terms of each explanatory variable by joint F test; (2) long-run causality — the significance of the error–correction terms by t-test; and (3) short-run adjustment to re-establish long-run equilibrium — the joint significance of the sum of lagged terms of each explanatory variable and the Error Correction Term (ECT) by joint F test. The lag of the system is decided by AIC criterion as 5.

Short-run causality is found only from PAX to GDP, but not the reverse, i.e. there is unidirectional Granger causality. The coefficient of the ECT is found to be significant in GDP equation, which shows that given any deviation in the ECT, both variables in the ECM would interact in a dynamic fashion to restore long-run equilibrium. Results of the significance of interactive terms of change in PAX, along with the ECT in the GDP equation are consistent with the existence of Granger-causality running from PAX to GDP. These indicate that whenever there is the presence of a shock to the system, PAX would make short-run adjustment to re-establish long-run equilibrium.

Table 3 Estimation results of Error Correction Model for logarithmic series of GDP and PAX

	Source of causality				
	Short-run		Error Correction Term	Joint short/long term test	
Variables	ΔGDP	ΔPAX	ΔGDP	ΔGDP	ΔGDP
	F-statistics		t-statistics	F-statistics	
ΔGDP	–	3.3538**	–2.1018**	–	8.1637***
ΔPAX	0.7169	–	–0.0580	0.6502	–

Note: ΔGDP and ΔPAX are the first difference series of GDP and PAX respectively. **is 5% critical level and *** is 1% critical level.
Source: World Development Indicators, own construction

4.5 Cointegration Equation Estimation

Cointegrating equation is estimated using recently developed econometric methodologies, namely: fully modified ordinary least squares (FMOLS) of Phillips and Hansen (1990), dynamic ordinary least squares (DOLS) technique of Stock and Watson (1993) and Conical Cointegration Regression (CCR) of Park (1992). These methodologies provide a check for the robustness of results and have the ability to produce reliable estimates in small sample sizes.

[3] For more see Bouzid (2012).

4.5.1 Fully Modified Ordinary Least Squares (FMOLS)

On the basis of VAR model results, cointegrating regression is estimated. In a situation, where the series are cointegrated at first difference 'I(1)', Fully modified ordinary least square (FMOLS) is suitable for estimation. FMOLS is attributed to Phillips and Hansen (1990) to provide optimal estimates of cointegrating regressions. FMOLS modifies least squares to explicate serial correlation effects and for the endogeneity in the regressors that arise from the existence of a cointegrating relationship.[4]

$$X_t = \hat{\Gamma}_{21} D_{1t} + \hat{\Gamma}_{21} D_{1t} + \hat{\epsilon}_t, \tag{8}$$

or directly from the difference regressions:

$$\Delta X_t = \hat{\Gamma}_{21} \Delta D_{1t} + \hat{\Gamma}_{21} \Delta D_{1t} + \hat{\upsilon}_t. \tag{9}$$

Let $\hat{\Omega}$ and $\hat{\Lambda}$ be the long-run covariance matrices computed using the residuals $\hat{\upsilon}_t = (\hat{\upsilon}_{1t}, \hat{\upsilon}_{2t}')'$. Then we may define the modified data:

$$y_t^* = y_t - \hat{\omega}_{12} \hat{\Omega}_{22}^{-1} \hat{\upsilon}_2. \tag{10}$$

An estimated bias correction term:

$$\lambda_{12}^* = \lambda_{12} - \hat{\omega}_{12} \hat{\Omega}_{22}^{-1} \hat{\Lambda}_{22}. \tag{11}$$

The FMOLS estimator is given by:

$$\hat{\theta} = \begin{bmatrix} \hat{\beta} \\ \hat{\gamma}_1 \end{bmatrix} = (\Sigma_{t=1}^{T} Z_t Z_t')^{-1} \left(\Sigma_{t=1}^{T} Z_t y_t^* - T \begin{bmatrix} \hat{\lambda}_{12}^* \\ 0 \end{bmatrix} \right), \tag{12}$$

where $Z_t = (X_t', D_t')'$. The key to FMOLS estimation is the construction of long-run covariance matrix estimators $\hat{\Omega}$ and $\hat{\Lambda}$. Before describing the options available for computing $\hat{\Omega}$ and $\hat{\Lambda}$, it will be useful to define the scalar estimator:

$$\hat{\omega}_{1.2} = \hat{\omega}_{11} - \hat{\omega}_{12} \hat{\Omega}_{22}^{-1} \hat{\omega}_{21}, \tag{13}$$

which may be interpreted as the estimated long-run variance of υ_{1t} conditional on υ_{2t}. We may, if desired, apply a degree-of-freedom correction to $\hat{\omega}_{1.2}$.

4.5.2 Dynamic Ordinary Least Square (DOLS)

Dynamic Ordinary Least Squares (DOLS) is attributed to Saikkonen (1991) and Stock & Watson (1993). DOLS is a simple approach to constructing an asymptotically efficient estimator that eliminates the feedback in the cointegrating system. Technically speaking, DOLS involves augmenting the cointegrating regression with lags and leads of so that the resulting cointegrating equation error term is orthogonal to the entire history of the stochastic regressor innovations:

$$y_t = X_t' \beta + D_{1t}' \gamma_1 + \sum_{j=-q}^{r} \Delta X_{t+j}' \delta + \upsilon_{1t}. \tag{14}$$

[4] See Phillips and Hansen (1990) and Hansen (1995) for details.

Under the assumption that adding q lags and r leads of the differenced regressors soaks up all of the long-run correlation between υ_{1t} and υ_{2t}, least-squares estimates of $\theta = (\beta', \gamma')'$ have the same asymptotic distribution as those obtained from FMOLS and Conical Cointegration Regression (CCR).

An estimator of the asymptotic variance matrix of $\hat{\theta}$ may be computed by computing the usual OLS coefficient covariance, but replacing the usual estimator for the residual variance of υ_{1t} with an estimator of the long-run variance of the residuals. Alternately, you could compute a robust HAC estimator of the coefficient covariance matrix.

4.5.3 Conical Cointegration Regression (CCR)

The CCR estimator is based on a transformation of the variables in the cointegrating regression that removes the second-order bias of the OLS estimator in the general case. The long-run covariance matrix can be written as:

$$\Omega = \lim_{n\to\infty} \frac{1}{n} E\left(\sum_{t=1}^{n} u_t\right)\left(\sum_{t=1}^{n} u_t\right)' = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}. \tag{15}$$

The matrix Ω can be represented as the following sum:

$$\Omega = \Sigma + \Gamma + \Gamma', \tag{16}$$

where:

$$\Sigma = \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} E\left(u_t u_t'\right), \tag{17}$$

$$\Gamma = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n-1} \sum_{t=k+1}^{n} E\left(u_t u_{t-k}'\right), \tag{18}$$

$$\Lambda = \Sigma + \Gamma = (\Lambda_1, \Lambda_2) = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}. \tag{19}$$

The transformed series is obtained as:

$$y_{2t}^* = y_{2t} - (\Sigma^{-1}\Lambda_2)' u_t, \tag{20}$$

$$y_{1t}^* = y_{1t} - (\Sigma^{-1}\Lambda_2\beta + (0, \Omega_{12}\Omega_{22}^{-1})')' u_t. \tag{21}$$

The canonical cointegration regression takes the following form:

$$y_{1t}^* = \beta' y_{2t}^* + u_{1t}^*, \tag{22}$$

where:

$$y_{1t}^* = u_{1t} - \Omega_{12}\Omega_{22}^{-1} u_{2t}. \tag{23}$$

Therefore, in this context the OLS estimator of (22) is asymptotically equivalent to the ML estimator. The reason is that the transformation of the variables eliminates asymptotically the endogeneity caused by the long-run correlation of y_{1t} and y_{2t}. In addition (23) shows how the transformation of the variables eradicates the asymptotic bias due to the possible cross correlation between u_{1t} and u_{2t}.

4.6 Comparison of the Cointegration Regression Estimates

Estimates of the three estimates techniques are summarized in the Table 4:

Table 4 Comparison of the Cointegration Regression Estimates using Three Different Techniques

Technique	Coefficient	S.E.	Adj. R^2	Remarks
Fully Modified OLS	1.6958***	0.0204	0.6861	Significant & positive relationship
Dynamic OLS	1.7029***	0.0211	0.5101	Significant & positive relationship
Conical Cointegration Regression	1.7003***	0.0204	0.7116	Significant & positive relationship

Note: All the constants and coefficient estimates are significant at 1%, indicated by***.
Source: World Development Indicators, own construction

Results of all three estimation techniques (FMOLS, DOLS & CCR) for cointegrating regression shows a positive relationship between GDP and PAX. However, DOLS has increased explanatory power of PAX while the adjusted R^2 is highest using CCR. Our major concern, however, is to find the nature of relationship between GDP and PAX, that is found to be positive and significant using all three cointegration equation estimation techniques.

Table 5 Granger Causality Test Results

Null Hypothesis	F-Statistic	Prob.
GDP does not Granger Cause PAX	3.1048	0.022
PAX does not Granger Cause GDP	1.9468	0.115

Source: World Development Indicators, own construction

Results of Granger causality (in-sample approach), in table 5, show that GDP has the tendency to boost the number of passengers carried by aviation sector. While, the causality does not run in opposite direction. This implies that increase in economic activity in the Czech Republic outgrows the economic opportunities of local and international trade that lead to increased mobility of passengers via aviation. As evident from the analysis, economic growth holds valuable information to forecast aviation demand.

CONCLUSION

This paper investigated the cointegration and causality relationships between demand for aviation and economic growth in the Czech Republic. The outcome of this paper implies that aviation and economic growth are cointegrated in the long run and the relationship holds in the short run as well. This can be translated into a multiplier effect. Our innovation introduced into the empirical analysis of estimation of cointegrating vector using FMOLS, DOLS and CCR corroborates the findings in Marazzo et al. (2010) and Mehmood and Kiani (2013).

The positive relationship can be attributed to direct and indirect effects of aviation. Direct effects include transportation of labour force (implicitly of services) and goods. Indirect benefits include benefits that accrue to other industries through backward and forward linkages of aviation industry. This produces further impetus on economic activity and hence growth. In the case of the Czech Republic, the data reveals two breaks during 1989 and 1990, these can be attributed to after effects of oil shocks and increase in flight fares. Further research can be focused on capturing effects of such issues using statistical tools like Andrews (1993). However, this study has pioneering the research in the field of aviation using other sophisticated tools like Fully modified OLS, Dynamic OLS and Conical cointegration regression (CCR). The events that aviation industry should get policy attention to play its further ameliorated role in determining economic growth. Formal incentives should be given to aviation industry

to increase its macroeconomic contribution. The scope of research on aviation can be extended by using cross country analysis.

References

ANDREWS, D. W. Tests for Parameter Instability and Structural Change with Unknown Change Point. *Econometrica*, 1993, Vol. 61, No. 4, pp. 821–856.

BENEŠ, L., BŘEZINA, E., BULICEK, J., MOJŽÍŠ, V. The Development of Transport in the Czech Republic. *European Transport*, 2008, Vol. 39, pp. 33–43.

DICKEY, D., FULLER, W. Distributions of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*, 1979, Vol. 74, pp. 427–431.

ENGLE, R., GRANGER, C. Cointegration and Error Correction: Presentation, Estimation and Testing. *Econometrica*, 1987, Vol. 55, pp. 251–276.

JOHANSEN, S., JUSELIUS, K. Maximum likelihood Estimation and Inference on Cointegration with Application to the Demand for Money. *Oxford Bulletin of Economics and Statistics*, 1990, Vol. 52, pp. 162–211.

MARAZZO, M., SCHERRE, R., FERNANDES, E. Air Transport Demand and Economic Growth in Brazil: A Time Series Analysis. *Transportation Research*, 2010, Vol. 46, pp. 261–269.

MEHMOOD, B., KHAN, A., KHAN, A. Empirical Scrutiny of Demographic Dividend of Economic Growth: Time Series Evidence from Pakistan. *Romanian Review of Social Sciences*, 2012, Vol. 2, pp. 3–11.

MEHMOOD, B., KIANI, K. An Inquiry into Nexus between Demand for Aviation and Economic Growth in Pakistan. *Academicia*. 2013, Vol. 3, No. 10, pp. 200–211.

Oxford Economic Forecasting. *Economic Benefits from Air Transport in the Czech Republic*, London, UK, 2009.

PHILLIPS, P. C., PERRON, P. Testing For a Unit Root in Time Series Regression. *Biometrika*, 1988, Vol. 75, No. 2, pp. 335–346.

PHILLIPS, P. C., HANSEN, B. E. Statistical Inference in Instrumental Variables Regression with I(1) Processes. *The Review of Economic Studies*, 1990, Vol. 57, No. 1, pp. 99–125.

PARK, J. Y. Canonical Cointegrating Regressions. *Econometrica*, 1992, Vol. 60, No. 1, pp. 119–143.

SAIKKONEN, PENTTI. Asymptotically Efficient Estimation of Cointegration Regressions. *Econometric Theory*, 1991, Vol. 7, No. 1, pp. 1–21.

STOCK, J. H., WATSON, M. W. A Simple Estimator of Cointegrating Vectors in Higher Order Integrated Systems. *Econometrica*, 1993, Vol. 61, No. 4, pp. 783–820.

A Simulation Study Comparing Two Methods of Handling Missing Covariate Values when Fitting a Cox Proportional-Hazards Regression Model

Ali Satty[1] | *Elneelain University, Khartoum, Sudan*

Abstract

Missing covariate values is a common problem in a survival data research. The aim of this study is to compare the use of the multiple imputation (MI) and last observation carried forward (LOCF) methods for handling missing covariate values in the Cox proportional hazards (PH) regression model. The comparisons between the methods are based on simulated data. The missingness mechanism is assumed to be missing at random (MAR). Missing covariate values are generated under different missingness rates. The results from both methods are compared by assessing the bias, efficiency and coverage. The simulation results in general revealed that MI is likely to be the best under the MAR mechanism.

Keywords

Missing covariate values, multiple imputation (MI), Cox proportional hazard model , Last observation carried forward (LOCF), missing at random (MAR)

JEL code

C15

INTRODUCTION

One of the challenges in modeling practice is missing data. A problem occurs when some data on covariates are missing in survival analysis, where the Cox proportional-hazards (PH) model (Cox, 1972) is usually used for analysis. Covariate observations may be missing for some individuals, for whatever reason. An important concept with missing data, specifically where there are multiple covariates with missing values, relates to the mechanism of missing data. Rubin (1976, 1987) classified these mechanisms into three basic categories: missing completely at random (MCAR), meaning that the missingness process does not depend on the observed responses, missing at random (MAR), when the missingness process depends on the observed responses and probably on measured covariates but not on the unob-

[1] School of Statistics and Actuarial Science, Elneelain University, Khartoum, Sudan. E-mail: Alisatty1981@gmail.com.

served responses, and missing not at random (MNAR) which allows the missingness process to depend on the unobserved responses as well as on the observed responses.

Given the problems that can arise in the Cox PH model when there are missing covariate values, the following question is forced upon researchers. What methods can be utilized to handle these potential pitfalls? The goal is to use approaches that better avoid the generation of biased results. There are several ways to deal with missing covariate values in Cox PH model. One recommendation is to discard subjects with incomplete sequences, and then analyze only the units with complete data. Method that uses this solution is called complete case analysis (CC) (Little and Rubin, 1987). However, this method has numerous disadvantages leading to reduction in the sample size, which reduces the precision of estimates and therefore can lead to biased results (Schafer and Graham, 2002).

In contrast to the CC analysis, there are other ways that can help to tackle the problem of missing covariate values in Cox PH model. There are methods that do generate possible values for the missing covariates. These methods are called imputation methods, where one fills-in (imputes) the missing covariate values to obtain a full dataset, and the resultant data are then analyzed by standard statistical methods without concern as if the set represented the true and complete dataset (Rubin, 1987; Little and Rubin, 1987). This is the key idea behind commonly used procedures for imputation which include, simple and multiple imputations (Little and Rubin, 1987). There are different simple imputation methods. In this study however we restrict ourselves to outlining one of them, which is called last observation carried forward (LOCF). LOCF substitutes one value for every missing covariate value in the dataset (Little and Rubin, 1987, 2002). Under certain restrictive circumstances, LOCF can produce unbiased results. In addition, in some situations, LOCF does not produce conservative results. However, this approach can still provide conservative results, under some specific circumstances. The method will be readdressed in detail in the following section. In contrast to the LOCF method, MI fills in more than one value for each missing covariates item and carries out the analysis as if the imputed values were observed data to allow for the appropriate evaluation of imputation uncertainty (Rubin, 1987; Little and Rubin, 1987). MI was proposed by Rubin (1978) and described in detail by Little and Rubin (1987). Considerable research has focused on MI for handling missing covariate data in Cox PH model (See, Paik, 1997; van Buuren et al.1999; Brazi and Wooward, 2004; White and Royston, 2009).

This study deals with the problem of missing covariate values in the Cox regression model. It is devoted to a comparison of two imputation techniques or methods. The methods that were compared include multiple imputation (MI) and last observation carried forward (LOCF). The main objective of this paper is to study imputation techniques and compare them with others to estimate Cox PH model parameters with missing covariates values. The missing data mechanism is assumed to be MAR. The comparisons are based on a simulation study. The comparisons are made through the evaluation of bias, efficiency, and coverage. The rest of this paper is organized as follow: Section 1 describes the notation and model assumptions. An overview of methods for analyzing missing covariate values is also given. Section 2 presents the simulation study scheme including the study design, data generation and the evaluation criteria used in the analysis. The results from the simulations of the two methods are presented in section 3. Finally, a brief discussion and concluding remarks are provided in the last section.

1 METHODS

1.1 Notation and model assumptions

Assume there are n independent individuals. For each individual, $i = 1,..., n$. Let c and T be the censoring and failure, respectively. Now, we assume the hazard for individual i follows a Cox proportional hazards regression model:

$$\lambda (t \mid xi) = \lambda_0 (t) \exp (\beta' x_i), \quad (1)$$

where $\lambda 0(t)$ is an unspecified baseline hazard function, x represent a set of independent covariates that may be categorical or continuous, and β is a $p \times 1$ parameter vector. In this study however an application will be confined to the continuous covariates case (i.e., x are continuous covariates). The vector of observed time to follow-up was obtained by $T = \min(T, c)$, and failure indicator vector δ by $\delta = 1$ if $T \leq c$ and $\delta = 0$ if censored. We suppose that x, T and c are independent. We restrict ourselves to consider that the survival time is fully observed, while some of the covariates x_i contains missing values. Now, partition the covariate vector x_i into its observed covariates and missing covariates, such that $x_i = (x^{obs}, x^{mis})$. Let R be a vector that represents the missing covariate process, with $R = 1$ if the covariate is observed (i.e. x^{obs}), and $R_{ij} = 0$ if the covariate is missing (i.e. x^{mis}). When MAR holds, the missing covariate mechanism is determined by the conditional distribution of R conditional upon (Z, δ, x^{obs}), which is Bernoulli with probability $\hbar = P(R = 1 \mid Z, \delta, x^{obs})$, where Z denotes the survival outcome. For each individual, let $(Z_i, \delta_i, x_i^{obs}, x_i^{mis}, R_i)$ denote *i.i.d* copies of $(Z, \delta, x^{obs}, x^{mis}, R)$. Thus the observed covariate data being analyzed are $(Z_i, \delta_i, x_i^{obs}, x_i^{mis})$ if $R = 1$, and $(Z_i, \delta_i, x_i^{obs})$ if $R=0$. There are a variety of methods that can be used to deal with missing covariate values (x^{mis}).The subsections that follow provide a review of the methods that are used in this study.

1.2 Multiple imputation (MI)

Following is a brief description of MI and its application. According to Rubin (1987), MI consists of three steps. First, each missing value is replaced by $M \geq 2$ simulated values. Each of these sets of plausible values can be used to fill-in the missing values and create a completed dataset. This method is valid under the MAR mechanism (Little and Rubin, 1987). Further, when MAR holds, for univariate x^{mis} and given the observed data (Z, δ, x^{obs}), sets of plausible values for missing observations (x^{mis}) can be created to reflect uncertainty about the stochastic non-response model. This can be done using an appropriate imputation model $P(x^{mis} \mid Z, \delta, x^{obs})$. In doing so, SAS PROC MI can be used. PROC MI fills in the missing covariate values and therefore the above univariate method can be conducted to each missing covariate x^{mis} in turn. This can be achieved using all the imputed values of the other missing covariates in case of creating new values of x^{mis}. This process is repeated until a suitable convergence criterion is satisfied. Second, each of the M complete datasets are analyzed using standard statistical methods, such as Cox proportional regression model. The use of the number of imputations M needs not be very large since, in practice, 3-10 imputations often provided satisfactory results (Schafer, 1997; Schafer and Olsen, 1998). Finally, the M results are combined using methods that allow for uncertainty regarding the imputation to be taken into account. The steps described earlier are repeated independently M times, resulting in β_m^*, where β_m^* is the parameter estimate of interest from imputation $m = 1,\ldots, M$. Steps 1 and 2 are referred to as the imputation task, and step 3 is the estimation task. Finally, we combine the estimates obtained after M imputations. The results of the M separate analyses (e.g. parameter estimates) are then combined into a single value as:

$$\beta_m^* = \frac{1}{M}\sum_m^M \beta_m^* , \tag{2}$$

where β_m^* is the parameter estimates of interest from imputation $m=1, 2..., M$. The variance for these estimates is composed of two parts: the between imputation variance and within imputation variance. Between imputation variance takes the form:

$$B = \sum_{m=1}^{M} \frac{(\beta_m^* - \beta_m^*)\,(\beta_m^* - \beta_m^*)'}{M-1} . \tag{3}$$

The within imputation variance, U^{-}, is the mean of estimated variances across the M imputations. The total variance for MI is then calculated as:

$$T = \overline{U} + \left(1 + \frac{1}{M}\right) B\,, \tag{4}$$

where:

$$\overline{U} = \sum_{m=1}^{M} \frac{T_m}{M}\,. \tag{5}$$

The MI inference assumes that the analysis model is the same as the model used to impute missing values (the imputation model). Practically, the two models might not be the same (Meng, 1994; Schafer, 1997). The quality of the imputation model influences the quality of the analysis model results and therefore it is important to carefully consider the design of the imputation model. In this study, the imputation model is based on the Cox proportional hazards regression model (1). However, the imputation model for missing covariates requires a valid characterization of the conditional distribution of missing covariates conditional upon the observed data. This problem of the conditional distribution poses a major complication under a Cox PH model. White and Royston (2009) stated that such conditional distribution did not have standard and closed forms for Cox PH model. Thus, one recommendation is to use some of the common regression models to approximate the covariate distribution (Lihong et al., 2009). Following van Buuren et al. (1999) and White and Royston (2009), we used the linear regression model to impute the continuous covariate data. The linear regression model provides an appropriate imputation model for a continuous x_i^{mis}, that is $x^{\text{mis}} \sim \beta 0 + \beta 1\, Z + \beta 2\, \delta + \Delta_3^{T}\, x^{\text{mis}}$. This model includes the following variables as predictors: the survival outcome Z, censoring δ, and the observed covariate x^{obs}. This means we used all the available data (including the outcome variable - survival time) to predict the missing covariate values to make the MAR assumption more plausible as well as to improve the accuracy and efficiency of the imputation. The survival time variable was included in the analysis as the outcome should be included in the imputation model (Moons et al., 2006). This was done to avoid the outcome-covariate association that might be biased toward null using the imputed data (Collins et al., 2001).

1.3 Last observation carried forward (LOCF)

The simplest imputation approach is the LOCF method in which every missing covariates value is replaced by the last observed covariates value from the subject or time series, i.e. it is a method that assumes that the outcomes would not have changed from the last observed value. We refer to Siddiqui and Ali (1998) and Satty and Mwambi (2012) for more details, and where insightful illustrations of the issues of this method are provided in Kenward, and Molenberghs (2009). It is a general and flexible technique for handling missing data, and can be implemented quickly in several statistical softwares. However, with respect to accurately reproducing known population results (parameter estimates and standard errors), the LOCF method has been found to be inadequate (Schafer and Graham, 2002). It shares with other single imputation methods that it tends to create inflated artificial values than truly expected, since imputed values are treated as observed values (Kenward and Molenberghs, 2009). Hence, the variability of the estimators is also underestimated. The problems linked with LOCF include: (1) the performance of this method is poor even when the ignorable missing data mechanism (MCAR or MAR) holds, a situation that limits their suitability to quite a restricted set of assumptions (Allison, 2002); (2) it produces seriously biased results that may or may not be predictable; (3) when using this technique, the standard errors and standard deviations tend to be underestimated, and, therefore, there is a great-

er likelihood of committing type-I error (see, Schafer and Graham, 2002). However, despite these shortcomings, the LOCF method has been recognized as a popular technique in dealing with missing data for the following reasons: its simplicity, in that the method can be quite effective and may be satisfactorily used with small amounts of missing data (Unnebrink and Jurgen, 2001), it is easy to carry out in most statistical software packages but it has varying details of implementations, and in some applications it makes sense to use this technique. LOCF does well when the missingness mechanism is assumed to be MCAR (Unnebrink and Jurgen, 2001). However, because such circumstance is rare, Kenward and Molenberghs (2009) advise that one should avoid this method whenever possible. In general, LOCF might become attractive under specific circumstances.

2 SIMULATION STUDY

We carried out a simulation study to compare the performance of the MI and LOCF methods. The simulations were conducted with 100 replications and sample size n = 1 000 for each replication. We simulated the survival time z_i from an exponential distribution using the following hazard:

$$\eta z = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}), \tag{6}$$

where ($\beta 0, \beta 1, \beta 2, \beta 3$) = (–1.5, 0.5, 1.0, 1.0). That is, the survival time for each individual was distributed according to equation (1). The covariates x_1, x_2 and x_3 were generated from the multivariate normal distributions, i.e., x_1~N (10, 0.25), x_2~N (10, 0.20) and x_3~N (10, 0.20). We assume that z_i observations are randomly censored with probability 0.20. Let R_i be a vector that represents the missing data process, with R_{ij} = 1 if the jth covariate is observed for individual i, and R_{ij} = 0 otherwise, where $i = 1, ..., n$, $j = 1, ..., p$. Let, too, Ri2 and Ri3 represent whether x_{i2} and x_{i3} are unobserved. We created missing covariate mechanism according to the following models:

$$p(R_{i2=0} \mid h_i(z), x_{i,obs}, \theta_2) = \frac{\exp(\theta_{20}+\theta_{21h_i(z)}+\theta_{22x_{i1}})}{1+\exp(\theta_{20}+\theta_{21h_i(z)}+\theta_{22x_{i1}})}, \tag{7}$$

$$p(R_{i3=0} \mid h_i(z), x_{i,obs}, x_{i2}, \theta_3) = \frac{\exp(\theta_{30}+\theta_{31h_i(z)}+\theta_{32x_{i1}}+\theta_{33x_{i2}})}{1+\exp(\theta_{30}+\theta_{31h_i(z)}+\theta_{32x_{i1}}+\theta_{33x_{i2}})}, \tag{8}$$

where θ denotes the parameters of the missingness distribution, $h(z)$ is the observed event times, θ_2 = (–2, 1.5, 2.5) and θ_3 = (–1, 0.5, 0.5, 0.5). We created missing covariate observations under MAR mechanism. Namely, the probability of having a missing covariate values depends on an observed covariate values. The MAR mechanism was generated with the fraction of missing covariates set to 10%, 20% and 30%. Now, after the missing covariate values had been generated, MI was carried out using SAS PROC MI. With PROC MI, we considered the linear regression (Little, 1988) as an imputation model for continuous missing covariates data. PROC MI was applied to generate M = 5 complete datasets. These 5 imputations are often sufficient to obtain satisfactory results (Rubin, 1987; Schafer, 1997). Note that the choice of M =5 was considered adequate and the efficiency of the parameter estimate based on imputation given by $(1+{}^{\nu}/_{M})^{-1}$ here ν is the rate of missing data (Rubin, 1987). This formula shows that the relative efficiency of the MI inference is related to the missingness rate (ν) in combination with the number of imputations (M). For 10%, 20% and 30% rates of missing data and estimates based on M = 5 implies we achieve at least 98%, 96% and 94% efficiency, respectively. A Cox PH model was then fitted to each completed dataset using SAS procedure PHREG to estimate the overall parameters. A Cox PH model that we considered is based on (6). Thereafter, results of the analysis from these 5 completed (imputed) datasets were combined into a single inference using SAS PROC MIANALYZE.

The simpler LOCF technique replaced the missing covariate values by the last available observed values, and once the dataset has been completed in this way, it is analyzed as if it were fully observed. LOCF was conducted by using a macro in SAS software. After applying LOCF, as above introduced, the same model (6) as before being fitted is analyzed. In model (6), if x_i has missing covariate value, it will be filled in by the previous observed covariate value x_{i-1}. Comparisons of MI and LOCF were assessed using criteria recommended in Schafer and Graham (2002): (1) Bias of the estimates: the difference between the average of the 1000 coefficient estimates and the corresponding true coefficient. Thus a better approach that does on the average presents the population value with less bias. (2) The efficiency: the variability of the estimates around the true population coefficient. It was measured in this study by the average width of the 95% confidence interval. Thus, a wider interval implies a less efficient technique. (3) The coverage of the confidence interval: the percentage of 95% confidence intervals estimates across 1000 replicates. If a method is working well, the actual coverage should be close to the nominal rate (95%).

3 RESULTS

The results obtained from a Cox PH model (6) for the bias, efficiency and coverage of the MI and LOCF methods, under different missing covariate values rates are presented in Tables 1, 2 and 3. Note that the largest bias and less efficiency for each given estimate appear in bold.

Table 1 Bias, Efficiency and Coverage of MI and LOCF, under 10% missing covariate values

Rate	Method	Parameter	Bias	Efficiency	Coverage true
	MI	$\beta 1$	0.006	1.158	0.971
		$\beta 2$	**0.018**	**1.113**	0.974
		$\beta 3$	0.017	1.116	0.967
10%					
	LOCF	$\beta 1$	**0.051**	**1.176**	0.902
		$\beta 2$	0.011	1.112	0.911
		$\beta 3$	**0.022**	**1.171**	0.908

Note: MI=multiple imputation; LOCF=last observation carried forward.
Source: Own construction

Under 10% missing covariates rate, the results of MI and LOCF in terms of bias, efficiency and coverage, are displayed in Table 1. By looking at this table we find the following. With respect to biasdness of the estimates, the performance of MI was unsurprisingly, better than that for LOCF. However, the LOCF based estimates were closer to those based on MI, and only slightly less biased in estimating x2. Efficiency estimates associated with LOCF were slightly elevated when compared to those with MI. The MI method was more efficient in most cases, except for x2. For coverage criterion, according to Schafer and Graham (2002), the performance of a method can be regarded to be poor if its coverage drops below 90%, and hence leads to substantially increased Type-I error rate. By this rationale, both approaches yielded acceptable coverage of parameters. Their coverage rates were consistently above 90%.

An examination of Table 2, for 20% missing covariate rate, reveals that among the methodologies examined here, LOCF was notable for consistently producing the most biased estimates vis-a-vis those in the MI method. Namely, treating the data with MI appears to have resulted in fairly minor bias. MI yielded equally acceptable performance across all covariates. Comparing the efficiency results, just as was the case in Table 1, efficiency by LOCF appeared to be independent of the missing covariate rates,

meaning the MI method yielded more efficient estimates under 20%. MI resulted in smaller estimates than estimates of LOCF. Differences in efficiency estimates between the 10% and 20% missing covariate rates were more pronounced for LOCF than for MI. Coverage rates obtained by the LOCF method in all cases were unsatisfactory, as its coverage rates were less than 90%.

Table 2 Bias, Efficiency and Coverage of MI and LOCF, under 20% missing covariate values

Rate	Method	Parameter	Bias	Efficiency	Coverage true
	MI	$\beta 1$	0.011	1.177	0.960
		$\beta 2$	0.064	1.181	0.966
		$\beta 3$	0.051	1.152	0.957
20%					
	LOCF	$\beta 1$	**0.067**	**1.801**	0.891
		$\beta 2$	**0.089**	**1.811**	0.881
		$\beta 3$	**1.030**	**1.852**	0.889

Note: MI=multiple imputation; LOCF=last observation carried forward.
Source: Own construction

Considering the 30% missing covariate values, the results shown in Table 3 reveal that in nearly all cases, LOCF consistently produced the most biased estimates. The efficiency performance was acceptable for MI but low for all parameters under LOCF. In general, the MI method tends to have the smallest estimates for efficiency condition. Thereby, it was more efficient than LOCF. With respect to coverage condition investigated, similar to the findings obtained under 10% and 20% missing covariate values, MI produced uniformly acceptable coverage; none was less than 90%. The LOCF's coverage at 95% was consistently lower than 90%. This coverage was indicative a seriously low level of coverage as 90% corresponds to a doubling of the nominal rate of error (0.05). As can be seen in the results, the low coverage rates by LOCF can also be attributed to its large biases.

Generally speaking, across all missing covariate rates, the worst performance for analyses run with LOCF occurred for the highest missing covariate rate, and declined in relative magnitude as the missingness rate decreased. In other words, when the missing covariate rate decreased to 10%, the results from LOCF became nearly closer to those of MI, but for 20% and 30%, it has seriously less efficient estimates.

Table 3 Bias, Efficiency and Coverage of MI and LOCF, under 30% missing covariate values

Rate	Method	Parameter	Bias	Efficiency	Coverage true
	MI	$\beta 1$	0.054	1.801	0.951
		$\beta 2$	0.098	1.826	0.942
		$\beta 3$	0.102	1.900	0.938
30%					
	LOCF	$\beta 1$	**1.124**	**2.522**	0.862
		$\beta 2$	**1.021**	**2.091**	0.859
		$\beta 3$	**1.205**	**2.511**	0.849

Note: MI=multiple imputation; LOCF=last observation carried forward.
Source: Own construction

DISCUSSIONS AND CONCLUSION

This study has discussed the performance of using the MI and LOCF methods for handling missing covariate values in survival analysis. The main objective was to address and compare the use of these methods when there are missing covariate values in Cox PH regression model. The methods were compared on simulated data. Missing covariate values were generated under three missingness rates. The missing data mechanism was assumed to be MAR. The comparisons between the two methods were made through the evaluation of bias, efficiency and coverage. Based on the simulation results, we reached the following conclusions:

- The results in general revealed that MI is likely to be the best under the MAR mechanism. MI consistently outperformed LOCF in terms of bias, efficiency and coverage. This advantage for the MI method is well documented in terms of the MAR mechanism (Little and Rubin, 1987; Schafer, 1997).
- The findings further suggested the inappropriateness of LOCF analysis. LOCF can lead loss in power of the covariates and imprecise parameter estimates. To avoid this problem, an application of MI can be utilized to handle this potential pitfall. Moreover, it appeared that no strong differences were seen between MI's results and those for LOCF when the missing data rate was low (10%). This indicates that the LOCF method can be applied if the proportion of missing covariate values is low. This LOCF situation is well stated in Unnebrink and Jurgen (2001) and Halabi et al. (2003). It would appear that Kenward and Molenberghs's (2009) recommendation to avoid the LOCF analysis whenever possible is supported by the current analysis.
- As missingness mechanism was simulated to be MAR, the current simulation results has shown clearly that the LOCF's performance was unsatisfactory under this assumption. This situation can be justified by some previous studies which show that LOCF is more widely used under MCAR than under MAR (See, Siddiqui and Ali, 1998; Halabi et al., 2003; Kenward and Molenberghs, 2009). Therefore, the better ways of dealing with missing covariate values in Cox PH model and the best method should be dependent on the nature of the missing covariate values mechanism. Conse quently, one needs to know why are there missing covariate values, and under which mechanism they are missing.
- In conclusion, we recommend that some techniques or methods use different approaches to address missing covariates in Cox PH model. The literature presents various techniques that can be used to deal with missing covariate values in Cox PH model, and these range from simple classical ad hoc methods to model-based methods. These methods should be fully understood and appropriately characterized in relation to missing data and should be theoretically proved before they are used practically. Additionally, each method is based on a specific missingness mechanism, but one needs to realize that at the heart of the missingness problem it is impossible to identify the missing data mechanism.

References

ALLISON, P. D. *Missing data*. Thousand Oaks, CA: Sage, 2002.

BARZI, F., WOODWARD, M. Imputations of missing values in practice: results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, 2004, 160, pp. 34–45.

COLLINS, L. M., SCHAFER, J. L., KAM, C. M. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 2001, 6, pp. 330–351.

COX, D. R. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, Series B, 1972, 34, pp. 187–220.

HALABI, S., WUN, C., DAVIS, B. R. Analysis of survival data with missing measurements of a time-dependent binary covariate. *Journal of Biopharmaceutical Statistics*, 2003, 13, pp. 253–270.

KENWARD, M. MOLENBERGHS, G. Last observation carried forward: A crystal ball. *Journal of Biopharmaceutical Statistics*, 2009, 872, pp. 872–888.
LIHONG, QI., YING-FANG, W., YULEI, HE. A comparison of multiple imputation and fully augmented weighted estimators for Cox regression with missing covariates. *Statistics in Medicine*, 2009, 29, pp. 2592–2604.
LITTLE, R., RUBIN, D. B. *Statistical analysis with missing data*. New York: John Wiley, 1987.
LITTLE, R. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 1988, 6, pp. 287–296.
LITTLE, R., RUBIN, D. B. *Statistical analysis with missing data* (2nd ed.). New York: John Wiley and Sons, 2002.
MENG, X. L. Multiple imputation inferences with uncongenial sources of input (with discussion). *Statistical Science*, 1994, 10, pp. 538–573.
MOONS, K. G., DONDERS, R. A., STIJNEN, T., HARRELL, JR. F. E. Using the outcome for imputation of missing predictor values was preferred. *Journal of Clinical Epidemiology*, 2006, 59, pp. 1092–1101.
PAIK, M. C. Multiple imputation for the Cox proportional hazards model with missing covariates. *Lifetime Data Analysis*, 1997, 3, pp. 289–298.
RUBIN, D. B. Inference and missing data. *Biometrika*, 1976, 63, pp. 581–592.
RUBIN, D. B. Multiple imputation in sample survey. *Proc. Survey Res. Meth. Sec.*, Am. Statist. Assoc., 1978, pp. 20–34.
RUBIN, D. B. *Multiple imputation for non-response in surveys.* Wiley: New York, 1987.
SATTY, A. MWAMBI, H. Imputation methods for estimating regression parameters under a monotone missing covariate pattern: A comparative analysis. *South African Statistical Journal*, 2012, 46, pp. 327–356.
SCHAFER, J. L. *Analysis of incomplete multivariate data*. London: Chapman and Hall, 1997.
SCHAFER, J. L., GRAHAM, J. W. Missing data: Our view of the state of the art. *Psychological Methods*, 2002, 7, pp. 147–177.
SIDDIQUI, O., ALI, M. W. A comparison of the random-effects pattern mixture model with last observation carried forward (LOCF) analysis in longitudinal clinical trials with dropouts. *Journal of Biopharmaceutical Statistics*, 1998, 8, pp. 545–563.
UNNEBRINK, K., JURGEN, W. Intention-to-treat: methods for dealing with missing values in clinical trials of progressively deteriorating diseases. *Statistics in Medicine*, 2001, 20, pp. 3931–3946.
VAN BUUREN, S., BOSHUIZEN, H. C., KNOOK, D. L. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 1999, 18, pp. 681–694.
WHITE, I. R., ROYSTON, P. Imputing missing covariate values for the Cox model. *Statistics in Medicine*, 2009, 28, pp. 1982–1998.

Cluster Analysis of Economic Data

Hana Řezanková[1] | *University of Economics, Prague, Czech Republic*

In the paper, some classical and recent approaches to cluster analysis are discussed. Over the last decades researchers focused mainly on categorical data clustering, uncertainty in cluster analysis and clustering large data sets. In this paper some of the recently proposed techniques are introduced, such as similarity measures for data files with nominal variables, algorithms which include uncertainty in clustering, and the method for data files with many objects.

Keywords

Cluster analysis, similarity measures, hierarchical clustering, k-clustering, fuzzy clustering, large data sets

JEL code

C10, C38

INTRODUCTION

Cluster analysis is a strong tool of the multivariate exploratory data analysis. It involves a great amount of techniques, methods and algorithms which can be applied in various fields, including economy. However, in most of research papers containing cluster analysis of economic data the classical basic approaches are only applied. In this paper some clustering algorithms proposed in the last decades are introduced.

The aim of cluster analysis is to identify groups of similar objects (countries, enterprises, households) according to selected variables (unemployment rate of men and women in different countries, deprivation indicators of households, etc.). The basic approaches are hierarchical clustering and k-means clustering. There are many types of these techniques.

Agglomerative hierarchical clustering, which is usually applied, starts with objects regarded as individual clusters. The clusters are stepwise linked until all objects are connected in one cluster. In k-means clustering, objects are assigned to a certain number of clusters. In both methods the analyst needs to have some tools for determining the number of clusters. In hierarchical cluster analysis it can be done intuitively via a dendrogram, in *k*-means clustering the objects are usually assigned to different numbers of clusters and according to selected criteria, see e.g. (Gan et al., 2007), the suitable number is chosen.

The basic term in cluster analysis is a *similarity*. An attempt to formalize the similarity measure and relation between similarity and distance is given in (Chen et al., 2009). Let $\mathbf{x}_i$ be a vector of variable values, which characterizes the *i*th object. If variables are quantitative then the distance between the *i*th and *j*th objects can be calculated e.g. as the Euclidean distance between vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ (in the following text an object and a representing vector will be considered as synonyms), i.e.

$$d(\mathbf{x}_i, \mathbf{x}_j) = d_{ij} = \sqrt{\sum_{i=1}^{m} (x_{ii}, x_{jl})^2} = ||\mathbf{x}_i - \mathbf{x}_j||, \tag{1}$$

[1] Department of Statistics and Probability, University of Economics, Prague, nám. W. Churchilla 4, 130 67 Prague, Czech Republic. E-mail: hana.rezankova@vse.cz, phone (+420)224095483.

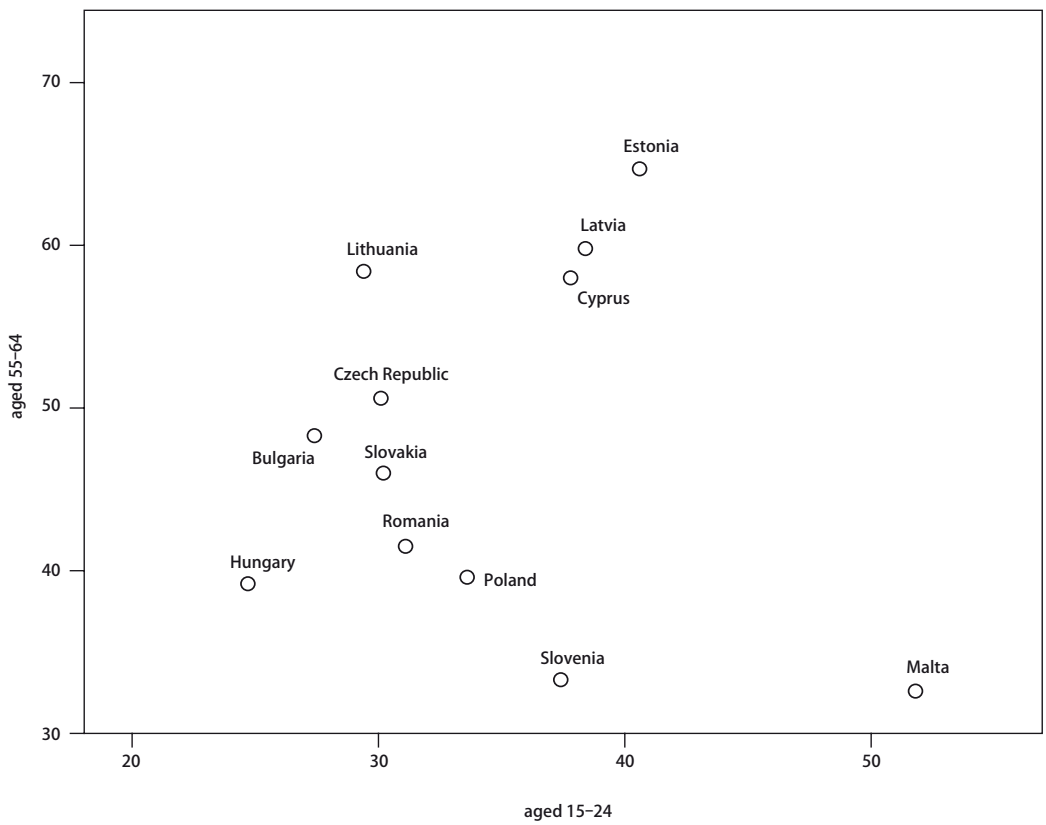
where m is the number of variables (e.g. economic indicators) and x_{il} is the value for the ith object and the lth variable. It is supposed that the data set $\mathbf{X}$ consisting of the vectors $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_n$, where n is the number of objects, should be partitioned into k clusters C_1, C_2…, C_k.

The main tasks for the cluster analysis research of the last decades has been clustering large data sets, clustering data files with categorical variables, fuzzy clustering and other techniques expressing uncertainty. Some related problems are solving, an outlier detection, determining the number of clusters, etc. Although the tasks mentioned above were solved at the beginning of the cluster analysis development, at the end of the 20th century and at the beginning of the 21st century the interest in these methods is growing in connection with the development of data mining techniques. The new algorithms for cluster analysis are proposed not only by statisticians, but also by computer science researchers. In this article the development of selected types of clustering is discussed.

1 HIERARCHICAL CLUSTER ANALYSIS

Probably the most applied method in economy is agglomerative hierarchical cluster analysis. It is based on a proximity matrix which includes the similarity evaluation for all pairs of objects. It means that various similarity or dissimilarity measures for different types of variables (quantitative, qualitative and binary) can be used. Moreover, different approaches for evaluation of the cluster similarity (single linkage, complete linkage, average linkage, Ward's method, etc.) can also be applied.

Figure 1 Scatter plot for countries characterized by economic activity rate in 2011 (IBM SPSS Statistics)



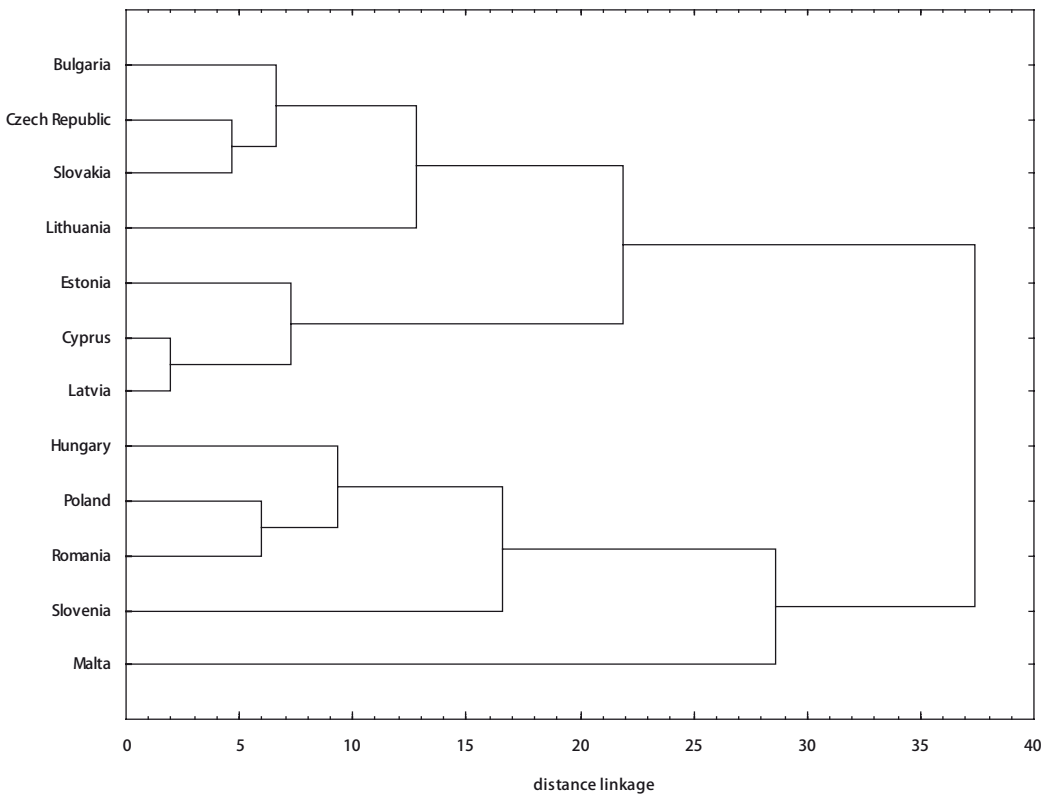
Source: *Slovensko v EÚ 2012 – Trh práce*. Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

Apart from giving a possibility to analyze data files with qualitative variables, the main advantage of this type of analysis is a graphical output in the form of a dendrogram. However, this graph is useful mainly for relatively small data files. In large files (with many objects) individual objects cannot be identified. Another disadvantage is a need to create a proximity matrix in the beginning of the analysis, what may cause a problem for large files.

Hierarchical cluster analysis can be illustrated by grouping selected countries of the European Union (new members of EU from 2004 and 2007 were selected). Let us consider three variables concerning the economic activity rate in 2011 according to the age (aged 15–24, 25–54, 55–64). Two of them can be represented by points in a scatter plot, see Figure 1.

With using the Euclidean distance and the complete linkage method the dendrogram in Figure 2 is obtained. We can see that Cyprus and Latvia are the most similar considering three studied variables (the countries are linked as the first; it is indicated by the smallest distance linkage in the dendrogram), then the Czech Republic and Slovakia are linked, etc. On the basis of the dendrogram we can identify two main clusters, which can be further divided to obtain a larger number of clusters.

Figure 2 Dendrogram for countries characterized by economic activity rate in 2011 (according to the age) obtained by the complete linkage method (STATISTICA)



Source: *Slovensko v EÚ 2012 – Trh práce.* Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

For example by cutting the dendrogram according to distance linkage 20 we obtain four clusters. In the first one there are the Czech Republic, Slovakia, Bulgaria and Lithuania, in the second one Cy-

prus, Latvia and Estonia are placed. The third cluster includes Poland, Romania, Hungary and Slovenia. In the fourth cluster there is only Malta. Minimum and maximum values of the analyzed variables characterizing four clusters are in Table 1.

Table 1 Characteristics of four clusters of countries obtained by the complete linkage method according to the economic activity rate in 2011

Cluster number	Aged 15–24		Aged 25–54		Aged 55–64	
	Min	Max	Min	Max	Min	Max
1	27.4	30.2	82.4	90.0	46.0	58.4
2	37.8	40.6	87.6	88.3	58.0	64.7
3	24.7	37.4	79.1	90.1	33.3	41.5
4	51.8	51.8	74.7	74.7	32.6	32.6

Source: Own calculation based on the data from publication: *Slovensko v EÚ 2012 – Trh práce*. Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

Malta has the highest value of the economic activity rate for the aged 15–24 group and the smallest value for the aged 55–64 group. The second cluster is characterized by high values both for the aged 15–24 group and for the aged 55–64 group. The first and the third clusters differ in the values of the aged 55–64 group.

If a data file contains nominal variables, some special measure must be used for similarity evaluation. The basic measure is the *simple matching coefficient*, which is also called the *overlap* measure. Let us denote the similarity of vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ as s_{ij}. For its calculation the values in the ith and jth rows of the input matrix are compared for all variables. Evaluation of relationships of the values for the lth variable is denoted as s_{lij}. If $x_{il} = x_{jl}$, then $s_{lij} = 1$, otherwise $s_{lij} = 0$. Similarity s_{ij} is calculated as the arithmetic mean, i.e.

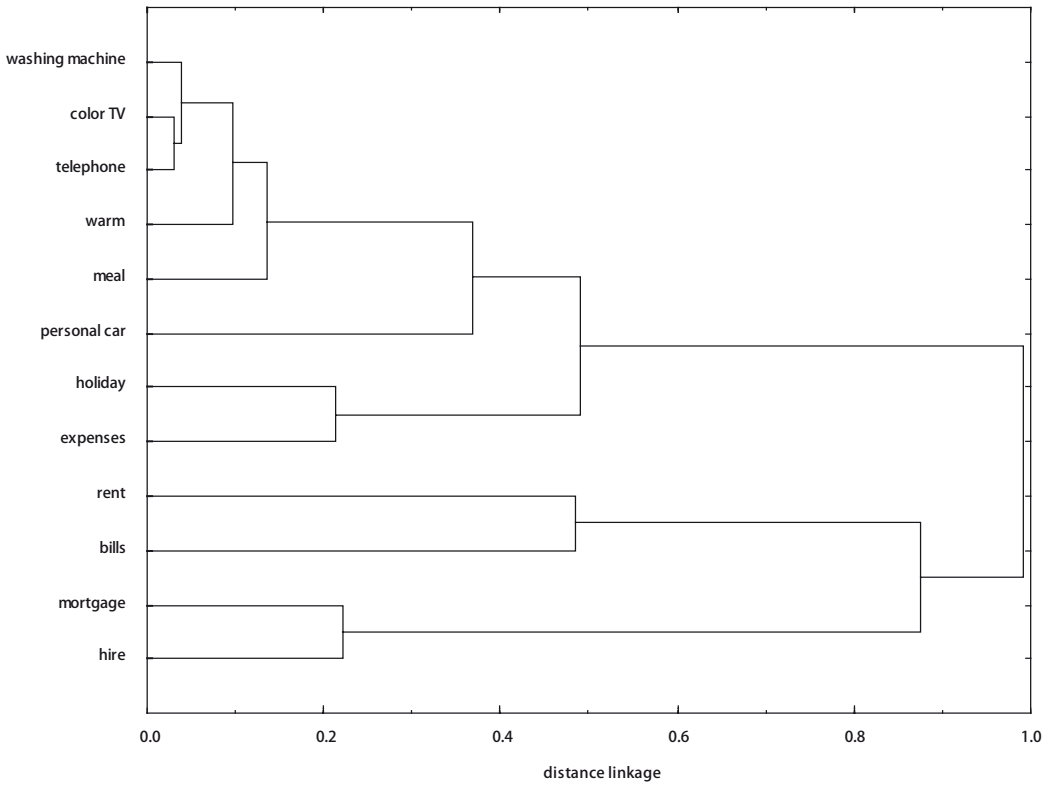
$$s_{ij} = \frac{\sum_{l=1}^{m} s_{lij}}{m}. \tag{2}$$

Hierarchical cluster analysis can be also based directly on a proximity matrix that evaluates the relationship of all pairs of variables. In a dendrogram similarity of variables and groups of variables can be identified.

Clustering of nominal variables will be illustrated by the data from the EU-SILC surveys in the Czech Republic (survey Living Condition 2011, the part "households"). There are nine indicators of material deprivation – eight indicators are answers to questions and the ninth one is composed of four answers. The data set with 12 original variables was analyzed (the number of households was 8 866). The variables contain values indicating whether or not the household can afford: a *washing machine*, a *color TV*, a *telephone*, a *personal car*, keeping the home adequately *warm*, a *meal* with meat, fish or vegetarian equivalent every second day, one week annual *holiday* away from home, coping with unexpected *expenses*, avoiding arrears in *rent*, utility *bills*, *mortgage* and *hire* purchase installments (the name of variables are in italic). These variables are nominal and they have different numbers of categories. The first four variables have three categories, the next four variables have two categories and the last four variables have three categories.

Figure 3 shows that the most similar answers concern a color TV and a telephone. The answers concern a washing machine are also alike (97–98% of the households own these durables). Three separated pairs of variables can be seen: *holiday* and *expenses* (56 and 58% of the households answered *yes*), *rent* and *bills*, and *mortgage* and *hire*.

Figure 3 Dendrogram for indicators of material deprivation of households, survey Living Condition 2011 (STATISTICA)



Source: the EU-SILC 2011 data

The *overlap* measure does not take into account different numbers of categories for individual variables. Recently, several similarity measures for objects characterized by nominal variables were proposed to deal with this problem. In the following text four measures for object similarity evaluation will be introduced. For the first three of them Equation (2) is applied, but the s_{lij} values are calculated differently.

The *Eskin measure* was proposed by Eskin et al. (2002). It assigns higher weights to mismatches which occur on variables with more categories. Let us denote the number of categories of the lth variable as n_l. If $x_{il} = x_{jl}$, then $s_{lij} = 1$, otherwise $s_{lij} = n_l^2/(n_l^2 + 2)$.

The *OF measure* (*Occurrence Frequency*) assigns higher weights to more frequent categories in case of mismatch, see (Boriah et al., 2008). Let us denote the frequency of the category (of the lth variable) equal to the value x_{il} as $f(x_{il})$. If $x_{il} = x_{jl}$, then $s_{lij} = 1$, otherwise $s_{lij} = 1/(1 + \ln(n/f(x_{il}))\cdot\ln(n/f(x_{jl})))$.

The *IOF measure* (*Inverse Occurrence Frequency*), see (Boriah et al., 2008), includes the opposite system of weights to OF. It evaluates mismatches of more frequent categories by lower weights. If $x_{il} = x_{jl}$, then $s_{lij} = 1$, otherwise $s_{lij} = 1/(1 + \ln f(x_{il})\cdot\ln f(x_{jl}))$.

The *Lin measure* (Lin, 1998) assigns higher weights to more frequent categories in case of matches and lower weights to infrequent categories in case of mismatches. Let us denote a relative frequency

of the category equal to the value x_{il} as $p(x_{il})$. If $x_{il} = x_{jl}$, then $s_{lij} = 2 \cdot \ln p(x_{il})$, otherwise $s_{lij} = 2 \cdot \ln(p(x_{il}) + p(x_{jl}))$. The similarity measure for two objects is then computed as

$$s_{ij} = \frac{\sum_{l=1}^{m} s_{lij}}{\sum_{l=1}^{m} (\ln p(x_{il}) + \ln p(x_{jl}))}. \tag{3}$$

The measures mentioned above and some other measures have been reviewed e.g. in (Boriah et al., 2008) and (Chandola et al., 2009). Some other similarity measures have been proposed, e.g. in (Le et Ho, 2005) and (Morlini et Zani, 2012).

2 K-CLUSTERING

In *k*-clustering the set of objects is divided to a certain number (*k*) clusters. We can distinguish different approaches from different points of view. The first classification is for hard and fuzzy clustering. In the first one, an object is assigned exactly to one cluster. The result is a membership matrix for objects and clusters with ones (the object is assigned to the cluster) and zeroes (the object is not assigned to the cluster). In the second approach membership degrees are calculated for all cluster-object pairs. Moreover, some other approaches to expressing uncertainty in cluster analysis have been proposed, see below.

The second classification is for *k*-centroid and *k*-medoids clustering. In the former, the center of a cluster is represented by a vector of variable characteristics (e.g. vector of means for quantitative variables). In the latter, the center of a cluster is represented by a selected object (by a vector from the input matrix).

The most applied *k*-centroid technique is the *k*-means (also HCM – *hard c-means*) algorithm (MacQueen, 1967), which analyzes the data set with the aim to minimize the objective function

$$J_{HCM} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{hard,\,ih} d_{ih}^2, \tag{4}$$

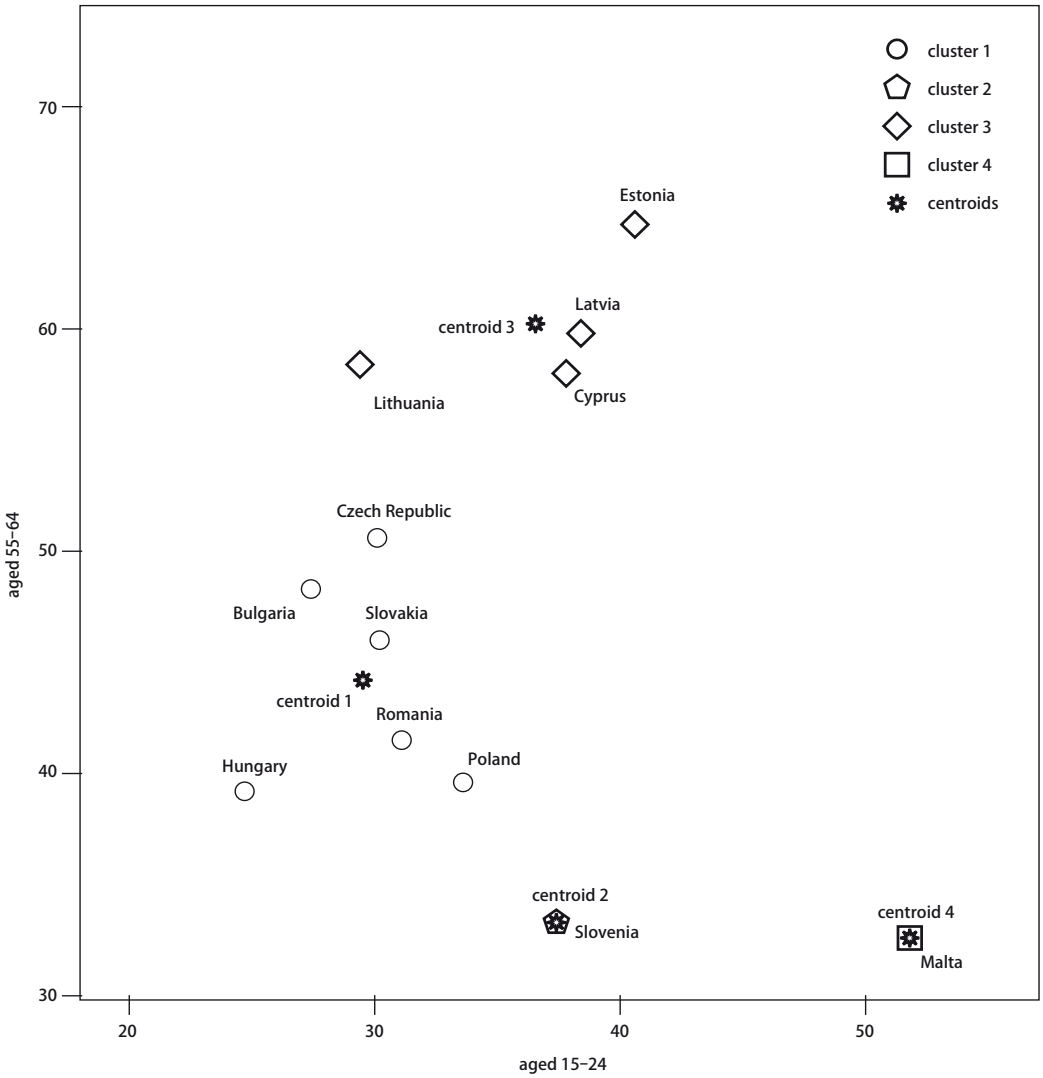
where the elements $u_{hard,ih} \in \{0, 1\}$ indicate the assignment of object vectors to clusters (1 means the assignment) and d_{ih} is the Euclidean distance between the *j*th object and the center (a vector of means) of the *h*th cluster. Further, the following conditions have to be satisfied:

$$\sum_{h=1}^{k} u_{hard,\,ih} = 1 \text{ for } i = 1, 2, ..., n \text{ and } \sum_{i=1}^{n} u_{hard,\,ih} > 0 \text{ for } h = 1, 2, ..., k.$$

Let us analyze the data file with three variables concerning the economic activity rate in 2011 according to age (see Section 1, Figure 1). With using the k-means algorithm for clustering countries to four clusters we obtain two one-element clusters (Slovenia and Malta). All four clusters and their centroids are presented in Figure 4 and the obtained clusters are characterized in Table 2. They differ from results obtained by the complete linkage method (Figure 2) but they are consistent with the results obtained by the average linkage method (these results are not presented in this paper).

According to the studied variables Malta is significantly different from the other countries regardless of the method used. Slovenia is characterized by the low value of the economic activity rate for the age group 55–64 and the highest value for the age group 25–54. The first and the third clusters differ mainly in the values of the age group 55–64.

Figure 4 Scatter plot for countries characterized by economic activity rate in 2011 with centroids of four clusters obtained by *k*-means clustering (IBM SPSS Statistics)



Source: *Slovensko v EÚ 2012 – Trh práce.* Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

The advantage of *k*-centroid clustering is a possibility to apply it to large data sets. The disadvantage is its instability; for different orders of object vectors different assignments of objects to clusters can be obtained. Further, the result of clustering depends on a type of initialization (determination of *k* initial centroids), which is the first step of the algorithm. *K*-clustering methods search for the optimal solution, but the optimum can only be local, not global. Despite some negative properties these methods play an important role in the exploratory data analysis.

Table 2 Characteristics of four clusters of countries obtained by k-means clustering according to the economic activity rate in 2011

Cluster number	Aged 15–24		Aged 25–54		Aged 55–64	
	Min	Max	Min	Max	Min	Max
1	24.7	33.6	79.1	88.0	39.2	50.6
2	37.4	37.4	90.1	90.1	33.3	33.3
3	29.4	40.6	87.6	90.0	58.0	64.7
4	51.8	51.8	74.7	74.7	32.6	32.6

Source: Own calculation based on the data from publication: *Slovensko v EÚ 2012 – Trh práce.* Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

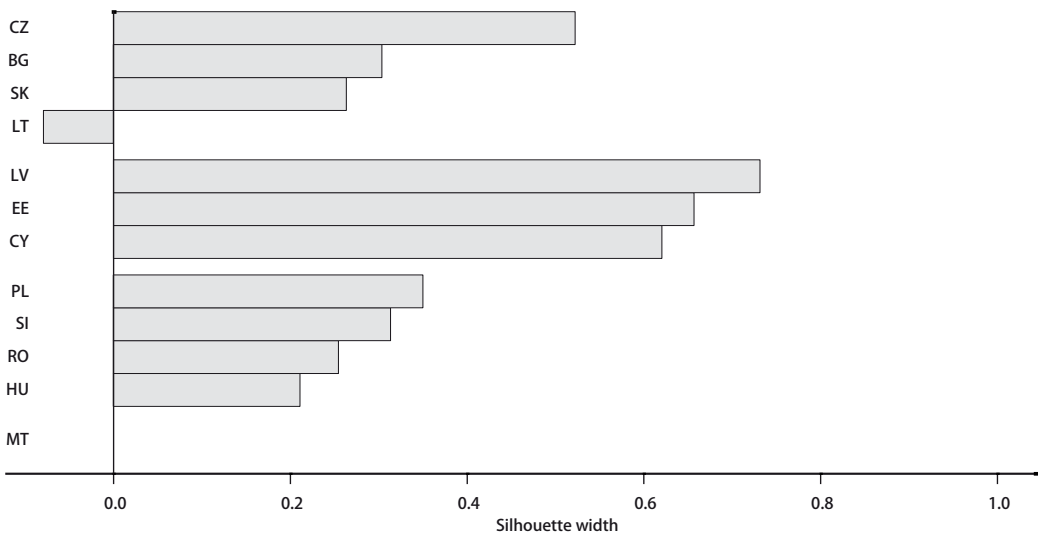
In the hard *k-medoids* (also *PAM – Partitioning Around Medoids*) algorithm, see (Kaufman et Rousseeuw, 2005), the objective function

$$f_{KM} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{hard,\,ih} ||\mathbf{x}_i - \mathbf{m}_h|| \tag{5}$$

is minimized, where $\mathbf{m}_h$ is a medoid of the hth cluster and for values $u_{hard,ih}$ the same conditions as in case of k-means clustering must be satisfied.

With using the k-medoids algorithm for clustering countries to four clusters we obtain one one-element cluster (Malta). All four clusters are presented in Figure 5 in the form of a silhouette plot. The silhouette widths are computed on the basis of distances of an object from the other objects from both the same cluster and the other clusters. The first object in the cluster is a medoid. In Figure 5 the medoids are the Czech Republic, Latvia, Poland and Malta. An opposite direction (a negative value) in case of Lithuania (belonged to the first cluster) means that this country is closer the objects from other clusters (according to the special measure).

Figure 5 Silhouette plot for countries characterized by economic activity rate in 2011 for four clusters obtained by the PAM algorithm (S-PLUS)



Source: *Slovensko v EÚ 2012 – Trh práce.* Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

From Figure 5 it is obvious that generally some objects can be assigned to two (or more) clusters. In this case an uncertainty can be expressed in results of clustering. One of the approaches how to express an uncertainty in cluster analysis is a fuzzy assignment of objects to clusters. It is applied in fuzzy cluster analysis. This technique is based on the theory of fuzzy sets (Zadeh, 1965). Fuzzy clustering has been studied very intensively in the past decades. A lot of papers have been published in journals, conference proceedings and in some monographs, e.g. (Abonyi et Feil, 2007) and (Höppner et al., 2000). There are many different algorithms used for fuzzy (soft) cluster analysis. Fuzzy *k*-means is one of them, see e.g. (Kruse et al., 2007). It is based on a generalization of the *k-means* (HCM) algorithm.

The *fuzzy k-means* (frequently FCM – *fuzzy c-means*) algorithm (Bezdek, 1981) minimizes the objective function

$$J_{FCM} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{ih}^{q} d_{ih}^{2}, \tag{6}$$

where the elements $u_{ih} \in \langle 0, 1 \rangle$ are membership degrees, and the parameter q ($q > 1$) is called a fuzzifier or weighting exponent (usually $q = 2$ is chosen). Furthermore, the following conditions have to be satisfied:

$$\sum_{h=1}^{k} u_{ih} = 1 \text{ for } i = 1, 2, ..., n \text{ and } \sum_{i=1}^{n} u_{ih} > 0 \text{ for } h = 1, 2, ..., k.$$

We can again illustrate the application of fuzzy cluster analysis to the data on selected countries of the European Union (see Section 1, Figure 1). Using the FANNY algorithm in the S-PLUS statistical software, see (Kaufman et Rousseeuw, 2005), we obtain the results in Table 3 with the assignment of countries to four clusters.

Table 3 Country membership degrees based on economic activity rate in 2011 for four clusters obtained by the FANNY algorithm (S-PLUS)

Country	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster number
Bulgaria	**0.51**	0.11	0.28	0.10	1
Czech Republic	**0.71**	0.10	0.12	0.07	1
Estonia	0.15	**0.66**	0.11	0.09	2
Cyprus	0.09	**0.82**	0.05	0.04	2
Lithuania	0.35	**0.36**	0.17	0.12	2
Latvia	0.05	**0.89**	0.03	0.02	2
Hungary	0.22	0.09	**0.53**	0.15	3
Malta	0.21	0.18	0.28	**0.34**	4
Poland	0.21	0.08	**0.48**	0.23	3
Romania	0.12	0.05	**0.74**	0.08	3
Slovakia	**0.62**	0.08	0.20	0.09	1
Slovenia	0.03	0.02	0.04	**0.90**	4

Source: Own calculation based on the data from publication: *Slovensko v EÚ 2012 – Trh práce*. Štatistický úrad Slovenskej republiky. ISBN 978-80-8121-123-2

According to the highest value of membership degrees over clusters (bold figures in Table 3), the first cluster is created by the Czech Republic, Slovakia and Bulgaria. In the second cluster there are Latvia, Cyprus, Estonia and Lithuania. It can be noticed Lithuania has similar membership degrees to the first and the second clusters (0.35 and 0.36). The third cluster is created by Romania, Hungary and Poland, and the fourth cluster contains Slovenia and Malta. However, membership degrees are very various – it is 0.9 for Slovenia and 0.34 for Malta.

The fuzzy *k*-means algorithm is sensitive to noise and outliers. Let us suppose clustering to two clusters C_h and C_g. If $\mathbf{x}_i$ is equidistant from centroids c_h and c_g then $u_{ih} = u_{ig} = 0.5$, regardless whether the actual distance is large or small. A similar situation can be mentioned in the *fuzzy k-medoids* algorithm (Krishnapuram et al, 2001), in which the objective function

$$f_{FKM} = \sum_{h=1}^{k} \sum_{i=1}^{n} u_{ih}^{q} \left\| \mathbf{x}_i - \mathbf{m}_h \right\|^2 \tag{7}$$

is minimized under the same condition as in the fuzzy *k*-means algorithm.

For the reason of the negative property of fuzzy clustering, different approaches were proposed later, see (Bodjanova, 2013). One of them is the *possibilistic k-means* (also PCM – *possibilistic c-means*) algorithm (Krishnapuram et Keller, 1993). It minimizes the objective function

$$J_{PCM} = \sum_{h=1}^{k} \sum_{i=1}^{n} w_{ih}^{q} d_{ih}^{2} + \sum_{h=1}^{k} \gamma_h \sum_{i=1}^{n} (1 - w_{ih})^q , \tag{8}$$

where the elements $w_{ih} \in \langle 0, 1 \rangle$ are membership degrees, q is a fuzzifier, and the following conditions have to be satisfied

$\sum_{h=1}^{k} u_{ih} = 1$ for $i = 1, 2, ..., n$ and $\sum_{i=1}^{n} u_{ih} > 0$ for $h = 1, 2, ..., k$. and γ_h is a user defined constant (scale parameter). It can be computed e.g. as

$$\gamma_h = \frac{\sum_{i=1}^{m} u_{ih}^{q} d_{ih}^{2}}{\sum_{i=1}^{m} u_{ih}^{q}} .$$

This algorithm is very sensitive to initialization.

Since both the FCM and the PCM algorithms have some negative properties, the combination of both algorithms has been developped in results of the *FPCM algorithm* (Pal et al., 2005). It minimizes the objective function

$$J_{FPCM} = \sum_{h=1}^{k} \sum_{i=1}^{n} (a u_{ih}^{q_1} + b w_{ih})^{q_2} + \sum_{h=1}^{k} \gamma_h \sum_{i=1}^{n} (1 - w_{ih})^{q_2} , \tag{9}$$

where a, b, q_1, q_2 and γ_h are positive constants. Constants a and b define the relative importance of probabilistic and possibilistic memberships. The *fuzzy-possibilistic c-medoids* algorithm has been also proposed (Maji et Pal, 2007a).

Other approaches which are alternative to hard clustering are techniques based on the *rough set theory* (Pawlak, 1982). The basic technique is the *rough k-means* (or RCM – *rough c-means*) algorithm (Lingras et West, 2004). In this algorithm each cluster C_h is defined by the lower approximation $A_{low}(C_h)$ and the upper approximation $A_{up}(C_h)$. The object $\mathbf{x}_i$ can be a part of most one lower approximation.

If the object $\mathbf{x}_i$ is a part of a certain lower approximation then is also a part of the upper approximation. If $\mathbf{x}_i$ is not a part of any lower approximation then it belongs to two or more upper approximation. A special technique for the cluster mean computation is applied.

In the RCM algorithm two values characterizing the membership for a certain object and a certain cluster are calculated. There are the low membership $u_{low,ih}$ and the up membership $u_{up,ih}$. If $u_{low,ih} = 1$ then the ith object certainly belongs to the hth cluster. If $u_{low,ih} = 0$ then the assignment of the ith object depends on the value of $u_{up,ih}$. If $u_{up,ih} = 1$ then the ith object possibly belongs to the hth cluster. If $u_{up,ih} = 0$ then the ith object does not belong to the hth cluster. The following conditions have to be satisfied:

$$u_{low,ih,}\, u_{up,\,ih} \in \{0, 1\}\, u_{low,ih} \leq u_{up,ih,} \text{ and } \sum_{h=1}^{k} u_{low,ih} \leq 1 \text{ for } i = 1, 2, ..., n.$$

The result of a combination of rough and fuzzy approaches is the *rough-fuzzy k-means* (or RFCM – *rough-fuzzy c-means*) algorithm (Mitra et al., 2004). Moreover, Maji and Pal (2007b) proposed the *rough-fuzzy-possibilistic k-means* (or RFPCM – *rough-fuzzy-possibilistic c-means*).

In rough-based techniques the lower approximation of each cluster depends on a fixed threshold *TH* which is defined by the user. For this reason a modification of these approaches based on the *shadowed sets* (Pedrycz, 1998) was proposed by Mitra et al. (2010). This algorithm is called *shadowed k-means* (SCM – *shadowed c-means*). It provides the dynamical evaluation of thresholds for each cluster individually, based on the original data.

3 OTHER APPROACHES

Besides of hierarchical clustering and *k*-clustering, there are some other approaches proposed e.g. for large data files (with many objects), for data files with categorical variables, and also for data files of both types. We can mention *two-step cluster analysis* implemented in the IBM SPSS Statistics software as an example of the procedure which can cluster large data sets with both quantitative and qualitative variables. This method is based on the BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*) algorithm, see (Zhang et al., 1996).

The algorithm arranges objects of the data set into subclusters, known as cluster features (CFs). These cluster features are then clustered into *k* groups using a traditional hierarchical clustering procedure. A CF represents a set of summary statistics on a subset of the data. The algorithm consists of two phases. In the first one, an initial CF tree is built. In the second one, an arbitrary clustering algorithm is used to cluster the leaf nodes of the CF tree. The disadvantage of this method is its sensitivity to the order of the objects.

In two-step cluster analysis, the user can apply either the Euclidean distance for the quantitative data or the log-likelihood distance measure which is determined for the data files with the combination of quantitative and qualitative variables (Chiu et al., 2001). In the second case the dissimilarity of two clusters is expressed as the difference between a variability of the cluster created by linking of the studied clusters and a sum of the variability of individual clusters. A variability is calculated as a combination of values of the variance (for quantitative variables) and the entropy (for qualitative variables).

The application of this method will be illustrated to the EU-SILC data (Living Condition 2011). After clustering 8 866 households (i.e. large data set for cluster analysis) according to 12 nominal variables analyzed in Section 1, the procedure determines two clusters of households as optimal (the average silhouette width is calculated on the basis of the silhouette widths, see Figure 5).

The output for two clusters indicates that the most important variables for clustering are *personal car, holiday,* and *expenses*. For three clusters, variables *mortgage* and *hire* were added as important. Variables *warm* and *meal* were added as important for four clusters. The relative frequencies of categories for variables mentioned above are in Tables 4 and 5.

Table 4 Relative frequencies of answers in individual clusters obtained by two-step cluster analysis based on indicators of material deprivation, survey Living Condition 2011

Cluster number (size)	Warm		Meal		Holiday		Expenses	
	Yes	No	Yes	No	Yes	No	Yes	No
1/2 (49.9%)	85.9%	14.1%	76.4%	23.6%	29.7%	70.3%	30.9%	69.1%
2/2 (50.1%)	99.8%	0.2%	99.8%	0.2%	83.0%	17.0%	84.2%	15.8%
1/3 (54.8%)	87.5%	12.5%	79.2%	20.8%	32.2%	67.8%	36.1%	63.9%
2/3 (20.1%)	98.8%	1.2%	97.8%	2.2%	68.1%	31.9%	63.6%	36.4%
3/3 (25.1%)	100.0%		100.0%		100.0%		100.0%	
1/4 (37.8%)	99.8%	0.2%	99.6%	0.4%	39.9%	60.1%	43.5%	56.5%
2/4 (19.6%)	99.4%	0.6%	99.5%	0.5%	69.6%	30.4%	54.7%	35.3%
3/4 (17.5%)	60.5%	39.5%	33.7%	66.3%	14.9%	85.1%	19.7%	80.3%
4/4 (25.1%)	100.0%		100.0%		100.0%		100.0%	

Source: the EU-SILC 2011 data

Table 5 Relative frequencies of answers in individual clusters obtained by two-step cluster analysis based on indicators of material deprivation, survey Living Condition 2011

Cluster number	Personal car			Mortgage			Hire		
	Own	Cannot afford	Other	Yes	No	Other	Yes	No	Other
1/2	29.8%	22.5%	47.7%	1.0%	4.4%	94.6%	2.4%	10.8%	86.9%
2/2	98.0%	0.5%	1.5%	0.1%	19.9%	80.0%	0.1%	19.3%	80.7%
1/3	38.1%	19.4%	42.5%	0.8%	1.9%	97.2%	2.1%	5.7%	92.3%
2/3	89.7%	4.2%	6.1%	0.6%	55.4%	44.1%	0.4%	59.4%	40.2%
3/3	100.0%					100.0%			100.0%
1/4	37.5%	14.6%	47.9%		0.1%	99.9%		1.1%	98.9%
2/4	89.1%	3.7%	7.2%	0.6%	54.6%	44.8%	0.3%	58.7%	41.0%
3/4	41.3%	29.9%	28.8%	2.6%	7.9%	89.5%	6.5%	17.7%	75.8%
4/4	100.0%					100.0%			100.0%

Source: the EU-SILC 2011 data

If the households are clustered to two clusters, they can be characterized in the following way. One cluster includes mostly the households that own a personal car and have no problems neither with paying holiday nor with unexpected expenses. The second cluster represents the households which have not a personal car from other reason than financial and have problems to pay holiday and unexpected expenses. Similarly the results of clustering to three and four clusters can be described.

Another application of two-step cluster analysis to the EU-SILC data is described in (Řezanková et Löster, 2013). For the analysis of large data files with quantitative variables, methods *k*-clustering can

be applied, either classical algorithms or their modifications. We can mention the *CLARA* (*Clustering LARge Applications*) algorithm as an example (Kaufman et Rousseeuw, 2005). It is based on the *k*-medoid algorithm and implemented in the S-PLUS system.

The principles of methods proposed for large data files (both with many objects and many variables) are reviewed e.g. in (Kogan, 2007). An example of techniques for clustering in case of high-dimensional data is the R package BCLUST (Partovi Nia et Davison, 2012). The approaches to clustering categorical data are summarized e.g. in (Řezanková, 2009). If a data set contains mixed-type variables, one possibility is to cluster objects according groups of variables of the same type and then combine of individual solutions by cluster ensembles, e.g. by package CLUE for R, see (Hornik, 2005).

CONCLUSION

In the paper selected approaches to cluster analysis were introduced. For cluster analysis of objects which are characterized by values of nominal variables, the analyst can use recently proposed similarity measures. Performed experiments showed (Šulc et al., 2013) that clustering with using some of these measures give better clusters than the overlap measure from the point of view of the within-cluster variability.

Recent methods which include uncertainty are a promising tool to give better results than basic fuzzy cluster analysis. For the data from some surveys, e.g. the EU-SILC, the techniques for large data sets is useful.

It is regrettable that commercial software products react to the recently proposed methods very slowly, or do not react at all. They rarely include measures for nominal variables, fuzzy cluster analysis and methods for large data files. If the software system offers some of these possibilities, it is usually just one of them and the analysts need to use several software products to perform modern analyses.

References

ABONYI, J., FEIL, B. *Cluster Analysis for Data Mining and System Identification*. Berlin: Birkhäuser Verlag AG, 2007.

BEZDEK, J. C. *Pattern Recognition with Fuzzy Objective Function Algorithm*. New York: Plenum Press, 1981.

BODJANOVA, S.: Fuzzy sets and rough sets in prototype-based clustering algorithms. In *Olomoucian Days of Applied Mathematics 2013 – Presentations* [online]. Olomouc: Palacky University in Olomouc, 2013. [cit. 11.11.2013]. <http://odam.upol.cz/downloads/presentations/2013/Bodjanova.pdf>.

BORIAH, S., CHANDOLA, V., KUMAR, V. Similarity measures for categorical data: A comparative evaluation. In *Proceedings of the 8th SIAM International Conference on Data Mining.* SIAM, 2008, pp. 243–254.

CHANDOLA, V., BORIAH, S., KUMAR, V. A framework for exploring categorical data. In *Proceedings of the 9th SIAM International Conference on Data Mining.* SIAM, 2009, pp. 187–198.

CHEN, S., MA, B., ZHANG, K. On the similarity metric and the distance metric. In Formal Languages and Applications: A Collection of Papers in Honor of Sheng Yu. *Theoretical Computer Science*, 2009, 24–25, pp. 2365–2376.

CHIU, T., FANG, D., CHEN, J., WANG, Y., JERIS, C. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* New York: ACM, 2001, pp. 263–268.

ESKIN, E., ARNOLD, A., PRERAU, M., PORTNOY, L., STOLFO, S. A geometric framework for unsupervised anomaly detection. In BARBARA, D., JAJODIA, S., eds. *Applications of Data Mining in Computer Security*, pp. 78–100. Norwell, MA: Kluwer Academic Publishers, 2002.

GAN, G., MA, C., WU, J. *Data Clustering: Theory, Algorithms, and Applications.* Philadelphia: ASA-SIAM, 2007.

HÖPPNER, F., KLAWON, F., KRUSE, R., RUNKLER, T. *Fuzzy Cluster Analysis. Methods for Classification, Data Analysis and Image Recognition.* New York: John Wiley & Sons, 2000.

HORNIK, K. A CLUE for CLUster Ensembles. *Journal of Statistical Software,* 2005, 14(12), pp. 1–25.

KAUFMAN, L., ROUSSEEUW, P. *Finding Groups in Data: An Introduction to Cluster Analysis.* Hoboken: Wiley, 2005.

KOGAN, J. *Introduction to Clustering Large and High-Dimensional Data.* New York: Cambridge University Press, 2007.

KRISHNAPURAM, R., KELLER, J. M. A possibilistic approach to clustering. *IEEE Trans. Fuzzy Syst.*, 1993, 1(2), pp. 98–110.

KRISHNAPURAM, R., JOSHI, A., YI, L. Fuzzy relative of the k-medoids algorithm with application to web document and snippet clustering. In *IEEE International Conference on Fuzzy Systems 3*. Institute of Electrical and Electronics Engineers Inc., 1999, pp. III-595–III-607.

KRUSE, R., DÖRING, C., LESOT, M.-J. Fundamentals of Fuzzy Clustering. In OLIVEIRA, J. V., PEDRYCZ, W., eds. *Advances in Fuzzy Clustering and Its Applications*. Chichester: John Wiley & Sons, 2007, pp. 3–30.

LE, S. Q., HO, T. B. An association-based dissimilarity measure for categorical data. *Pattern Recognition Letters*, 2005, 26(16), pp. 2549–2557.

LIN, D. An information-theoretic definition of similarity. In *ICML '98: Proceedings of the 15th International Conference on Machine Learning*. San Francisco: Morgan Kaufmann Publishers Inc., 1998, pp. 296–304.

LINGRAS, P., WEST, C. Interval set clustering of web users with rough k-means. *Journal of Intelligent Information Systems*, 2004, 23, pp. 5–16.

MACQUEEN, J. B. Some methods for classification and Analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.* University of California Press, 1967, pp. 281–297.

MAJI, P., PAL, S. K. Protein sequence analysis using relational soft clustering algorithms. *International Journal of Computer Mathematics*, 2007a, 84(2), pp. 599–617.

MAJI, P., PAL, S. K. Rough set based generalized fuzzy c-means algorithm and quantitative indices. *IEEE Trans. Syst., Man and Cybernetics Part B*, 2007b, 37(6), pp. 1529–1540.

MITRA, S., BANKA, H., PEDRYCZ, W. Rough-fuzzy collaborative clustering. *IEEE Trans. Syst., Man and Cybernetics, Part B*, 2006, 36(4), pp. 795–805.

MITRA, S., PEDRYCZ, W., BARMAN, B. Shadowed c-means: Integrating fuzzy and rough clustering. *Pattern Recognition*, 2010, 43, pp. 1282–1291.

MORLINI, I., ZANI, S. A new class of weighted similarity indices using polytomous variables. *Journal of Classification,* 2012, 29(2), pp. 199–226.

PAL, N. R., PAL, K., KELLER, J. M., BEZDEK, J. C. A possibilistic fuzzy c-means clustering algorithm. *IEEE Trans. Fuzzy Syst.*, 2005, 13 (4), pp. 517–530.

PARTOVINIA, V., DAVISON, A. High-dimensional Bayesian clustering with variable selection: the R package bclust. *Journal of Statistical Software*, 2012, 47(5), pp. 1–22.

PAWLAK, Z. Rough sets. *International Journal of Computer and Information Sciences*, 1982, 11, pp. 341–356.

PEDRYCZ, W. Shadowed sets: representing and processing fuzzy sets. *IEEE Trans. Syst., Man and Cybernetics,* Part B, 1998, 28(1), pp. 103–109.

ŘEZANKOVÁ, H. Cluster analysis and categorical data. *Statistika*, 2009, 3, pp. 216–232.

ŘEZANKOVÁ, H., LÖSTER, T. Shluková analýza domácností charakterizovaných kategoriálními ukazateli. *E+M Ekonomie a Management*, 2013, 3, pp. 139–147.

ŠULC, Z., ŘEZANKOVÁ, H., MOHAMMAD, A. Comparison of selected approaches to categorical data clustering. In *AMSE 2013*. Banská Bystrica: Univerzita Mateja Bela, 2013, p. 25.

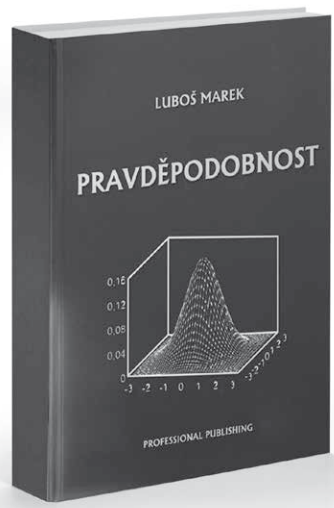
ZADEH, L. A. Fuzzy sets. *Information and Control*, 1965, 8, pp. 338–353.

ZHANG, T., RAMAKRISHNAN, R., LIVNY. M. BIRCH. An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD Conference on Management of Data*. Montreal: ACM, 1996, pp. 103–114.

PRAVDĚPODOBNOST (PROBABILITY): CRITICAL REVIEW

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

MAREK, L. *Pravděpodobnost*. Prague: Professional Publishing, 2012. ISBN 978-80-7431-087-4.



The book deals with the probability theory up to limit theorem and enables the reader to learn the subject in Czech language. It offers a complete, carefully structured and continuous explanation which enables the reader to follow the process of building the above mechanism from the basis up to complex problems of probability convergences in limit theorems. The book can be easily studied by anyone who mastered basics of university mathematics and has especially active knowledge of matrix algebra and differential and integral calculus. The selected access can make the study of theoretically difficult methods easier to those who are lacking more profound mathematical basis.

The text is structured into seven parts which clearly and consistently divide the subjects into consecutive units and the explanation then proceeds from the probability definition up to limit theorems. In the first part the concept of random trial is discussed in detail and possible definitions of probabilities are given. The author deals with properties of probability, examines carefully the independence of events and related definition of conditional probability. The following two chapters are devoted to random variables and random vectors, their distribution and numerical characteristics. These problems are followed by examining transformations of random variables (and vectors) which are very useful and are covered in chapter four. In the following two chapters the readers may find a review of most commonly used discrete and continuous distributions of random variables and two multivariate distributions (multinominal distribution and multivariate normal distribution). The review of random variable contains also sample distributions used for inductive inference in mathematical statistics. The last chapter regarding limit theorems introduced convergences in probability and in distribution and deals with the formulation and use of selected laws of large numbers and central limit theorems.

[1] Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: malai@vse.cz.

The scope of information in the book will enable the reader to use with comprehension the probability models or to follow statistical literature. The monograph can be recommended both to those who just begin with the study of the subject and to the users of statistical methods wishing to complete their fundamental knowledge on which these methods are based. The book will help such reader to study the literature and to use the methods of mathematical statistics with better comprehension. The book may also serve as valuable study aid for students who study the subject at the university in bachelor's or master's programmes. It seems to me that the book can be successfully used as a reference source of information which the users of probability and mathematical statistics need constantly.

The book offers every knowledge and information necessary for mastering the subject, unfortunately, the intention which the author wished to follow suggests that the reader does not have a chance to test his/her knowledge and comprehension of explanation on examples or problems for which they may seek solution independently. The reader of this type of monograph would definitely appreciate more detailed bibliography which would make further study of this extensive subject easier.

Taking into account the above we may conclude that the monograph can be recommended to anyone who mastered at least fundamental knowledge of university mathematics and wishes to properly learn classical theory of probability.

Recent Publications and Events

New Publications of the Czech Statistical Office

Cizinci v České Republice 2013 (Foreigners in the Czech Republic 2013). Prague: CZSO, 2013.
Česká republika v mezinárodním srovnání 2013 (Czech Republic in International Comparison 2013). Prague: CZSO, 2014.
DUBSKÁ, D., KAMENICKÝ, J., KUČERA, L. *Vývoj ekonomiky České republiky v 1. až 3. čtvrtletí 2013* (Development of the Czech Economy in the first three Quarters of 2013). Prague: CZSO, 2013.
KAMENICKÝ, J. *Vybrané aspekty vývoje hospodaření vládního sektoru v zemích EU* (Selected Aspects of the Development of the Government Sector in the EU). Prague: CZSO, 2013.
Náboženská víra obyvatel podle výsledků sčítání lidu 2011 (Religious Belief of the Population According to Results of the Census 2011). Prague: CZSO, 2014.
Sčítání lidu, domů a bytů 2011 (Population and Housing Census 2011). Prague: CZSO, 2013.
Spotřeba potravin 1948–2012 (Food Consumption 1948–2012). Prague: CZSO, 2013.

Other Selected Publications

BARTOŠOVÁ, J. *Finanční potenciál domácností. Kvantitativní metody a analýzy* (Financial Potential of Households. Quantitative Methods and Analyses). Prague: Professional Publishing, 2013.
HEŘMANOVÁ, E. *Koncepty, teorie a měření kvality života* (Concepts, Theories and Measurement of Life Quality). Prague: SLON, 2012.
MELOUN, M., MILITKÝ, J., HILL, M. *Statistická analýza vícerozměrných dat v příkladech* (Statistical Analysis of Multidimensional Data in the Examples). Prague: Academia, 2012.
NEUBAUER, J., SEDLÁČEK, M., KŘÍŽ, O. *Základy statistiky* (Introduction to Statistics). Prague: Grada, 2012.
VENKATESH, S. S. *The Theory of Probability.* Cambridge University Press, 2013.

Conferences

The ***21st International Conference on Computational Statistics COMPSTAT 2014*** will be held ***from 19th to 22nd August 2014*** at the International Conference Centre in ***Geneva, Switzerland***. The conference aims at bringing together researchers and practitioners to discuss recent developments in computational methods, methodologies for data analysis and applications in statistics. More information available at: *http://compstat2014.org*.

The ***17th International Scientific Conference AMSE 2014*** (Applications of Mathematics and Statistics in Economics) will take place ***from 27th to 31st of August 2014*** in ***Jerzmanowice, Poland***. The purpose of the conference is to acquaint the participants of the conference with the latest mathematical and statistical methods that can be used in solving theoretical and practical economic problems. The conference main sections are: Macroeconomics, Public Economics and Methodological Issues of Economics; Social Economics, Economic Sustainability and Demographic Economics; Financial

Markets, Risk Measurement and Insurance; Microeconomic Issues; Multidimensional Statistics in Economics. AMSE 2014 is organized by the University of Economics, Prague, Czech Republic (Faculty of Informatics and Statistics, Department of Statistics and Probability), Matej Bel University, Banská Bystrica, Slovakia (Faculty of Economics, Department of Quantitative Methods and Information Technology) and Wroclaw University of Economics, Wroclaw, Poland (Department of Statistics). More information available at: *www.amse.ue.wroc.pl*.

The ***60th World Statistics Congress ISI 2015*** will be held during ***26–31 July 2015*** in ***Rio de Janeiro, Brazil***. The congress will bring together members of the statistical community to present, discuss, promote and disseminate research and best practice in every field of Statistics and its applications. More information available at: *http://www.isi2015.ibge.gov.br*.

Papers
We publish articles focused at theoretical and applied statistics, mathematics and statistical methods, econometrics, applied economics, 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 *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 field of the official statistics. Reviews shall have up to 600 words or one (1) 1.5-spaced page.

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 (except headings).

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: Chráněná grafická dílna Slunečnice, David Hošek
Print: Czech Statistical Office

94th 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