STATISTIKA

STATISTICS AND ECONOMY JOURNAL

VOL. 95 (4) 2015

CZECH
STATISTICAL
OFFICE

EDITOR-IN-CHIEF

EDITORIAL BOARD

EXECUTIVE BOARD

MANAGING EDITOR

CONTENTS

ANALYSES

METHODOLOGY

INFORMATION

About Statistika

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

Publisher

The Czech Statistical Office is an official national statistical institution of the Czech Republic. The Office 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

Income Inequality by Highest Attained Education in the Czech Republic

Michaela Brázdilová[1] | *Czech Statistical Office; University of Economics, Prague, Czech Republic*
Petra Švarcová[2] | *University of Economics, Prague, Czech Republic*

Abstract

Income distribution strongly affects the value of risk-of poverty, what could explain small values of poverty rate in the Czech Republic. Therefore it is important to examine the level of income inequality in society and find out the socio-economic characteristics of people affecting the overall income inequality. The factor showing the biggest influence on the income level is education, so it is meaningful to examine the relationship between income inequality and poverty rate of each group of people by their highest attained education. One appropriate approach is quantification of each group's contribution to the overall income differentiation by decomposition of some income inequality indicators. This decomposition enables also to identify the reason the value of each contribution according to various aspects, such as the group size or total volume of groups incomes. The development of overall income inequality in the last year is a necessary condition for the prediction in the future, so the trends of time series of some inequality indicators were analyzed. The whole analysis enables to complete a view on income level and its inequality in the society, which are important indicators measuring the living standards of people.

Keywords

Income inequality, poverty rate, highest attained education, Theil index

JEL code

I00, I24

INTRODUCTION

Income is one of the appropriate indicators for evaluating of living standards of people, which provides the information about the economic well-being of individuals. It is an important component by assessing the overall quality of life of people, so it is important to examine the distribution of income and its degree of differentiation in the society. Is it possible to observe an impact of the level of education on the income situation of people? It is necessary to identify the dependence, how does the factor of education influence the overall income distribution and then the contributions of groups according to level of education to overall income inequality could be measured.

[1] Czech Statistical Office, Na padesátém 81, 100 82 Prague 10, Czech Republic; University of Economics, Prague, Faculty of Informatics and Statistics, Department of Economic Statistics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: michaela.brazdilova@vse.cz.

[2] University of Economics, Prague, Faculty of Informatics and Statistics, Department of Economic Statistics, W. Churchill Sq. 4, 130 67 Prague 3, Czech Republic. E-mail: petra.svarcova@vse.cz.

This paper is based on the assumption that there exists a directly proportional relationship between the level of poverty rate and income inequality. The main objective of this research is to find the differences between the level of income inequality among individual groups of people. The classification criterion of people was variable called "highest attained education" because of its clear influence on income level of people in society.

First, we find out the values of poverty rate and income inequality within each group. Then, we provide an analysis of income inequality value for all groups together. Like this, we can identify, which groups of people contribute to income inequality the most and which as the least of all. It is also possible to get the information about between-groups contribution to the overall income inequality. Finally, we can observe the situation within the groups over time.

1 MAIN PUBLICATION IN THE FIELD

Many articles in academic literature deal with poverty, income inequality, material deprivation etc. but there are many ways how inspect these matters. At first it can be viewed from a global perspective, as described Martin Ravallion in his article about globalization and poverty (2003). Poverty is usually understood as a level of living and can be perceived absolutely or relatively. According to (Ravallion, 2003) absolute poverty means a certain level of purchasing power. Each country shows a different absolute poverty line because their different purchasing power. In this case we talk about relative poverty, which depends on each area.

The poor people are those, whose consumption is below a given level of need. This poverty line is typically determined relatively to mean income, so the poverty rate depends on income distribution in society. The statement of Ravallion (2003) "How much more is held by rich people than by poor people" shows disparities in level of living, which is called inequality. Inequality can be also understood absolutely or relatively. Absolute inequality is given by absolute differences in level of living. On the other hand: "Relative inequality depends on the ratios of individual incomes to the overall mean." (Ravallion, 2003).

The measurement of poverty rate is derived from income distribution, so the poverty rate shows similar behavior as income inequality indicators. "The poverty measure can be hardly considered as sufficient statistics for judging the quality of people lives," (Ravallion, 2003). The poverty rate reflects more the income distribution in society, so it makes sense to identify the reason why the number of people living below given poverty line in the Czech Republic is measured by values of income inequality indicators.

The other article (Sirovátka & Mareš, 2006) focuses directly on poverty in the Czech Republic. The poverty rate indicates the percentage of people living below 60 per cent of the national income median. The data published on the Eurostat (2014) website show that the Czech Republic still achieves the lowest rate in Europe.

According to (Sirovátka & Mareš, 2006) it could be attributed to by relatively low national income median resulting in lower purchasing power, and a narrow income distribution. They claim that in the Czech Republic many people are between the 60 per cent and 70 per cent of threshold. This fact is also contributed to by former egalitarian character of the Czech social structure, where the redistribution of income is applied. The level of income distribution is presented by Eurostat (2014) and income inequality is, on average, higher in Europe than in the Czech Republic, which also proves income equality there.

According to (Marek, 2011) this lower value of the Gini index is caused by smaller amount of redistribution in the Czech Republic. In addition, this index is maintained for the last 10 years still at an approximately constant level. Also no significant differences are observed in accordance to sex or age. Different level of income inequality shows regional classification, where Gini index of Prague is comparable with values of other countries in European Union.

Dependence of income situation on education is described by (Finardi, Fischer, Mazouch, 2012). Significant differences in income level are observed especially by different study fields. Values of private rate of return on human capital vary considerably between various study fields. Whatever, in the Czech Republic this rate in general is higher than the OECD average, what is caused by tuition fees. This is reason, why the influence of education on income distribution in society is important to observe.

2 DATA AND METHODOLOGY

In order to examine the standard of living, we need to know the income situation of the population. Such information is available in EU-SILC survey (European Statistics on Income and Living Conditions), the most famous research, which collects data about households and persons living in the household. In this paper all computations are based on the data from research EU-SILC 2013 (ČSÚ-sk, 2014). These facts reflect income distribution for reference year 2012, for which these data were collected and officially presented on the Czech Statistical Office website (ČSÚ-ep, 2014).

This survey observes actually available income in each of households, which is called equalized disposable income. It is also appropriate to consider the average households income per consumption unit, which reflects the diversity of the economic structure of the household. According to (Jílek & Moravová, 2007): "The scale of consumption units for individuals is defined as relative volume of consumption (income) of various types of people, based on the consumption (income) of the selected type of person." The design of these consumables (equivalent) units reflects savings from the cost of items of mass consumption realized multi-households. For comparison of household incomes in the EU-SILC survey, the average income per modified consumption unit is used, because this most reflects the size and demographic composition of the household (Schechtman & Yitzhaki, 2007). For the first member whole one unit is considered, but other adults in households are weighted only by half a unit and for child under 13 years weight of 0.3 is used.

This equalized disposable income per consumption unit in household is allocated to each member of households. Then, the income situation can be compared between groups of people in society according to their social-economic characteristics.

For comparison of income distribution in society the classification of people by factor education representing the level of highest attained education was used. For our purposes, a detailed division was grouped into larger units in accordance with the classification of ISCED, and subsequently formed the groups:

- Primary = People with attained education of first or second grade of elementary school,
- Lower secondary = People with lower secondary education without leaving exam,
- Higher secondary = People with secondary education with graduation, or post-secondary courses,
- Tertiary = People with tertiary education (bachelor's, master's and doctoral graduates), including higher vocational schools.

Completely omitted is a group of children under 18 years and actively studying people under 26 years. They have not yet any own income and therefore their income is derived entirely from the earnings of their parents, or more precisely, it is budgeted to them from the total household income by the number of its economic consumer units. Therefore, their inclusion in the analysis of incomes by groups of education would not be relevant.

Between these education groups of people it is possible to observe differences in income distribution, income level and also income differentiation. It is appropriate to detect the level of income inequality in each of education groups and between these groups.

The growth of income inequality in society can be monitored by changes in the characteristics of variability and many income inequality indicators. Most commonly used measure of the concentration is the Gini coefficient with its graphical representation called Lorenz curve (Moravová & al., 2000). The more this curve deviates from the axis of the quadrant downwards the higher is degree of inequality in society.

Gini coefficient is the numerical representation which takes values from 0 to 1 and also higher values indicating larger income inequality. Extremes would be an absolute inequality ($G = 1$ – all incomes are held by one person) at the rate of 100% concentration (Moravová et al, 2000).

The rate of income inequality can also be described using the coefficient of income inequality (Jílek & Moravová, 2007). It measures proportion of the volume of income received by people in the top quintile and volume of income of people in the bottom quintile. The top (the fifth) quintile includes 20% of those with the highest income of ordered set of people by the size of income per modified consumption unit. The ratio of these values is noted as S80/S20 (Income quintile share ratio). The greater is the value of this coefficient, the larger income inequality exists. It indicates how many times larger income receives one-fifth of households with the highest incomes on average compared to a fifth of households with the lowest incomes (ČSÚ-mv, 2014).

Commonly used methods for assessing income inequality are taken from paper of Hesmati (2004) and several of these methods are also applied in this work.

The degree of income inequality in society can also be determined by using the Theil index of inconsistency. The formula of Theil index is presented by (Moravová et al, 1996):

$$\mathrm{T} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \left(- \frac{x_i}{\sum_{i=1}^{k} x_i} . \operatorname{In} \frac{\bar{x}_i}{\bar{x}} \right), \tag{1}$$

where *xi* presents total income of group *i*, $\bar{x}_i$ means average income in group *i*, $\bar{x}$ is average income in society and k represents the number of groups.

The advantage of this index is the possibility of its decomposition into subgroups (Moravová et al, 1996). This feature enables the extended use of Theil index according to (Ferreira, 2000) for examining the differences between income distributions of different groups of people. The decomposition of this index into groups can be based on measuring the income volumes and number of people in each group. If all groups showed the same population share as the income share, the overall index would be equal to 0 and it would be absolute equality. The index takes values in the interval (0, ln (*k*)), where the value of ln (*k*) means, that one person owns all income (Jílek & Moravová, 2007).

The following calculation formula of Theil index is appropriate for this purpose. It consists of two parts, the first is the sum of contributions of each group to the overall income inequality and the second indicates the contribution of income inequality between groups (Novotný, 2007). The formula is presented by (Ferreira, 2000):

$$\mathrm{T} = \sum_{\mathrm{i}=1}^{\mathrm{k}} w_i T_i + \sum_{\mathrm{i}=1}^{\mathrm{k}} w_i \operatorname{In} \frac{\mathrm{w}_i}{\mathrm{n}_i}, \tag{2}$$

where T_i is within-group Theil index, w_i represents income share and n_i means population share.

The Theil index value thus depends both on within-group variability and on between-group variability of income and not least on group size by volume incomes occurring. This index has the ability to decompose its value between multiple summands as the only one of its kind, thanks to the properties of logarithms therein used (Ferreira, 2000).

3 INCOME INEQUALITY BY LEVEL OF EDUCATION

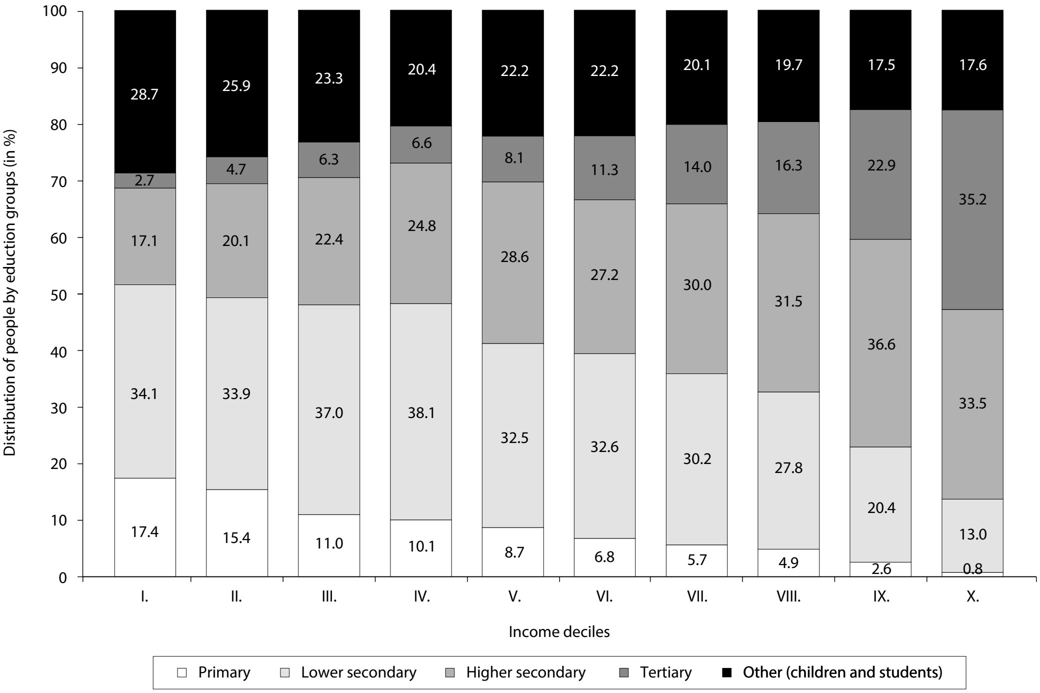
3.1 Income distribution within each group by level of education

Income level and income inequality within each of education groups is the first step in observing the differences between people according to their highest attained education. The above mentioned classification of people into the education groups was used in this analysis. The reasons for choosing the four education groups and omitting a group of children under 18 years and active studying people under 26 years have been explained above. The total equalized disposable income per consumption unit was recalculated for such an adjusted population.

The distribution of people into income deciles by their level of education is observable in Figure 1. People with the lowest (primary) education can be found most often in the lowest income groups

(I to III. decile). These people represent 17.4% of all people in the first decile. With higher income their share in groups is declining. In the highest income group (X. decile) they represent only 0.8%. Most often these are women aged over 50, who are either working or retired and live in multiple-member households. A similar trend can be seen among people with lower secondary education without leaving exam. Their share also falls with growing income. The opposite situation occurs by representation of people with upper secondary education with leaving exam or the highest (tertiary) education, which includes graduates from universities and higher vocational schools. In the first decile, the proportion of people with such education represents only 2.7% and with increasing income their representation in groups grows. Particularly, women aged 25 to 49 years living in numerous families, are in the first decile. Mostly they work as self-employed or are otherwise inactive persons. The highest influence has tertiary education on placements in the highest decile. The proportion of tertiary educated people is here just 35.2%, which means more than 10 times higher representation than in the first decile.

Figure 1 Distribution of people into income deciles by education groups (in %) in 2012



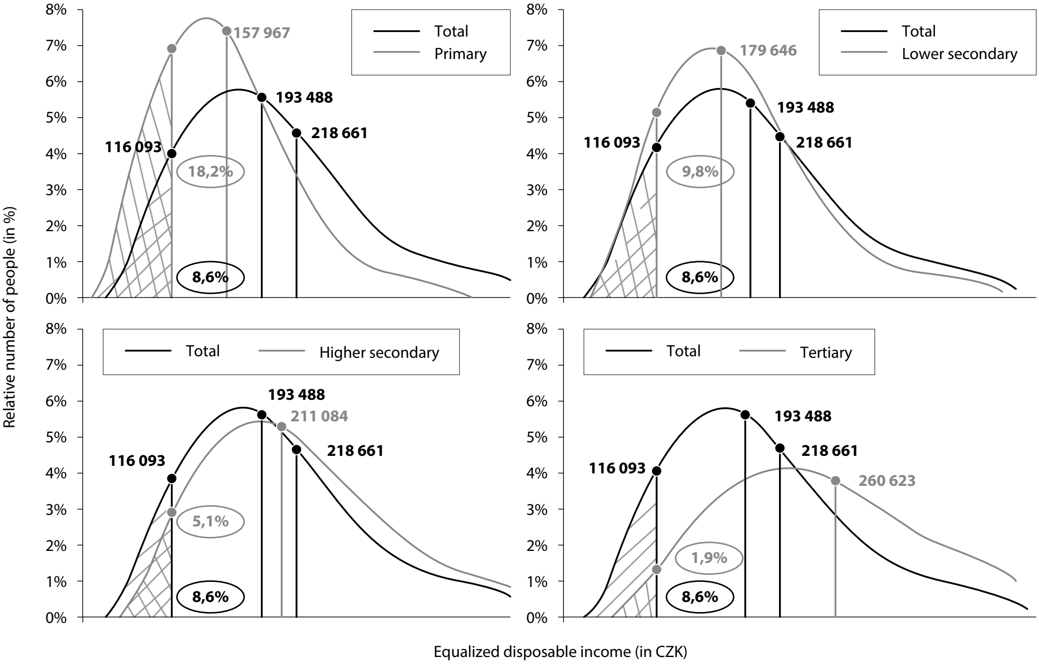
Source: Own calculations and creation in MS Excel using data from EU-SILC 2013

The distribution of people by income in total and subsequently also distribution in different groups by highest attained education shows Figure 2. The bold black line indicates the distribution of income among the total, where the average income for year 2012 CZK 218 661/year and median with obviously lower value of CZK 193 488/year. Other curves describe the situation in groups. In the second graph just their median income is used for transparency.

There is an obvious crosshatched area with people at risk of poverty, indicating those whose income is below the poverty line, which is defined as 60% of the median of equalized disposable income.

In this case it means the value of CZK 116 093/year. The poverty rate, or, more precisely, the percentage of people at risk of poverty against all persons, is 8.6%. The color areas are plotted for the groups, where the lowest level of poverty can be found among people with tertiary education with a value of just 1.9%, which is mainly due to their high income. Half of these people take more (median is CZK 260 623) than the average income of all members in society. Other extreme is the group of people with primary education, which are threatened with 18.2% of income poverty. This indicates their already low average and median income (CZK 157 967), both are also even less than the total median. Average income of secondary educated people is very close to the overall median and their poverty risk is about 9.8%.

Figure 2 Distribution of people according to equal. disposable income (CZK/year) by education groups



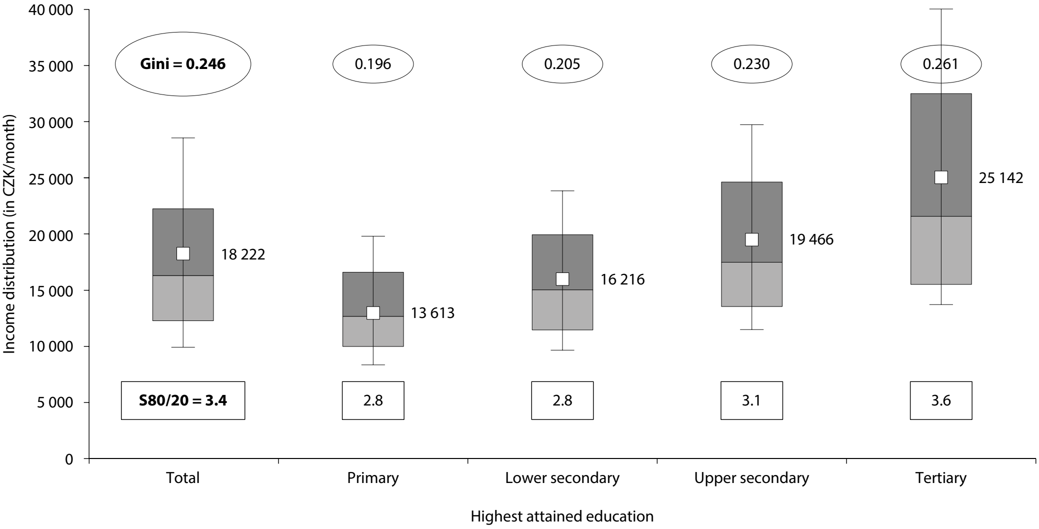
Source: Own calculations and creation in MS Excel using data from EU-SILC 2013

The inequality of income distribution among each of groups by the highest attained education of people shows Figure 3, it is created by using the box-plots. The little white squares indicate the level of average month income in the group, which is always higher than the median.

There are also boxes with middle six deciles. The coefficient of income inequality can be obtained by using the ratio of the top quintile to the bottom quintile. It is in examined year 2012 at overall size of 3.4, which indicates that persons in the upper two deciles take 3.4 times more than those in the two lowest deciles. This indicator corresponds to further indicator of income inequality, which is the Gini coefficient, here at the level of 0.246. It expresses how much the current situation differs from the absolute equality of incomes. The higher the greater the inequality is. Graphical representation using the Lorenz curve is shown in the Figure 4.

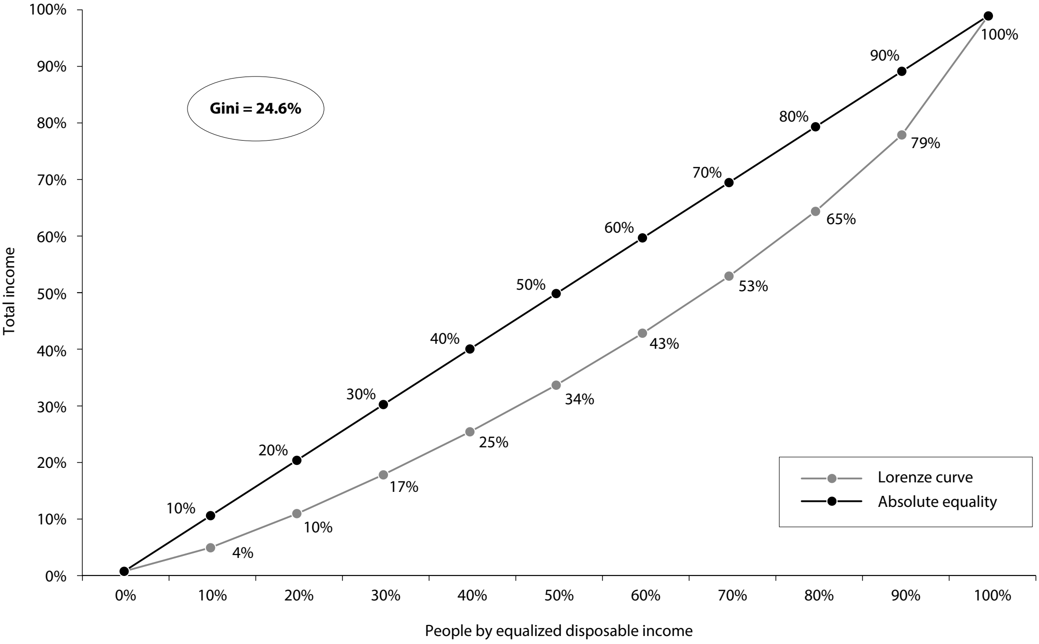
The Figure 3 shows that the higher level of education means the higher level of income and especially the higher level of inequality. Tertiary educated people have a great income range and Gini coefficient

Figure 3 Differentiation of equalized disposable income per month by group of education in 2012



Source: Own calculations and creation in MS Excel using data from EU-SILC 2013

Figure 4 Lorenz curve of inequality of equalized disposable income in 2012



Source: Own calculations and creation in MS Excel using data from EU-SILC 2013 in their group is at the level of 0.261. The richest fifth of them takes 3.6 times more incomes than the poorest fifth. Large differences in their income levels are given by their greater chances on the labor market, where tertiary educated people can easily evaluate the price of their work themselves. More aspects, than only wage or salary, are taken in consideration when choosing the work, because they can make decisions based on more opportunities.

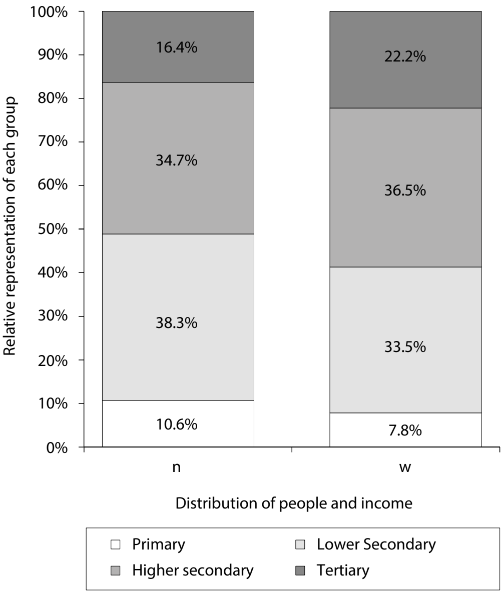
3.2 Decomposition of Theil index into groups by education

So far, analysis of income inequality was carried out in each group separately, which is useful for getting an idea of the differences between groups. However, to get the information about the contributions of the groups to overall income inequality is not enough. It offers analysis using Theil index, which allows us to perform the decomposition of income inequality in selected groups. We get not only an overview of how much a particular group contributes, but also about the influence of level of disparity between selected groups to overall inequality. Therefore we can see, how the between-groups variability participates in the total.

This index is based on comparison of the proportion of the number of persons in one group with the share of total incomes within this group. Does the amount of income correspond to the size of the group in there? The detection of just this inconsistency is then an obvious proof of income inequalities between persons in total, which are only classified into groups according to certain criteria.

The Figure 5 shows the differences between income and people distribution in each group of education for the year 2013. The two groups of people with primary and secondary education represent smaller volume of income than their frequency. Conversely, the population includes approximately 16% of people with tertiary education, by which we find more than 22% of total income volume.

Figure 5 Distribution of the group size (n) and total incomes (w) into groups by education in 2012

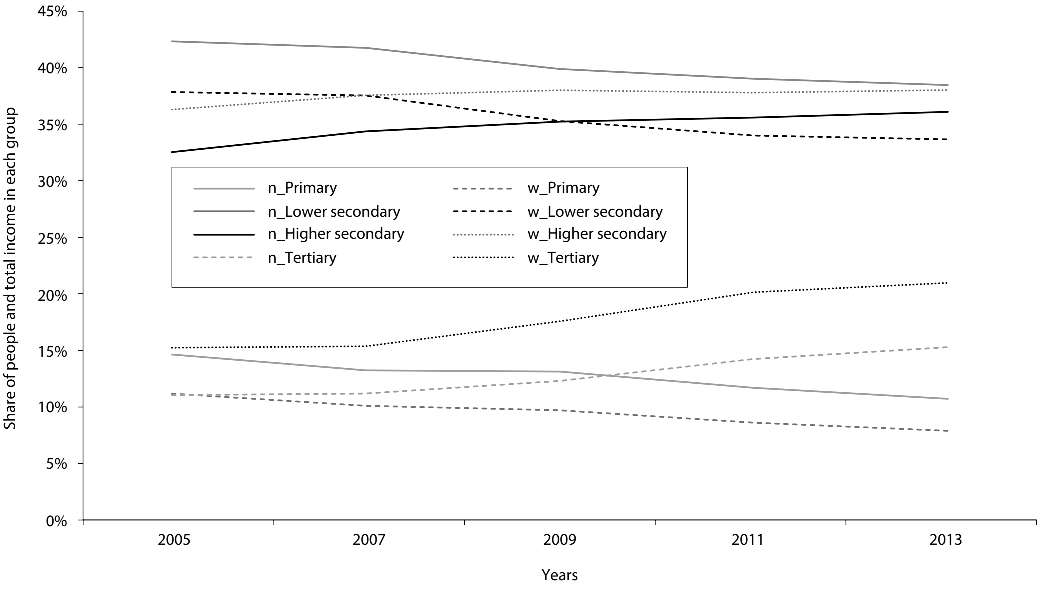


Source: Own calculations and creation in MS Excel using data from EU-SILC 2013

This finding is supported also by analysis of some previous years in Figure 6. Here we can see that the two first mentioned groups have a solid line indicating the relative proportion of people placed above than the dotted line expressing the share of total income attributable to the group. By the other two groups, the opposite is true. We can also observe the trends of changing of representation in these groups. The only obvious trend is the faster growth of incomes among the people with tertiary education than their number and, on the other hand, the convergence of proportion of the people with upper secondary education with their relative incomes. Otherwise, the two curves more or less correspond to each other.

This information is an important basis for the calculation of the Theil index. Its numerical expression has in itself almost no explanatory power, the follow size of the contributions of each group to the total value of income inconsistency are important. Contributions of each group to this index in relative expression are shown in Figure 7. The group of people with higher secondary education contributes to the total income inequality the most due to their one-third representation

Figure 6 Development of group size (n) and total income (w) into groups by education between 2005–2013



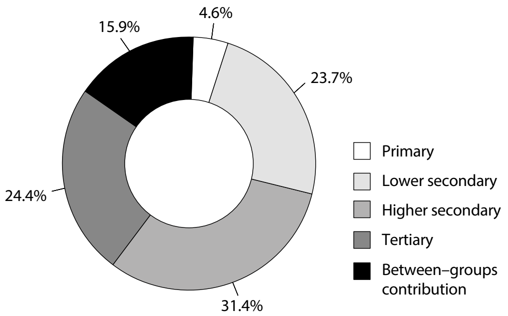
Source: Own calculations and creation in MS Excel using data from EU-SILC 2013

in society. Second highest value presents the group of people with tertiary education despite their small representation amounting to 16.4%. It is because of larger proportion of total equalized income, of which they dispose, and because of their huge income inequality within the group. This is the reason, why their contribution is even higher than that of persons with lower education, whose representation in the society is more than twice bigger than that of persons with tertiary education.

Between-group contribution with level of 15.9% occupies the fourth position between all contributions, which means that its value is not that significant. Income inequality within education groups indicates higher contributions than inequality between these groups.

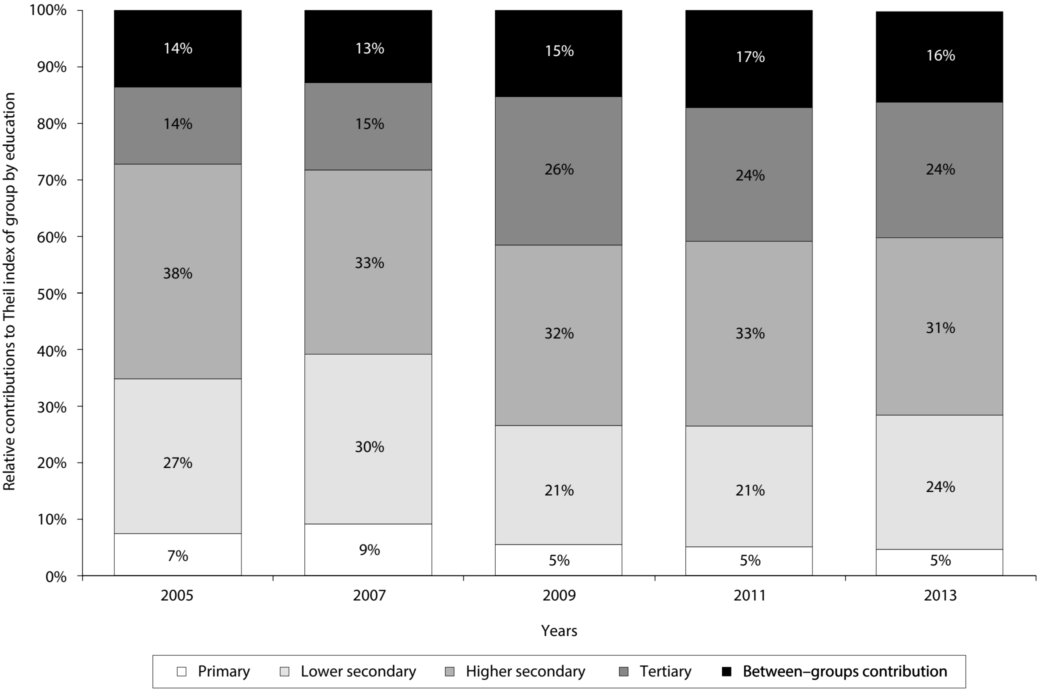
A significant increase of the contribution of people with tertiary education is evident if we focus on the development of these relative shares over the years in Figure 8, compared to a decline of contributions of persons with lower secondary education. No clear trend in the development of between group variability could be observed, its value rather fluctuates between years. However, there is an obvious increase of income inequality in 2009. The contribution of between-group inequality and contribution of persons with tertiary education increased at the expense of three lowest income groups. From this year, income inequality has not been increasing as fast as until the 2013, a moderate leveling of income has been recorded.

Figure 7 Contributions of education groups to Theil index in 2012



Source: Source: Own calculations and creation in MS Excel using data from EU-SILC 2013

Figure 8 Development of contributions of education groups to Theil index between the years 2005–2013



Source: Own calculations and creation in MS Excel using data from EU-SILC 2013

Based on the trends of income inequality observable in the last year we can create some opinions about the future development. Further increase of the number of people with tertiary education can be expected, which can cause their higher contribution to the total inequality due to greater proportion in the population and their even higher volume of total incomes, that is predictable based on Figure 6. The change of income inequality within the group can show different development trends. On one side, this can be increasing because of more various work requirements of people, but on the other hand, it can decrease due to leveling of work opportunities. This is connected with between-group variability, which could be declining in importance. Greater number of people with tertiary education possibly causes certain reduction of the education factor significance, so the differences between persons with higher secondary and tertiary education would be mitigated and their chances on the labor market could be equalized. These are just assumptions of possible development of income inequality, that it will be dependent on many other factors and especially on the development of the economic situation in society.

4 DISCUSSION

Some possible limitation can occur by providing a decomposition of income inequality. First of all, we consider the influence of population share in each group, what has an impact on the value of contribution and so distorts the effect of income inequality level within each group, It would be appropriate to achieve the same group volume, but this is impossible because of their classification according to education level. Other disadvantage of Theil index subsists in the fact, that its value is affected by the number of groups *k*. The relative contributions of each group to the total degree of income inequality are dependent on this. The variability of group averages increases with number of groups, so higher *k* means also higher between-groups contribution (Novotný, 2007).

The advantage of Theil index is the possibility of additional decomposition between subgroups and obtaining the within and between-groups inequality. Whatever, it is not limited by maximum values, so that interpretation of its size is very complicated. For measuring the income inequality just relative expression of group contributions are useful. According to European Commission (2010), Theil index presents a comprehensive, but complicated indicator of income inequality.

CONCLUSION

In this paper it was assumed, that income inequality affects the poverty rate, so detailed examination of income inequality was conducted. Thereafter, the large differences of level of income inequality between groups of people categorized by their highest attained education were detected. It is obvious, that the level of education has a significant impact on income inequality and the highest values of Gini coefficient are observed among persons with tertiary education.

The income inequality was confirmed by comparisons of each group size and income amount within this group. Among persons with tertiary education we find more than 22% of total equalized income amount while these persons represent only 16% of population. So this group makes the biggest contribution to overall income inequality. Conversely, persons with primary education do not have a significant influence on inequality.

By multiple comparisons, the within-group as well as the between-group contributes to overall income inequality were detected. The largest contribution to the Theil index occurs by the group of persons with higher secondary education, which is because of their most frequent representation in the population. Persons with tertiary education follow. They dispose of greater proportion of total income and show higher income inequality. Similarly, persons with lower secondary education contribute to income inequality more than those with primary education. It is because of their larger group size, despite their high income equality.

The between-group contribution represents almost 16%, so between-group variability has also certain impact to overall inequality. Over time, no significant trends in development of contributions of groups to overall income inequality are monitored.

It would be appropriate to produce also an analysis according to classification of people by their social status in society.

ACKNOWLEDGEMENTS

The paper has been prepared with the support of the University of Economics, Prague – Internal Grant Agency; project No. IGA 39/2015 "Estimate a reasonable rate of price regulation of housing expenditures". Many thanks belong also to the Czech Statistical Office for providing data necessary for the analysis according to the needs of the topic, which had vital importance for completion of this work.

References

ČSÚ-ep. (2014). *Příjmy a životní podmínky domácností 2013 – Ediční plán* [online]. Prague, 2014. [cit. 20.4.2015]. Available: <http://www.czso.cz/csu/2014edicniplan.nsf/p/160021-14>.

ČSÚ-mv. (2014). *Příjmy a životní podmínky domácností 2013 – Metodické vysvětlivky* [online]. Prague, 2014. [cit. 10.4.2015]. Available: <ttp://www.czso.cz/csu/2014edicniplan.nsf/t/E400292733/$File/16002114mc.pdf>.

ČSÚ-sk. (2014). *Příjmy a životní podmínky domácností 2013 – Stručný komentář* [online]. Prague, 2014. [cit. 15.3.2015]. Available: <http://www.czso.cz/csu/2014edicniplan.nsf/t/E400292725/$File/16002114kc.pdf>.

EUROPEAN COMMISSION. (2010). *Social mobility and Intra-regional income distribution across EU member states* [online]. DG Regional Policy – final report, July 2010. Available: <http://ec.europa.eu/regional_policy/sources/docgener/studies/pdf/sm_es_08072010_en.pdf>.

EUROSTAT. (2014). *European Commision. Eurostat* [online]. [cit. 21.4.2015]. Available: <http://epp.eurostat.ec.europa.eu/data/database>.

FERREIRA, P. C. (2000). *The Young Person's Guide to the Theil index: suggesting Intuitive Interpretations and Exploring Analytical Applications* [online]. UTIP WP N. 14, February 2000. Available: <http://utip.gov.utexas.edu/papers/utip_14.pdf>.

FINARDI, S., FISCHER, J., MAZOUCH, P. (2012). Private Rate of Return on Human Capital Investment in the Czech Republic: Differences by Study Fields [online]. ISSN 0322-788X. *Statistika*, 2012/1, pp. 23–30. Available: <https://www.czso.cz/documents/10180/20550299/e-180212q1k2.pdf/7a362763-ae77-4187-9d67-f811defc3ec1?version=1.0>.

HESHMATI, A. (2004). *A review of Decomposition o Income Inequality (IZA DP No. 1221)* [online]. Germany, 2004. Available: <http://ftp.iza.org/dp1221.pdf>.

JÍLEK, J., MORAVOVÁ, J. (2007). *Ekonomické a sociální indikátory.* Prague: VŠE. ISBN 978-80-86844-29-9.

MAREK, L. (2011). Gini Index in Czech Republic in 1995–2010 [online]. ISSN 1804-8765. *Statistika*, 2011/2, pp. 42–48. Available: <https://www.czso.cz/documents/10180/20542065/180211q2.pdf/d0c9b50b-bd62-45e8-ac1a-40c2969b5473?version=1.0>.

MORAVOVÁ, J. et al. (2000). *Úvod do sociálněhospodářské statistiky.* Prague: VŠE. ISBN 80-245-0006-X.

MORAVOVÁ, J. et al. (1996). *Sociálněhospodářská statistika I.* Prague: VŠE. ISBN 80-7079-366-X.

NOVOTNÝ, J. (2007). *On the measurement of regional inequality: Does spatial dimension of income inequality matter?* [online]. March 2007. Available: <http://web.natur.cuni.cz/~pepino/NOVOTNY2007AnnalsofRegionalScience.pdf>.

RAVALLION, M. (2003). *Debate on Globalization, Poverty and Inequality: Why measurement Matters* [online]. Available: <http://www.onlinelibrary.wiley.com/doi/10.1111/1468-2346.00334/pdf>.

SCHECHTMAN, E., YITZHAKI, S. (2007). *The "Melting Pot": A Success Story?* [online]. CBS, November 2007. Available: <http://www1.cbs.gov.il/www/publications/pw32.pdf>.

SIROVÁTKA, T., MAREŠ, P. (2006). *Poverty, Social Exclusion and Social Policy in the Czech Republic* [online]. Available: <http://www.onlinelibrary.wiley.com/doi/10.1111/j.1467-9515.2006.00490.x/pdf>.

Environmental-Economic Accounts and Financial Resource Mobilization for Implementation of the Convention on Biological Diversity

Cesare Costantino[1] | *Rome, Italy*
Emanuela Recchini | *Italian National Institute of Statistics (Istat), Rome, Italy*

Abstract

At the Rio "Earth Summit" the Convention on Biological Diversity introduced a global commitment to conservation of biological diversity and sustainable use of its components. An implementation process is going on, based on a strategic plan, biodiversity targets and a strategy for mobilizing financial resources. According to target "2", by 2020 national accounts should include monetary aggregates related to biodiversity. Environmental accounts can play an important role – together with other information – in monitoring processes connected with target "20": contribute to identifying activities needed to preserve biodiversity, calculating the associated costs and eventually assessing funding needs. In particular, EPEA and ReMEA are valuable accounting tools for providing data on biodiversity expenditure. The high quality of the information provided by these accounts makes them good candidates for being adopted world-wide within the Convention's monitoring processes. Enhanced interaction between statisticians and officials from ministries of environment would be crucial to reach significant advancement towards standardization of the information used in support of the Convention.

Keywords

Conservation of biological diversity, Aichi Biodiversity Targets, environmental-economic accounts, environmental expenditure, standardization

JEL code

Q56, Q57

INTRODUCTION

At the UN 1992 Conference on Environment and Development – the Rio "Earth Summit" (United Nations, 2015a) – the internationally agreed text of the Convention on Biological Diversity (CBD) was submitted for signature. It was one of the three "Rio Conventions", together with the United Nations Framework Convention on Climate Change and the United Nations Convention to Combat Desertification.

[1] Independent expert. Corresponding author: e-mail: costantino.cesare@alice.it, phone: (+39)3498828527.

At the completion of signature and ratification, in December 1993 the CBD entered into force (Convention on Biological Diversity, 2015a).

With the adoption of this international legal instrument, almost two hundred Parties[2] all over the world have committed themselves to the conservation of biological diversity and the sustainable use of its components. Hence a global long-term process for the implementation of the CBD is going on, based on the Decisions adopted by the Conference of the Parties (CoP), the CBD's governing body. The CoP has held twelve ordinary meetings, the last one in 2014; the next meeting is scheduled for 2016.

In 2014 the Mid-term Review of progress in implementation of the Strategic Plan for Biodiversity 2011–2020 was completed. An important effort for the collection and utilization of statistical data had been made world-wide in order to provide suitable information to support the assessments made on the occasion of the Mid-term Review. Actually, statistical information suitable for taking decisions and monitoring the attainment of agreed targets is an important point in the global process for the implementation of the CBD. The relevant data-sets cover a variety of domains, ranging from various fields of environmental sciences to economic aspects such as the costs of biodiversity conservation.

Financial aspects related to biodiversity are considered to be very important for the CBD, in the same way as the many distinct domains concerning environmental issues. As a matter of fact, the lack of sufficient financial resources has turned out to be one main obstacle in achieving the internationally agreed objectives. The main aspects to be taken into account include financial resources globally mobilized and expenditures spent in all countries committed to the CBD's implementation, as well as the amount of financial resources which is estimated as necessary to carry out the activities that are needed to that end.

Within statistics, environmental expenditure is one specific field which is thoroughly investigated: extensive statistical information on this topic is currently available. In particular, within official statistics, data on expenditure for the conservation of biodiversity and the sustainable use of its components is a mature statistical domain, more advanced as compared to other domains focused on the measurement of environmental phenomena. Expenditure aggregates derived from environmental accounts have an additional merit: they are calculated according to a system approach.

In general, when considering the economic statistical information available for the purposes of the CBD, the potential of environmental accounts should not be ignored. Environmental accounting within official statistics, in fact, can play an important role as a tool for providing indicators for decision makers as well as data for analytical work.

The subsequent paragraphs are focused on information concerning the mobilization of financial resources and environmental expenditures actually carried out for the achievement of the CBD's objectives. By reviewing developments that have taken place since the last decade, first an overview of processes related to the implementation of the CBD is given. Then the use of data on financial resources and expenditures for biodiversity is discussed. In that context, the system of integrated environmental-economic accounting is considered and its potential for the purposes of the CBD is highlighted; a specific focus is put on environmental expenditure aggregates. Some concluding remarks are emphasized mainly with the aim to encourage good interaction between official statisticians and those involved in political and administrative steps connected with targets of resource mobilization in support of the CBD.

1 THE ENDURING IMPLEMENTATION PROCESS OF THE CONVENTION ON BIOLOGICAL DIVERSITY

The implementation of the CBD is an enduring and complex process involving the engagement of international organizations and national governments all over the world. Like the international effort which

[2] According to the UN Glossary of terms relating to Treaty actions, Parties are the States as well as Organizations (for example the EU) that are bound by the CBD. All the States/Organizations that have either "ratified", "acceded to", "approved" or "accepted" the CBD are Parties to it (United Nations, 2015b).

had lead to the adoption of the CBD, its implementation is inspired by a global commitment to sustainable development. Mobilization of financial resources is a crucial element of this process.

A number of milestones have characterized the advancement of work carried out for the implementation of the CBD as well as the efforts put in place to ensure good governance for key processes. An overall strategic plan has been adopted for the conservation of biodiversity, as well as a strategy specific for mobilization of financial resources in support of the CBD. Work on indicators has been developed in the course of decades and a framework for reporting on financial aspects has recently been established.

1.1 The Convention on Biological Diversity and mobilization of financial resources in support of its implementation

The CBD's three objectives, as stated by Article 1, can be synthesized as follows: ensure that biodiversity is preserved by adopting economic and social development patterns that are environmentally sustainable and equitable at the same time.

As regards the financial resources that are necessary to implement the CBD, each country is committed to provide financial support in respect of its domestic activities intended to achieve the CBD's objectives, in accordance with its national plans, priorities and programs; furthermore, in order to help developing country Parties to fulfill the obligations deriving from the CBD, developed country Parties are committed to provide new and additional financial resources (Article 20 – Financial Resources).

Along with economic reasons which also exist, equity appears to be at the origin of the developed country Parties' additional commitment. Behind this there is the recognition that for developing country Parties economic and social development and eradication of poverty are priorities: in other words, following a strictly economic rationale, the opportunity costs of the conservation of biodiversity are particularly high for non affluent countries.

Having recognized the crucial importance of the financial resource mobilization undertaken both within countries and through international financial resource flows provided to help developing country Parties, the CoP has paid special attention to financial aspects. Through the CoP's Decisions, several elements have been put in place step by step to mobilize flows of money and eventually ensure that effective efforts are made in support of the CBD. Key steps have been the adoption of a strategy, a strategic plan, a set of indicators, a financial reporting framework.

In 2008, based on an in-depth review of the availability of financial resources for the purposes of the CBD, through Decision IX/11 the CoP encouraged the Parties and relevant organizations to improve the existing financial information through enhancing accuracy, consistency and delivery of existing data on biodiversity financing and improved reporting on funding needs (Convention on Biological Diversity, 2015b).

Furthermore, considering the urgency of coping with a difficult situation, through the same Decision IX/11 the CoP adopted the Strategy for resource mobilization in support of the achievement of the CBD's three objectives for the period 2008–2015. Strategic goals and objectives were defined, calling for concrete activities and initiatives to be developed to achieve the outlined goals; in addition, indicators were to be developed to monitor the implementation of the Strategy, which is noteworthy from a statistical viewpoint.[3] The first strategic goal was of particular relevance from the point of view of statisticians involved in the production of official statistics: according to Goal 1, the information base on funding needs and gaps – which also implies information on financial resources available – was to be improved.

[3] All this was to be done within appropriate timeframes, according to the Strategy.

1.2 The Strategic Plan 2011–2020 and the Aichi Targets

A very important step in the implementation of the CBD was the CoP's Decision "X/2 – The Strategic Plan for Biodiversity 2011–2020 and the Aichi Biodiversity Targets", adopted by the CoP in 2010 at its 10th meeting (Convention on Biological Diversity, 2015c).

The Strategic Plan, covering by definition all main aspects of the CBD, is based on five Strategic Goals.[4] Besides these, the Strategic Plan comprises a set of twenty biodiversity targets (Convention on Biological Diversity, 2015d) – known as Aichi Biodiversity Targets (ABTs) – which are organized under the Strategic Goals.[5] The CoP decided, through Decision X/3 of the same meeting, to adopt the ABTs at its next meeting, provided that robust baselines would have been identified and endorsed and that an effective reporting framework would have been adopted.

Though ambitious, the ABTs were considered to be achievable, some for 2015, others for 2020. One of them – ABT 2 – directly involves official statistics; another one – ABT 20 – implies the use of such statistics in one way or another, including for analytical work based on modeling.

ABT 2 is under Strategic Goal A ("Address the underlying causes of biodiversity loss by mainstreaming biodiversity across government and society"); it reads as follows: "By 2020, at the latest, biodiversity values have been integrated into national and local development and poverty reduction strategies and planning processes and are being incorporated into national accounting, as appropriate, and reporting systems". National accounts are mentioned explicitly in this target.

ABT 20 is under Strategic Goal E ("Enhance implementation through participatory planning, knowledge management and capacity building"). It reads: "By 2020, at the latest, the mobilization of financial resources for effectively implementing the Strategic Plan 2011–2020 from all sources and in accordance with the consolidated and agreed process in the Strategy for Resource Mobilization should increase substantially from the current levels. This target will be subject to changes contingent to resources needs assessments to be developed and reported by Parties". For ABT 20 also a baseline is needed – similarly to several other targets – for the purpose of measuring progress. Official statistics on flows of financial resources are involved by ABT 20; furthermore, any kind of official statistics – in particular national accounts, but also other official statistics – may be crucial to carry out analyses that are necessary for the foreseen assessments of resource needs.

In addition to what is reported above, through Decision X/10 the CoP also decided that the national reports due in 2014 should focus on the implementation of the 2011–2020 Strategic Plan and progress achieved towards the ABTs.

Concerning possible indicators in monetary terms for ABT 20, "Official Development Assistance provided in support of the Convention" was taken into consideration, but it was recognized that additional indicators could include the financial resources provided to developing countries which were dispersed through other mechanisms. Also, the global monitoring reports of the Strategy for resource mobilization were considered as useful to monitor the progress towards ABT 20.

As a matter of fact, through Decision X/3 a set of indicators was adopted to monitor the implementation of the Strategy for resource mobilization; several of them were in monetary units. Indicator 1 measured aggregated financial flows of biodiversity-related funding; it included both an overall amount, without double-counting, and the following categories: "Official Development Assistance" (ODA); "Domestic budgets at all levels"; "Private sector"; "Non-governmental

[4] The Strategic Goals (SGs) are as follows: SG A – "Address the underlying causes of biodiversity loss by mainstreaming biodiversity across government and society"; SG B – "Reduce the direct pressures on biodiversity and promote sustainable use"; SG C – "Improve the status of biodiversity by safeguarding ecosystems, species and genetic diversity"; SG D – "Enhance the benefits to all from biodiversity and ecosystem services"; SG E – "Enhance implementation through participatory planning, knowledge management and capacity building".

[5] While the goals and targets are intended for achievement at the global level, they also represent a flexible framework for the establishment of national or regional targets.

organizations, foundations, and academia"; "International financial institutions"; "United Nations organizations, funds and programs"; "Non-ODA public funding"; "South South cooperation initiatives"; "Technical cooperation". In addition to Indicator 1, the following indicators in monetary units were also adopted: Indicator 3 – "Amount of domestic financial support, per annum, in respect of those domestic activities which are intended to achieve the objectives of this Convention"; Indicator 4 – "Amount of funding provided through the Global Environment Facility[6] and allocated to biodiversity focal area"; Indicator 11 – "Amount of financial resources from all sources from developed countries to developing countries to contribute to achieving of the Convention's objectives"; Indicator 12 – "Amount of financial resources from all sources from developed countries to developing countries towards the implementation of the Strategic Plan for Biodiversity 2011–2020". The CoP also set out a process for elaborating and implementing the set of fifteen indicators it had adopted, including an expert consultation aimed at developing methodological guidance (United Nations, 2015c).

It is worth noting that the monetary indicators quoted above – Indicator 1 to Indicator 12 – correspond to different subsets of the economy and as a whole they cover the entire economic system. This suggests, in principle, that the information derived as appropriate from the system of national accounts, in particular from environmental accounts, could play a significant role as basic data for these indicators.

1.3 Recent developments

A preliminary reporting framework was agreed in 2012 for the indicators adopted to monitor the implementation of the Strategy for resource mobilization (Convention on Biological Diversity, 2015e). This framework was aimed at ensuring that, after adoption of targets, the attainment of the same targets could be monitored conveniently. Through a review of the said indicators, progress had been made in understanding which basic data could be taken into account to calculate them. It had been noted, in particular, that many of the indicators in monetary units relied on overlapping information for their calculation.[7] With a view to reducing the risk of double-counting, a limited set of "data fields" required to provide the information needed for the entire set of indicators had been identified; the Preliminary Reporting Framework was developed based on these "data fields".

This preliminary framework was intended for use by Parties to provide data on resource mobilization according to the adopted indicators. As concerns the calculation of these indicators, one suggestion was to organize the requested information by indicator and relevant set of basic data. Flows of financial resources for biodiversity from developed to developing countries and financial resources available in each country for biodiversity were the two main sets of basic data required to calculate the monetary indicators to be used to monitor the Strategy for resource mobilization.[8] For these sets of basic data a brief description of the distinct categories comprised in them was provided, as well as an indicative list of activities that could be considered for each category, while Parties were encouraged to add further possible activities that they might want to take into account.

Parties were also encouraged to interact, in completing the reporting framework, with their respective statistical offices. It was argued that some information needed was probably already available and should

[6] The Global Environment Facility is a partnership for international cooperation where 183 countries work together with international institutions, civil society organizations and the private sector, to address global environmental issues (GEF, 2015).

[7] For Indicator 1 it had been highlighted that some of its components were sub-categories of other components, some overlapped one another and many of them overlapped, completely or partially, with the other indicators; furthermore, there was an additional risk of double-counting in as much as in some cases these components were related to the end use of a flow of an international financial support while in other cases they consisted of amounts of international financial flows. In addition to that, Indicator 3 overlapped largely with the sum of components of Indicator 1, while Indicator 11 and Indicator 12 overlapped with several components of the same Indicator 1 (Convention on Biological Diversity, 2015e).

[8] The same could be said as far as monitoring of the attainment of ABT 20 is concerned.

have been used, where possible, in order to reduce duplicity of efforts; furthermore, a joint effort with statistical offices could lead to an improvement in the quality of the information used.

At the meeting held in October 2014 in Pyeongchang[9] the CoP has established, through Decisions XII/1 to XII/6, the Pyeongchang Roadmap for the enhanced implementation of the Strategic Plan for Biodiversity 2011-2020 and the achievement of the ABTs. Concerning financial resources matters, Decisions XII/1 and XII/3 include developments that have an impact on work carried out within official statistics.

Through Decision XII/3 (Convention on Biological Diversity, 2015f) the Strategy for resource mobilization has been extended until 2020. Also, having reviewed the progress towards the achievement of ABT 20, the CoP has adopted final targets concerning this Strategy.[10] The key importance of domestic resource mobilization for implementation of the Strategic Plan for Biodiversity 2011–2020 has been recognized: according to one of the final targets Parties provided with adequate financial resources endeavour to report domestic biodiversity expenditures, as well as funding needs, gaps and priorities, by 2015.

Most importantly from a statistical viewpoint, at the same 2014 meeting the CoP has adopted the revised Financial Reporting Framework, as Annex II to Decision XII/3. This is intended for use by Parties to provide baseline information and report on their contribution to reach the global financial targets, under ABT 20, as adopted through the same Decision XII/3.

2 STATISTICAL DATA ON FINANCIAL RESOURCES MOBILIZATION FOR THE CONVENTION: CURRENTLY USED DATA AND ECONOMIC AGGREGATES FROM OFFICIAL STATISTICS

The information needed to monitor the mobilization of financial resources in support of the CBD includes data that may or may not be produced within official statistics; several kinds of data are provided by other sources.

In particular, information on biodiversity-related funding and expenditures includes, according to the Financial Reporting Framework mentioned in the previous paragraph, distinct categories of data with different characteristics and quality: on international financial flows, which partly are “official” and partly relate to resources mobilized e.g. by non-governmental organizations; on expenditures, as resulting e.g. from public budgets as well as from environmental accounts; on funding needs, as assessed in National Biodiversity Strategies and Action Plans.

2.1 Data on financial resources mobilization according to the Convention’s Financial Reporting Framework

At present, reporting on financial aspects for the purposes of the CBD is based on the Financial Reporting Framework adopted in 2014, which is discussed here with limitation to what concerns information to be provided on funding and expenditure flows; other matters, e.g. priorities and plans or other assessments, are not discussed here.

The Framework includes reporting on baseline and progress towards 2015.[11] To this end, monetary data on the following flows are to be taken into account: international financial resource flows to developing countries and countries with economies in transition; current domestic biodiversity expenditures; funding needs and gaps. To identify biodiversity-related activities and thereby the corresponding monetary flows, an indicative list of possible classifications is suggested in the Appendix of the Framework, where reference is made to international work on this matter such as e.g. the guidance provided by OECD.[12]

[9] Twelfth meeting, the last meeting held by the CoP. Its thirteenth meeting is scheduled for December 2016.

[10] Preliminary targets on resource mobilization had been agreed in 2012 at the CoP’s eleventh meeting (Decision XI/4).

[11] Reporting on this part of the Framework is scheduled for December 2015.

[12] See: <http://www.oecd.org/dac/stats/46782010.pdf>.

The baseline concerns the international financial resource flows to developing countries and countries with economies in transition and 2010 is the reference year. If data is not available for that year, it is to be provided for the most recent year prior to that, and if possible, for the period from 2006 to 2010.[13] As concerns progress towards 2015, the years 2011 to 2015 are to be covered.

The data to be provided concerning international financial flows include official financial flows and resources mobilized by the private sector as well as non-governmental organizations, foundations, and academia. Data on official financial flows are presented under two main headings: Official Development Assistance (ODA), i.e. flows of official financing aimed at promoting economic development and welfare of developing countries; Other official flows (OOF), i.e. transactions by the official sector with countries on the List of Aid Recipients which do not meet the conditions for eligibility as Official Development Assistance or Official Aid. In order to identify official financial flows, in past reporting under the preliminary reporting framework several Parties used "Rio markers".[14] Data on resources mobilized by the private sector as well as non-governmental organizations, foundations, and academia are under the heading Other flows.

Concerning current domestic biodiversity expenditures, what needs to be reported is the annual financial support provided to domestic activities related to the conservation of biodiversity carried out in the reference year by the different sectors of society. Several years should be covered, if possible, starting with the most recent year for which the data are available.The data to be provided cover all sectors of the economy, but at least data on central government budget outlays directly related to biodiversity should be provided. Expenditures financed by international sources are to be taken into account, while funding provided to other countries is excluded. In past reporting, under the preliminary reporting framework, Parties made use of public budget data and also of the information derived from environmental protection expenditure accounts included in their systems of environmental-economic accounts.

As concerns reporting on funding needs and gaps, the reference year should be the year which is most appropriate for national planning purposes. The information requested is normally included in National Biodiversity Strategies and Action Plans.

Reporting on progress towards 2020 is also due.[15] Two main sets of data are requested in this context. First, the information on international financial resource flows is to be provided through the same data as in the section on progress towards 2015; these data are requested for the years 2016–2019. Secondly, each country should provide data on funding needs and gaps; these data are connected with the implementation of a country's national finance plan, and they include: the country's funding gap; the resource mobilization from domestic sources and from abroad achieved by the country; the remaining gap.

2.2 Environmental accounts aggregates

Within official statistics the interaction between economy and environment is described by means of different statistical tools. Two main categories can be distinguished in this regard: environmental statistics and environmental-economic accounts. The former include data that in some cases relate to both environmental and economic aspects simultaneously, but it is the latter that regularly link environmental and economic dimensions. Environmental-economic accounts are national accounts that are satellites to the core accounts of SNA, the system of national accounts (European Commission et al, 2009); they

[13] If specific annual figures are not available, the best estimates of average figures for 2006 to 2010 would have to be delivered.

[14] These Parties were members of the Development Assistance Committee of the OECD, which monitors aid provided for the purposes of the Rio conventions (Biological Diversity, Climate Change, Desertification). "Rio markers" are policy markers: external development funding for biodiversity purposes is labeled, and this is done by using a scoring system that highlights whether the funding is targeting biodiversity as its "principal" objective or simply as a "significant" one.

[15] Reporting on this section will take place in conjunction with the sixth national reports. As concerns the last national reports, their submission to the CoP had been requested for March 2014.

are based on a system approach and compiled according to an overarching international framework: the System of Environmental-Economic Accounting (SEEA), endorsed by the UN Statistical Commission (United Nations, 2015d).

2.2.1 The system of integrated environmental-economic accounts

SEEA has been developed by the UN Statistical Commission as a follow up to an input from Agenda 21. At the Rio "Earth Summit" the importance of integrating the statistical evidence that informs policy decision-making had been highlighted and the idea had been shared that, to monitor the transition to sustainable development, a system approach would help significantly.

SEEA provides a comprehensive conceptual accounting framework based on the same basic principles, definitions and classifications of SNA, thus allowing proper linkages with economic accounting data and other official statistics. Environmental and socio-economic statistics are reconciled and organized within the various SEEA modules, highlighting the interrelationships between the different phenomena covered; this allows the construction of time series of consistent, comparable and comprehensive statistics and indicators to monitor the contribution of the environment to the economy and the pressure of the economy on the environment, as well as the state of the environment. As a result, the trade-offs of policy-makers' decisions affecting natural resources and associated services are made explicit. The different domains of the environmental debate are suitably dealt with by accounts compiled according to SEEA; biodiversity is one of such domains.

Within environmental accounting in a broad sense, some SEEA components as well as SEEA-related initiatives provide tools which may be of particular interest to deal with the biodiversity theme. One example is the SEEA publication called System of Environmental-Economic Accounting 2012 – Experimental Ecosystem Accounting (United Nations et al., 2014a); another one is a SEEA subsystem: System of Environmental-Economic Accounting for Agriculture, Forestry and Fisheries (United Nations, 2015e); the global partnership Wealth Accounting and the Valuation of Ecosystem Services – WAVES – is also relevant (WAVES, 2015), being focused on research work on ecosystems valuation.

The main SEEA publication – System of Environmental-Economic Accounting 2012 – Central Framework (SEEA-CF) – is nevertheless of crucial interest in general (United Nations et al., 2014b); it deals with issues related to the interaction between economy and environment without being limited to the biodiversity theme.[16] Agenda 21 had explicitly proposed to develop integrated environmental-economic accounts, and the release of SEEA-CF has been the main response of the official statistics community to this. In 2012, after a global consultation that involved UN member countries, UN agencies, World Bank, IMF, OECD and the European Commission, SEEA-CF was adopted as an international statistical standard, similarly to SNA.

The above mentioned System of Environmental-Economic Accounting 2012 – Experimental Ecosystem Accounting is not an international statistical standard like SEEA-CF, but it complements the latter by providing methodological guidelines specific for ecosystem accounting and of course it is relevant when biodiversity is at issue. In general, it deals with biodiversity-related aspects more in detail and more comprehensively as compared to SEEA-CF; however, it is not specialized on aspects related to financial resources. Two specific environmental-economic accounts derived from SEEA-CF, instead, provide economic aggregates on biodiversity-related expenditure, which are of particular interest with connection to financial targets under ABT 20.

[16] Another SEEA publication is the following one: System of Environmental-Economic Accounting 2012 – Applications and Extensions. This latter publication and those on SEEA Central Framework and on Experimental Ecosystem Accounting mentioned above are known as the three SEEA publications.

2.2.2 Environmental expenditure aggregates

The Environmental protection Expenditure Account (EPEA) and the Resource Management Expenditure Account (ReMEA), derived from SEEA-CF, are the proper accounting tools to describe, in a national accounting perspective, expenditures carried out for environmental purposes, including those for conservation of biodiversity.

EPEA describes expenditures and economic activities performed to protect the environment against pollution and degradation phenomena (including loss of biodiversity); ReMEA describes expenditures and economic activities carried out to manage natural resources (e.g. forest resources, wild flora and fauna) and to save the stock of these resources against depletion phenomena. The expenditures and economic activities taken into account are those realized by resident units of the national economy; the overall aggregate derived from each of these accounts, known as national expenditure, includes consumption of environmental services and investments for their production. The total amount of the two national expenditure aggregates derived from EPEA and ReMEA is an assessment of the total economic effort devoted by a country to preservation of the natural environment.

Among distinct environmental domains that are covered in these accounts, two are relevant in relation to ABT 20. According to the classifications used, they are labeled as follows: “Protection of biodiversity and landscapes” as far as EPEA is concerned (classification: CEPA) and “Management of wild flora and fauna” as concerns ReMEA (classification: CReMA).[17]

The implementation of EPEA and ReMEA is particularly advanced within EU member countries, were a legal basis is in place for mandatory production of national environmental-economic accounts in line with SEEA-CF: Regulation (EU) No 691/2011 of the European Parliament and of the Council of 6 July 2011 on European environmental economic accounts (European Union, 2011), amended by Regulation (EU) No 538/2014 (European Union, 2014).[18] This legal basis provides methodology, common standards, definitions, classifications and accounting rules for the compilation of accounts that are given highest priority in the EU according to the European Strategy for Environmental Accounts – ESEA (European Statistical Committee, 2014). As concerns quality criteria, Regulation No 223/2009 shall apply (European Union, 2009).

Like all figures delivered within the European Statistical System, the EPEA and ReMEA expenditure aggregates are produced in compliance with the European Statistics Code of Practice – ESCP (Eurostat, 2011), which in turn is aligned with the UN Fundamental principles of official statistics (United Nations, 2015f); this applies in particular to EPEA and ReMEA data concerning the two environmental domains mentioned above, which is the information relevant in relation to ABT 20.

[17] CEPA (Classification of Environmental Protection Activities and Expenditure) is an international statistical standard; its item 6 – Protection of biodiversity and landscapes refers e.g. to measures and activities aimed at the protection and rehabilitation of fauna and flora species, ecosystems and habitats as well as the protection and rehabilitation of natural and semi-natural landscapes; measurement, monitoring, analysis activities as well as administration, training, information and education activities are also included; excluded are e.g. the protection and rehabilitation of historic monuments or predominantly built-up landscapes, the control of weed for agricultural purposes. CReMA (Classification of Resource Management Activities) has been developed within the European Statistical System for compiling statistics on the Environmental Goods and Services Sector; its item 12 – Management of wild flora and fauna refers to activities aimed at the minimization of the intake of wild flora and fauna through in-process modifications as well as withdrawals, reduction and regulation measures; restoration activities (replenishment of wild flora and fauna stocks) are included when aiming at maintaining/increasing the consistency of stocks (otherwise they come under CEPA item 6); measurement, monitoring, analysis activities as well as administration, training, information and education activities are also included; excluded is the protection of biodiversity which concerns essentially threatened species (under CEPA item 6).

[18] Regulation No 538/2014, in particular, includes provisions for the production of EPEA aggregates. A similar approach would be appropriate for the calculation of ReMEA aggregates.

3 ISSUES AND PROSPECTIVE FUTURE DEVELOPMENTS

In order to allow decision-makers to make decisions with a solid knowledge basis, high-quality statistics are needed. The general public also needs high-quality statistics, because people want to evaluate the performance of politicians and other decision-makers. Quality is a crucial point which in principle distinguishes official statistics as compared to other statistical information; the former are based on a set of fundamental principles and follow international statistical standards. The foundation for all this is the idea that democratic societies hardly function properly without a solid basis of reliable and objective statistics.

The world-wide applied UN Fundamental principles of official statistics and ESCP, mentioned before, target both outputs of statistical production and processes used, as well as institutional and organisational aspects. As concerns ESCP, fifteen principles and a set of indicators of good practice for each principle are adopted, while mandatory quality assurance procedures and a quality reporting system are in place.

Among the ESCP principles, "Professional Independence" and "Impartiality and Objectivity" might deserve special attention in some cases when considering the statistical information used within processes for monitoring the attainment of the global financial targets under ABT 20. Sound Methodology – another fundamental principle of official statistics – might also be an issue in some cases. With connection to this, it appears to be very constructive that Parties have been encouraged to interact with their respective statistical offices, not only because there is a need to avoid duplication of work, but because special attention should be devoted to quality of the data used: statistical offices could help to that end.

A special effort to promote interaction with the official statistics community might end up, in practice, with an increased use of official statistics in support of CBD's processes. This applies in particular to environmental accounts. A possible issue, in this perspective, would be the possibility to introduce, in the CBD's complex negotiations concerning monitoring activities, the intention to arrive, within appropriate timeframes, at a point where EPEA/ReMEA-type aggregates are systematically used worldwide for monitoring current domestic expenditures as requested by the Financial Reporting Framework.

Furthermore, it should be taken into account that available data from official statistics – including EPEA and ReMEA data, but not only this data – is a relevant and valuable potential input to the analytical work that is necessary to estimate financial resource needs. With connection to this, another possible issue would be to examine the extent to which such an input is actually used in assessments of financial resources needs.[19]

In general, sound methodology and comparability at the international level is a crucial point for statistical data. This has been recognized also with regard to the implementation of the Strategy for resource mobilization, for which reliable statistical information is needed. Then, a more general issue would be whether to adopt thoroughly concepts, definitions and classifications of environmental accounts and other official statistics while preserving essential rules given by the Financial Reporting Framework.

CONCLUSION

CBD's overall goal is twofold: first, biodiversity is to be preserved world-wide; secondly, this is to be done in an equitable way. Accordingly, two main instruments are in place: an overall strategic plan and a strategy for mobilizing financial resources. The strategic plan includes ABTs, which correspond to all CBD's purposes; in particular, ABT 20 concerns financial aspects. The strategy for mobilizing financial resources takes into account the effort to preserve biodiversity world-wide and to assist non affluent countries in their own effort for conservation of biodiversity.

[19] Or, otherwise – as concerns the preparation of National Biodiversity Strategies and Action Plans – to examine the extent to which EPEA and ReMEA data, together with other official statistics, actually contribute to the preparation of those strategic documents, from which information may be derived according to the Financial Reporting Framework.

As far as financial aspects are concerned, the information requested to monitor the attainment of financial targets under ABT 20 includes two distinct sets of data: on actual expenditures for activities intended to achieve the CBD's objectives; on funding needs and gaps and on financial flows from developed country Parties to developing ones. This approach to the collection of the information needed is tailored on the main purposes of the CBD.

Concerning official statistics, in general it would be natural that data derived from SEEA and SNA would be used extensively in the context of the CBD, together with other official statistics as well as other information; this happened in past reporting to some extent. As national accounts are referred to in ABT 2, in a sense the usefulness of SEEA and SNA aggregates is out of discussion.

Indeed, as far as ABT 20 is concerned, SEEA and SNA aggregates may turn out to be essential, together with other data, in order to monitor the achievement of financial targets: this applies in particular at the stage of identifying activities that are needed for conservation of biodiversity, then for the calculation of the costs associated with these activities and eventually for assessing funding needs and gaps.

In past reporting, for the purposes of monitoring financial targets under ABT 20, some Parties have provided EPEA/ReMEA-type data on expenditures related to conservation of biodiversity. Such an exercise could be extended and refined, provided that there is room for improving interaction between ministries of environment and statistical offices.

Perhaps an ad hoc developmental work at the international level focused on the proper way to single out, as appropriate, data from EPEA and ReMEA for the Financial Reporting Framework, could help. Classification issues would deserve special attention, because the guidance provided by the Financial Reporting Framework as concerns the set of activities to be considered for the calculation of biodiversity-related expenditures does not ensure that standardized information is provided by Parties. Statisticians' understanding of the scope and breakdown of EPEA and ReMEA data would have to be shared with ministries' officials. The final goal would be to enhance the accuracy and consistency of the data used within the Convention's implementation processes.

The costs and benefits of such an endeavor would include an advancement towards standardization; furthermore, the fact that to enhance accuracy and consistency national accounting aggregates would be used would represent an additional benefit. From the cost side, there would be an additional charge on statisticians and ministries' officials; the importance of this extra cost, however, depends on the priority that statisticians give to environmental-economic accounts and ministries to biodiversity.

Overall, when reflecting on the importance of enhancing accuracy and consistency of data on biodiversity expenditure within the Convention's implementation processes, the importance of the utilization of the financial resources committed to biodiversity targets should be emphasized: in the end, activities actually carried out to preserve biodiversity is what really matters.

ACKNOWLEDGEMENTS

The paper is the result of a joint effort of the authors. Anyhow Section 1, Section 3 and Conclusion can be attributed to Cesare Costantino; Introduction and Section 2 to Emanuela Recchini.

References

CONVENTION ON BIOLOGICAL DIVERSITY (a). *Convention on Biological Diversity* [online]. [cit. 10.8.2015]. <https://www.cbd.int/>.

CONVENTION ON BIOLOGICAL DIVERSITY (b). COP 9 *Decision IX/11* [online]. [cit. 10.8.2015]. <https://www.cbd.int/decision/cop/?id=11654>.

CONVENTION ON BIOLOGICAL DIVERSITY (c). *Strategic Plan for Biodiversity 2011–2020, including Aichi Biodiversity Targets* [online]. [cit. 10.8.2015]. <https://www.cbd.int/sp/>.

CONVENTION ON BIOLOGICAL DIVERSITY (d). *Aichi Biodiversity Targets* [online]. [cit. 10.8.2015]. <https://www.cbd.int/sp/targets/default.shtml>.

CONVENTION ON BIOLOGICAL DIVERSITY (e). *Implementation of the Strategy for Resource Mobilization – Preliminary Reporting Framework* [online]. [cit. 10.8.2015]. <https://www.cbd.int/financial/oda/malawi-preliminary-reporting-framework-2014-en.pdf>.

CONVENTION ON BIOLOGICAL DIVERSITY (f). *COP 11 – Methodological and implementation guidance for the "Indicators for monitoring the implementation of the Convention's Strategy for resource mobilization"* [online]. [cit. 10.8.2015]. <https://www.cbd.int/doc/meetings/cop/cop-11/official/cop-11-14-add1-en.pdf>.

EUROPEAN COMMISSION, INTERNATIONAL MONETARY FUND, OECD, UNITED NATIONS, WORLD BANK. *System of National Accounts 2008*. New York, 2009.

EUROPEAN UNION. *Regulation (EC) No 223/2009 of the European Parliament and of the Council of 11 March 2009 on European statistics and repealing Regulation (EC, Euratom) No 1101/2008 of the European Parliament and of the Council on the transmission of data subject to statistical confidentiality to the Statistical Office of the European Communities, Council Regulation (EC) No 322/97 on Community Statistics, and Council Decision 89/382/EEC, Euratom establishing a Committee on the Statistical Programmes of the European Communities.* Official Journal of the European Union, L 87, 31.3.2009.

EUROPEAN UNION. *Regulation (EU) No 691/2011 of the European Parliament and of the Council of 6 July 2011 on European environmental economic accounts.* Official Journal of the European Union, L 192, 22.7.2011.

EUROPEAN UNION. *Regulation (EU) No 538/2014 of the European Parliament and of the Council of 16 April 2014 on European environmental economic accounts.* Official Journal of the European Union, L 158, 27.5.2014.

EUROPEAN STATISTICAL SYSTEM COMMITTEE. *21st Meeting of the European Statistical System Committee.* Item 24 of the agenda – European Strategy for Environmental Accounts Work Programme Objective 2.21, Luxembourg, 14th–15th May 2014.

EUROSTAT. *European Statistics Code of Practice for the National and Community Statistical Authorities.* Adopted by the European Statistical System Committee 28th September 2011.

GEF. *Global Environment Facility* [online]. [cit. 10.8.2015]. <https://www.thegef.org/gef/>.

THE ECONOMICS OF ECOSYSTEMS & BIODIVERSITY. *Ecological and Economic Foundations* [online]. [cit. 10.8.2015]. <http://www.teebweb.org/our-publications/teeb-study-reports/ecological-and-economic-foundations/#.Ujr1xH9mOG8>.

UNITED NATIONS (a). *UN Conference on Environment and Development (1992)* [online]. [cit. 31.7.2015]. <http://www.un.org/geninfo/bp/enviro.html>.

UNITED NATIONS (b). *Treaty Collection* [online]. [cit. 10.8.2015]. <https://treaties.un.org/pages/Overview.aspx?path=overview/glossary/page1_en.xml>.

UNITED NATIONS (c). *Indicators for Monitoring the Implementation of the Strategy for Resource Mobilization – Methodological Guidance – Draft Proposal for Expert Consultation, June 2011* [online]. [cit. 10.8.2015]. <https://www.cbd.int/financial/doc/indicator-methodological-informal-guidance-en.pdf>.

UNITED NATIONS (d). *System of Environmental-Economic Accounting 2012 (SEEA)* [online]. [cit. 10.8.2015]. <http://unstats.un.org/unsd/envaccounting/seea.asp>.

UNITED NATIONS (e). *SEEA Agricolture, Forestry and Fisheries (SEEA AFF)* [online]. [cit. 10.8.2015]. <http://unstats.un.org/unsd/envaccounting/aff/chapterList.asp>.

UNITED NATIONS (f). *Fundamental Principles of National Official Statistics* [online]. [cit. 10.8.2015]. <https://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx>.

UNITED NATIONS, EUROPEAN COMMISSION, FAO, IMF, OECD, WORLD BANK (a). *System of Environmental-Economic Accounting 2012 – Experimental Ecosystem Accounting.* United Nations, New York, 2014.

UNITED NATIONS, EUROPEAN COMMISSION, FAO, IMF, OECD, WORLD BANK (b). *System of Environmental-Economic Accounting 2012 – Central Framework*. United Nations, New York, 2014.

WAVES. *Wealth Accounting and the Valuation of Ecosystem Services* [online]. [cit. 10.8.2015]. <http://www.wavespartnership.org/en>.

Wavelet-Based Test for Time Series Non-Stationarity[1]

Milan Bašta[2] | *University of Economics, Prague, Czech Republic*

Abstract

In the present paper, we propose a wavelet-based hypothesis test for second-order stationarity in a Gaussian time series without any deterministic components or seasonality. The null hypothesis is that of a second-order stationary process, the alternative hypothesis being that of a non-stationary process with a time-varying autocovariance function (excluding processes with unit roots). The test is based on the smoothing of the series of squared maximal overlap discrete wavelet transform coefficients employing modern techniques, such as robust filtering and cross-validation. We propose several test statistics and use bootstrap to obtain their distributions under the null hypothesis. We examine the test in settings that may mimic the properties of economic time series, showing that it enjoys reasonable size and power characteristics. The test is also applied to a data set of the U.S. gross domestic product to demonstrate its practical usefulness in an economic time series analysis.

Keywords	***JEL code***
Wavelets, time series, non-stationarity, bootstrap, hypothesis test, gross domestic product	*C 12, C 15, C 22, C 49, E 23*

INTRODUCTION

When referring to stationarity in this paper, we will mean second-order stationarity (see Section 2). Processes that are not stationary will be called non-stationary.

In the analysis of economic, financial and demographic time series, non-stationary time series are often assumed to have either unit roots and/or deterministic trends. The approach to unit root non-stationarity testing was pioneered by Dickey, Fuller (1979). Similar or extended approaches can be found in Said, Dickey (1984), Said, Dickey (1985) or Phillips, Perron (1988). Non-stationarity is, however, a much broader term. In fact, any time series, whose mean function or autocovariance function (to be defined in Section 2) are time-varying, is necessarily non-stationary.

We propose a bootstrap wavelet-based hypothesis test where the null hypothesis is that of stationarity and the alternative hypothesis that of a non-stationary process with a time-varying autocovariance function (generally excluding processes with unit roots).[3] The size and power of the test are estimated by Monte Carlo simulations. The results are compared with those of a test available in the literature. A data set of the U.S. gross domestic product is used to illustrate the implementation of the test.

[1] This paper was written with the support of the Czech Science Foundation (GA CR), grant no. P402/12/P507, Modelling of Financial and Economic Time Series – Application and Comparison of Wavelet and Traditional Methods.

[2] Department of Statistics and Probability, Faculty of Informatics and Statistics, University of Economics, nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: milan.basta@vse.cz.

[3] The reason for this exclusion will be made clear in Section 3 and Section 4.

We run the Monte Carlo simulations and implement the test utilizing R software (R Core Team, 2014). The following contributed R packages were widely used during the preparation of the paper: *wmtsa* (Constantine, Percival, 2013), *locits* (Nason, 2013a) and *forecast* (Hyndman, 2015).

The paper is organized as follows. A literature review is given in Section 1. Section 2 defines the notion of second-order stationarity. Section 3 provides a short introduction to the maximal overlap discrete wavelet transform (MODWT) and the properties of MODWT coefficients. The hypothesis test is introduced in Section 4. The properties of the test (size and power) are studied in Section 5. The test is applied to the first difference of the logarithm of the U.S. gross domestic product in Section 6.

1 LITERATURE REVIEW

Regarding the tests for stationarity other than unit root ones, we can mention Grinsted et al. (2004), who followed the idea of Torrence, Compo (1998), having applied the continuous wavelet transform to an input time series, calculating the so-called wavelet power and comparing the power values to critical values obtained under the assumption that the input time series was generated by a stationary AR(1) process. This type of test detects the instances of non-stationarity localized in time and scale (in an "otherwise stationary" process).[4]

Another wavelet-based test for stationarity has been proposed by von Sachs, Neumann (2000), who used localized versions of periodogram and Haar wavelet coefficients of the periodogram to decide about the stationarity of the underlying stochastic process. A test similar to that of von Sachs, Neumann (2000) has been introduced by Nason (2013b), who studied whether a specific linear transformation of the evolutionary wavelet spectrum (Nason et al., 2000) is time-varying or not. This was accomplished by exploring Haar wavelet coefficients of the empirical wavelet periodogram. If the null hypothesis of stationarity is rejected in the test by Nason (2013b), a locally stationary wavelet model with a time-varying autocovariance function is suggested as an alternative.

Variability in the series of smoothed or averaged squared wavelet coefficients is perceived – in an exploratory and descriptive sense – as qualitative evidence against stationarity also in other papers (see, e.g., Jensen, Whitcher, 2014, Whitcher et al., 2000, Nason et al., 2000, Fryzlewicz, 2005). None of these studies, however, include a hypothesis test that would provide the significance of this evidence.

Similarly to the papers mentioned above, we also smooth the series of squared wavelet coefficients using, however, a different smoothing approach. More specifically, we note that the median function of the logarithm of squared MODWT wavelet coefficients is constant over time for stationary Gaussian processes, being, however, generally time-varying for non-stationary Gaussian processes with a time-varying autocovariance function. We propose to use practices common in the field of statistical learning (see, e.g., Hastie et al., 2011) and non-parametric regression, such as cross-validation and robust filtering, to estimate the median function. Moreover, we do not downsample the series of the logarithm of squared MODWT wavelet coefficients in order not to lose any valuable information. Further, we propose several measures of non-constancy of the estimated median function to be used as the test statistic. Because of the complexity of the estimation procedure and analytical intractability of the distribution of the test statistic under the null hypothesis, approximate p-values are found by bootstrap. This leads to a computationally expensive test which may – as demonstrated by the results of the Monte Carlo simulations – enjoy better size and power properties than the test proposed by Nason (2013b) for time series lengths common in economics. Moreover, the time series length is not required to be a power of two, which is the requirement for the practical implementation of the test by Nason (2013b) in R *locits* package

[4] As will be discussed later, in hypothesis testing in general, the alternative hypothesis need not be necessarily well-specified. This is also the case of the test of Grinsted et al. (2004), where no explicit statistical model associated with the alternative hypothesis is given.

(Nason, 2013a). Moreover, our test provides a single p-value (for a given test statistic), differing from the test by Grinsted et al. (2004) where each "time-scale cell" is tested for stationarity individually and where no "global" decision about stationarity is provided.

2 SECOND-ORDER STATIONARITY

Following Brockwell, Davis (2002), let us assume a stochastic process $\{X_t: t = \ldots, -1, 0, 1, \ldots\}$ with $E(X_t^2) < \infty$, $t = \ldots, -1, 0, 1, \ldots$. Further, let $\mu_t \equiv E(X_t)$, $t = \ldots, -1, 0, 1, \ldots$, be the mean function and

$$\gamma(t+h,t) \equiv E\left[(X_{t+h} - \mu_{t+h})(X_t - \mu_t)\right], \quad t,h = \ldots,-1,0,1,\ldots, \tag{1}$$

the autocovariance function of the process, the variance function being defined as the autocovariance function for $h = 0$. The process is defined to be second-order stationary if both the mean and autocovariance functions (the latter for each h) are independent of time t.

When referring to stationarity in further parts of this paper, we will mean second-order stationarity. Processes that are not stationary will be called non-stationary.

3 MAXIMAL OVERLAP DISCRETE WAVELET TRANSFORM

In this section, the notion of MODWT coefficients is introduced together with the properties of these coefficients for stationary, integrated and locally stationary wavelet processes. These properties provide the basis for the hypothesis test.

3.1 MODWT wavelet filters and coefficients

The jth level ($j = 1, 2, 3, \ldots$) MODWT wavelet filter, denoted as $\{h_{j,l}: l = 0, \ldots, L_j - 1\}$, where L_j is the filter length, is[5] an approximately ideal linear filter for the frequency range $[1/2^{j+1}, 1/2^j]$. The filter is constructed in a very special way, fulfilling the following properties:

$$\sum_{l=0}^{L_j-1} h_{j,l} = 0, \quad \sum_{l=0}^{L_j-1} h_{j,l}^2 = \frac{1}{2^j}, \quad \sum_{l=0}^{L_j-1} h_{j,l} h_{j,l+2^j n} = 0 \text{ (for } n \neq 0). \tag{2}$$

There are various "families" of filters, such as the Haar, D(4), LA(8), etc.

Let $\{X_t: t = \ldots, -1, 0, 1, \ldots\}$ be a stochastic process which need not be stationary. The jth level ($j = 1, 2, 3, \ldots$) MODWT wavelet coefficients for $\{X_t\}$ are denoted as $\{W_{j,t}: t = \ldots, -1, 0, 1, \ldots\}$ and obtained by linear filtering $\{X_t\}$ with $\{h_{j,l}\}$, i.e.

$$W_{j,t} = \sum_{l=0}^{L_j-1} h_{j,l} X_{t-l}, \quad t = \ldots,-1,0,1,\ldots . \tag{3}$$

Percival, Walden (2002) show that the jth level MODWT wavelet coefficients are closely related to *changes* between two adjacent weighted averages of $\{X_t\}$ values, the weighted averages being calculated on an effective scale 2^{j-1}.

From Equation 3 and the first identity given in Equation 2, it follows that the constant mean function of $\{X_t\}$ is a sufficient condition for the zero mean function of $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$). Equation 3 also directly implies that $\{W_{j,t}\}$ (for j = 1, 2, 3, …) is a Gaussian process provided so is $\{X_t\}$.

[5] $\{h_{j,l}\}$ characteristics are based on Percival, Walden (2002, Ch. 5).

In the next sections, we will discuss the variance function of $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$) for stationary, integrated and locally stationary wavelet processes. The characteristics of the variance function for these types of processes will provide the basis for our hypothesis test.

3.2 Variance function of $\{W_{j,t}\}$ for stationary $\{X_t\}$

Let us assume a stationary stochastic process $\{X_t\}$. Since $\{W_{j,t}\}$ is the output from linear filtering of $\{X_t\}$ with $\{h_{j,l}\}$, it is stationary too and has a zero mean function. The variance of $W_{j,t}$ is called the wavelet variance and is denoted by v_j^2, i.e.

$$\nu_j^2 = \text{var}(W_{j,t}) = E(W_{j,t}^2), \quad j = 1,2,\ldots;\ t = \ldots,-1,0,1,\ldots . \tag{4}$$

Percival, Walden (2002, pp. 296) reveal that

$$\text{var}(X_t) = \sum_{j=1}^{\infty} \nu_j^2, \quad t = \ldots,-1,0,1,\ldots . \tag{5}$$

v_j^2 generally differs across various stationary stochastic processes.

3.3 Variance function of $\{W_{j,t}\}$ for integrated processes

Since a stationary $\{X_t\}$ has a stationary $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$), it follows that a non-stationary $\{W_{j,t}\}$ necessarily implies a non-stationary $\{X_t\}$. However, not all non-stationary processes have non-stationary $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$). Percival, Walden (2002, Ch. 8.2) show that $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$) for processes integrated of order d is stationary provided "Daubechies" filters[6] with $L_1 \geq 2d$ are used in the analysis.[7] Moreover, this implies that the variance function of $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$) is constant over time in such a situation.

3.4 Variance function of $\{W_{j,t}\}$ for locally stationary wavelet processes

In this section, we discuss a certain class of non-stationary processes, namely locally stationary wavelet processes as well as the variance function of $\{W_{j,t}\}$ for such processes. Since a thorough discussion might be too theoretical exceeding the scope and extent of this paper, the interested reader is referred to Nason et al. (2000), where details can be found.[8]

Nason et al. (2000, Def. 1) construct locally stationary wavelet processes making use of wavelet filters with random amplitudes as building blocks. Locally stationary wavelet processes are defined in a way implying that their mean function is zero (see Nason et al., 2000, pp. 274) and so is therefore the mean function of $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$). Locally stationary wavelet processes are characterized by the so-called (evolutionary) wavelet spectrum (Nason et al., 2000, Def. 2), which is generally time-varying. Nason et al. (2000, pp. 278) reveal that the generally time-varying autocovariance function of locally stationary wavelet processes tends (in a sense rigorously described in Nason et al., 2000) to the so-called local autocovariance defined in Nason et al. (2000, Def. 4) which depends on the wavelet spectrum and is time-varying in general.

Since the mean function of $\{W_{j,t}\}$ (for j = 1, 2, 3, …) is zero for locally stationary wavelet processes,[9] the variance function of $\{W_{j,t}\}$ is equal to the mean function of $\{W_{j,t}^2\}$. Nason et al. (2000, Prop. 4) show

[6] Daubechies filters include, among others, the Haar, D(4) as well as LA(8) family of filters.

[7] $L_1 = 2$ for Haar, $L_1 = 4$ for D(4) and $L_1 = 8$ for LA(8) filters.

[8] Nason et al. (2000) utilize a different notation than that used in our paper.

[9] It follows from Section 3.1 that the mean function of $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$) would be zero even if a non-zero constant was added to the locally stationary process.

that the mean function of $\{W_{j,t}^2\}$ is (up to a remainder term) equal to a special linear transformation[10] of the wavelet spectrum. Because the wavelet spectrum is generally time-varying for locally stationary wavelet processes, so is the linear transformation. Let us denote this transformation by $\varphi_{j,t}$.

Nason et al. (2000, Prop. 3a) show that all stationary processes with an absolutely summable autocovariance function are locally stationary wavelet processes. Nason et al. (2000, Prop. 3b) also prove that any locally stationary wavelet process that has a wavelet spectrum independent of time – and fulfills an additional restriction stated in Nason et al. (2000, Prop. 3b) – is stationary with an absolutely summable autocovariance function. The constancy of the wavelet spectrum over time also implies the constancy of the local autocovariance as well as that of $\varphi_{j,t}$ over time which turns into v_j^2, as introduced in Section 3.2.

3.5 Smoothing of the series of squared wavelet coefficients

Let us assume that the mean function of $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$) is zero (see Section 3.1 for the sufficient condition for the zero mean function). From Section 3.2, it further follows that for stationary $\{X_t\}$, the mean function of $\{W_{j,t}^2\}$ (for $j = 1, 2, 3, \ldots$) is constant over time. From Section 3.3, it follows that for integrated processes, the mean function of $\{W_{j,t}^2\}$ is also constant over time provided a long enough wavelet filter is employed in the calculation of wavelet coefficients. From Section 3.4, it follows that the mean function of $\{W_{j,t}^2\}$ tends to $\varphi_{j,t}$ and is generally time-varying for locally stationary wavelet processes.

As a result, the characteristics of the mean function of $\{W_{j,t}^2\}$ distinguish between stationary and non-stationary processes such as locally stationary wavelet ones, but generally not between stationary and integrated processes. Exploring the behavior of squared wavelet coefficients can thus be used as a means of non-stationarity detection. Such reasoning has also been applied in Torrence, Compo (1998) or Grinsted et al. (2004) even though a different type of wavelet transform than MODWT has been used.

Since the task is to estimate the mean function of $\{W_{j,t}^2\}$ from the squared wavelet coefficients, various smoothing and averaging techniques can be utilized for this purpose. Nason et al. (2000) used denoising based on wavelet transform. We can also mention various approaches to smoothing and averaging of squared wavelet coefficients utilized in other papers as an exploratory and descriptive tool for the detection of non-stationarity. See, for example, the smoothing of squared wavelet coefficients applied in Jensen, Whitcher (2014), the averaging of squared wavelet coefficients separately within different calendar seasons (resulting in the so-called seasonal wavelet variances) utilized in the study of Whitcher et al. (2000), the application of cross-validation to the choice of the smoothing parameter while smoothing the downsampled squared wavelet coefficients in Fryzlewicz (2005). In all of these studies, the non-constancy of the smoothed series of squared wavelet coefficients or the differences between seasonal variances have been interpreted as evidence against stationarity. However, a hypothesis test is not included in these studies that would provide the significance of this evidence. Nason (2013b) does not explicitly smooth the series of squared wavelet coefficients, but calculates its Haar wavelet coefficients and provides a formal hypothesis test for stationarity. However, as demonstrated in Section 5, the test of Nason (2013b) does not seem to be suitable – due to a relatively low power – for time series lengths typical in economics.

3.6 Synchronization of wavelet coefficients and boundary effects

$\{h_{j,l}\}$ is a causal linear filter. Consequently, $\{W_{j,t}\}$ (for $j = 1, 2, 3, \ldots$) is not synchronized with $\{X_t\}$, lagging behind it. Advancing $\{W_{j,t}\}$ by $\delta_j \geq 0$ time units is a way to approximately synchronize $\{W_{j,t}\}$ with $\{X_t\}$. The value of δ_j (given in Wickerhauser, 1994, p. 341; or in Percival, Walden, 2002, p. 118) depends on both the shape of the first-level wavelet filter and j. As a result, the process $\{w_{j,t}: t = \ldots, -1, 0, 1, \ldots\}$ (for $j = 1, 2, 3, \ldots$) defined as

[10] The weights of the linear transformation are time-independent.

$$w_{j,t} \equiv W_{j,t+\delta_j}, \quad t = ..., -1, 0, 1, ... \tag{6}$$

is approximately synchronized with $\{X_t\}$.

It is noteworthy that the characteristics of $\{W_{j,t}\}$, which can be used to distinguish a stationary process from a locally stationary wavelet process with a time-varying autocovariance function (such as a constant vs. time-varying mean function of $\{W_{j,t}^2\}$), are shared by $\{w_{j,t}\}$.

In real life applications, the observed time series of a *finite* length N can be considered the realization of a *portion* of a stochastic process; let this portion be denoted as $\{X_t: t = 0, \ldots, N - 1\}$. As a consequence, the coefficient $W_{j,t}$ of Equation 3 can be defined (if no further ad hoc assumptions on X_t for $t < 0$ are introduced) only for times $t = L_j - 1, \ldots, N - 1$. Further, the coefficient $w_{j,t}$ can be defined only for times $t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j$. Moreover, since L_j increases with j, an upper limit on the j value, denoted as J, is obtained by requiring N to be larger than L_j. As a result, j = 1, 2, …, J in real life applications.

4 HYPOTHESIS TEST

In Section 4.1, we discuss the null and alternative hypothesis of our test. In Section 4.2, we propose an approach to smoothing the series of squared wavelet coefficients. The smoothed series will establish the basis for the calculation of the test statistic (Section 4.3). Bootstrap will be used to obtain the approximation of the test statistic distribution under the null hypothesis (Section 4.4).

4.1 Null and alternative hypothesis

The null hypothesis of stationarity (Section 2) will be tested. In hypothesis testing in general, the alternative hypothesis need not be well-specified (see, e.g., Efron, Tibshirani, 1994, Ch. 16 and p. 233), the hypothesis test providing evidence whether or not the data are in agreement with the null hypothesis not necessarily pointing to any specific alternative model.

However, we can make the alternative hypothesis more specific in our test. We will *assume* that the mean function of $\{X_t\}$ is constant over time, that $\{X_t\}$ is Gaussian and contains no seasonality.[11] As argued in Section 4.2, under these assumptions, the rejection of the null hypothesis will point to a process with a time-varying autocovariance function – except for integrated processes. A locally stationary wavelet process with a time-varying autocovariance function can thus provide an excellent example of the model under the alternative hypothesis.

4.2 Smoothing in the logarithmic scale

Let the jth level (for $j = 1, 2, \ldots, J$) wavelet coefficients for $\{X_t\}$ be calculated and synchronized with $\{X_t\}$. Sequences $\{w_{j,t}: t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j\}$ (for $j = 1, 2, \ldots, J$) are obtained as a result.

We assume that the mean function of $\{X_t\}$ is constant over time, $\{X_t\}$ being Gaussian. Consequently, $\{w_{j,t}: t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j\}$ (for $j = 1, 2, \ldots, J$) is a Gaussian sequence with a zero mean function (see Section 3.1). We can thus write (for $j = 1, 2, \ldots, J$; $t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j$)

$$w_{j,t}^2 = \mathrm{var}(w_{j,t})U_{j,t}^2 = E(w_{j,t}^2)U_{j,t}^2 = E(w_{j,t}^2) + E(w_{j,t}^2)\left(U_{j,t}^2 - 1\right), \tag{7}$$

where $U_{j,t}$ is a zero-mean unit-variance Gaussian variable. From Equation 7 it follows that $\{w_{j,t}^2\}$ is heteroskedastic, the variance function of $\{w_{j,t}^2\}$ being proportional to the square of the mean function

[11] The assumption of no seasonality is in agreement with the fact that the mean function of $\{X_t\}$ is supposed to be constant over time, which would not be the case if deterministic seasonality was present. Moreover, for seasonally integrated processes, $\{W_{j,t}\}$ is non-stationary with a variance function varying over time. However, the proposed test is neither intended nor designed to test for seasonal unit roots in $\{X_t\}$. The assumption of no seasonality in $\{X_t\}$ thus avoids the need to deal with the seasonal pattern.

of $\{w_{j,t}^2\}$. In such a situation, logarithmic transformation stabilizes the variance (for a general discussion on variance-stabilizing transformations, see, e.g., Rawlings et al., 2001, Ch. 12.3), i.e.

$$z_{j,t} \equiv \log(w_{j,t}^2) = \log\left(E(w_{j,t}^2)\right) + \log\left(U_{j,t}^2\right), \tag{8}$$

where log stands for the natural logarithm.

Because of the properties of the mean function of $\{w_{j,t}^2\}$, which follow from Section 3, we can note that the distributions of $z_{j,t}$ (for $t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j$) generally differ only in their locations for locally stationary wavelet processes with a time-varying autocovariance function and are identical for stationary processes (and potentially also for integrated ones).

$\{w_{j,t}\}$ is correlated. The correlation, however, is very close to zero for lags greater or equal to 2^j, which is often utilized in the sense that if downsampled by 2^j, $\{w_{j,t}\}$ can be approximately treated as a sequence of uncorrelated random variables (see, e.g., Fryzlewicz, 2005, pp. 213–214, or Percival, Walden, 2002, Ch. 9). Consequently, since $\{w_{j,t}\}$ is Gaussian, $w_{j,t}$ (for a given t) can be assumed to be approximately independent of the set $\{w_{j,t+\tau}: |\tau| \geq 2^j\}$. Thus, $z_{j,t}$ (for a given t) can be assumed to be approximately independent of the set $\{z_{j,t+\tau}: |\tau| \geq 2^j\}$. Similar arguments are applied by Fryzlewicz (2005, p. 214).

4.2.1 Robust smoothing

There are alternative ways how smoothing and averaging of wavelet coefficients can be performed, as demonstrated by the examples given in Section 3.5 and noted by Fryzlewicz (2005).

We propose to work in the logarithmic scale since the logarithmic transformation leads to variance stabilization which will be useful while implementing cross-validation in Section 4.2.2. In fact, Nason et al. (2000, p. 282) also noted that logarithmic transformation could be a useful transformation applicable prior to the smoothing procedure even though they used a different approach to smoothing in the end.

It follows from the previous section that exploring the constancy of any measure of location of $\{z_{j,t}: t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j\}$ can serve as a means to distinguish between stationary and locally stationary wavelet processes with a time-varying autocovariance function. We propose to study the constancy of the median function[12] of $\{z_{j,t}: t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j\}$. This choice is determined by the fact that the median function can be *robustly* estimated using a locally weighted median smoother (Härdle, Gasser, 1984, see also Fried et al., 2007, Sec. 2.1 and 2.2 for the notion of the weighted median and weighted median filtering). The robust aspect of the estimation is appealing and necessary because of a heavy left tail[13] of $z_{j,t}$. More specifically, a symmetric weighted *median* filter of an odd length D is used in the paper, with weights $\{u_k: k = -(D-1)/2, \ldots, 0, \ldots, (D-1)/2\}$ given as[14]

$$u_k = \frac{1}{\sqrt{1+|k|}}, \quad k = -\frac{D-1}{2}, \ldots, 0, \ldots, \frac{D-1}{2}. \tag{9}$$

Since the median of $z_{j,t}$ differs from $\log(E(w_{j,t}^2))$ only by a constant independent of t, the changes in the median of $z_{j,t}$ over time are directly related to those in $\log(E(w_{j,t}^2))$ over time, and thus also to the (percentage) changes in $E(w_{j,t}^2)$ over time. This enables us to qualitatively explain the temporal

[12] The median function of a sequence of random variables is defined to be a sequence of medians of the random variables.

[13] In real life applications, $w_{j,t}^2$ can also be equal to zero due to rounding issues, which consequently leads to "minus infinite" values of $z_{j,t}$.

[14] At the boundaries of $\{z_{j,t}\}$, the symmetric weighted median filter cannot be fully employed and an asymmetric weighted median filter is used instead, being constructed by assuming only that part of the symmetric filter for which data are available.

variations in the median of $z_{j,t}$ by means of the temporal variations in $E(w_{j,t}^2)$; this approach being applied in some of the following parts of the paper.

4.2.2 Cross-validation

We use cross-validation to select an "optimal" span over which smoothing is to be performed. Such an approach to smoothing is considered a very flexible "statistical learning" technique in general (Hastie et al., 2011), its use being also advocated by Fryzlewicz (2005) for the smoothing of squared wavelet coefficients. We differ from Fryzlewicz (2005) by performing smoothing in the logarithmic scale. Further, in contrast to Fryzlewicz (2005), we do not downsample the series by 2^j, but work with dependent data not to lose any information. We note that such an approach is legitimate if a modified version of cross-validation, such as the "leave-$(2r + 1)$-out" cross-validation, is utilized (see, e.g., Arlot, Celisse, 2010, Sec. 8.1 or Chu, Marron, 1991), where the validation set is constructed so that it can be considered independent of the training set. In accordance with the previous discussion on (approximate) independence in the sequence $\{z_{j,t}\}$, we put r equal to $2^j - 1$. Moreover, we do not use the usual "least squares" cross-validation criterion because of the heavy-tailed nature of $z_{j,t}$. Instead, we utilize a robust cross-validation criterion (for ideas on robust cross-validation see, e.g., Morell et al., 2013).

More specifically, we implement cross-validation as follows. For a given length of the weighted median filter and a given m, for $m = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j$, a total number of $(2r + 1)$ observations, namely $z_{j,m-r}, \ldots, z_{j,m}, \ldots, z_{j,m+r}$, are left out[15] from the sequence $\{z_{j,t}: t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j\}$. A prediction from the median filter for time m is constructed using the remaining observations from the sequence, e_m (the prediction error) being defined as $z_{j,m}$ minus the prediction. Subsequently, the sequence $\{e_m: m = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j\}$ is sorted in the ascending order, the lowest 12.5% and the highest 12.5% of observations in this sorted set being removed. The mean of the *absolute values* of the remaining e_m values is calculated and serves as a cross-validation criterion. The optimal D is selected as such a length of the weighted median filter which minimizes the criterion. The weighted median filter of the optimal length is used to smooth $\{z_{j,t}\}$.

4.3 Test statistic

Let $\{\text{med}(z_{j,t}): t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j\}$ denote the median function of $\{z_{j,t}: t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j\}$ and let $\{q_{j,t}: t = L_j - 1 - \delta_j, \ldots, N - 1 - \delta_j\}$ be the median function estimate made using the approach described above. As will be explained in this section, we can also be interested in a linear combination of median functions such as

$$\sum_{j=1}^{J} \alpha_j \text{med}(z_{j,t}), \quad t = L_{j*} - 1 - \delta_{j*}, \ldots, N - 1 - \delta_{j*}, \tag{10}$$

where α_j, for $j = 1, \ldots, J$, are real-valued weights of the linear combination and $j*$ is the maximum value of j for which α_j is non-zero.[16] The linear combination can be estimated by $\{q_t^{\alpha}: t = L_{j*} - 1 - \delta_{j*}, \ldots, N - 1 - \delta_{j*}\}$ defined as

$$q_t^{\alpha} = \sum_{j=1}^{J} \alpha_j q_{j,t}, \quad t = L_{j*} - 1 - \delta_{j*}, \ldots, N - 1 - \delta_{j*}. \tag{11}$$

[15] For $m < L_j - 1 - \delta_j + r$ or $m > N - 1 - \delta_j - r$, the observations have to be left out asymmetrically.

[16] The value of $L_j - 1 - \delta_j$ is increasing with j, whereas that of $N - 1 - \delta_j$ is decreasing with j.

Efron, Tibshirani (1994, Ch. 16) reveal that the two quantities are the components of a bootstrap hypothesis test, namely a test statistic (which need not be an estimate of any parameter) and an estimate of the probability model under the null hypothesis. The choice of the test statistic determines the power of the test. Despite some arbitrariness in its choice, the test statistic has to measure the discrepancy between the null hypothesis (stationarity) and data at hand. Consequently, any reasonable measure of non-constancy of $\{q_t^\alpha\}$ can be suggested as a test statistic in our case. Let the statistic be generally denoted as *s*.

For example, the test statistic can be defined as the standard deviation of $\{q_t^\alpha\}$, i.e.

$$s \equiv \sqrt{\frac{1}{N - L_{j*}} \sum_{t=L_{j*}-1-\delta_{j*}}^{N-1-\delta_{j*}} \left(q_t^\alpha - \overline{q}_t^\alpha\right)^2}, \tag{12}$$

where

$$\overline{q}_t^\alpha \equiv \frac{1}{N - L_{j*} + 1} \sum_{t=L_{j*}-1-\delta_{j*}}^{N-1-\delta_{j*}} q_t^\alpha. \tag{13}$$

Analogously, the Spearman's rank correlation coefficient between $\{q_t^\alpha\}$ and time *t* can be used as the test statistic, i.e.

$$s \equiv \frac{\sum_{t=L_{j*}-1-\delta_{j*}}^{N-1-\delta_{j*}} \left(R_t - \overline{R}\right)\left(t - \overline{t}\right)}{\sqrt{\sum_{t=L_{j*}-1-\delta_{j*}}^{N-1-\delta_{j*}} (R_t - \overline{R})^2 \sum_{t=L_{j*}-1-\delta_{j*}}^{N-1-\delta_{j*}} \left(t - \overline{t}\right)^2}}, \tag{14}$$

where

$$R_t \equiv \text{rank of } q_t^\alpha \text{ in} \{q_t^\alpha : t = L_{j*} - 1 - \delta_{j*}, \ldots, N - 1 - \delta_{j*}\}, \tag{15}$$

$$\overline{R} \equiv \frac{1}{N - L_{j*} + 1} \sum_{t=L_{j*}-1-\delta_{j*}}^{N-1-\delta_{j*}} R_t, \tag{16}$$

and

$$\overline{t} \equiv \frac{1}{N - L_{j*} + 1} \sum_{t=L_{j*}-1-\delta_{j*}}^{N-1-\delta_{j*}} t = \frac{L_{j*} - 1 - \delta_{j*} + (N - 1 - \delta_{j*})}{2}. \tag{17}$$

The choice of the weights α_j (for $j = 1, \ldots, J$) in the linear combination of Equation 11 can influence the power of the test and is dictated by the research purpose. The basic version of the test corresponds to the situation where α_j is non-zero for a particular *j* only and zero for other *j* values. If we assume that the variance of large-scale changes[17] associated, for example, with $j = 4$, decreases, whereas the variance of short-scale changes, associated with $j = 1$, increases, we can use the following weights:

[17] See Section 3.1 for the relation of wavelet coefficients to changes occurring on various scales.

$\alpha_1 = 1$, $\alpha_4 = -1$ and $\alpha_j = 0$ for j equal neither to 1 nor 4. These weights contrast the dynamics on large and short scales. Further illustrations are provided in Section 5 and Section 6.

The test statistic of Equation 12 goes hand in hand with an exploratory part of the analysis where the variability of the smoothed series (or a linear combination of the smoothed series) of (the logarithm of) squared wavelet coefficients is generally perceived as evidence against non-stationarity (see also Section 3.5). The choice of the Spearman's rank correlation coefficient as the test statistic (see Equation 14) is useful in situations where the occurrence of a non-constant, close-to-monotone (not necessarily highly variable) median function or a non-constant, close-to-monotone linear combination of median functions is expected (see also Section 5 and Section 6 for further illustrations).

4.4 Bootstrap approximation of the p-value

A range of well-founded practices used in the field of statistical learning and non-parametric regression has been employed to smooth the series of the logarithm of squared wavelet coefficients. Due to the complexity of the smoothing approach, its properties are not analytically tractable. And neither is the distribution of the test statistic under the null hypothesis. Analytical intractability is common in complex, albeit standard procedures (see e.g. Faraway, 1992 for an illustration in regression). Efron, Tibshirani (1994, Preface and Ch. 1) and Davison, Hinkley (2009, Ch. 1) suggest, however, that bootstrap can be used to tackle analytically intractable problems. They reveal that bootstrap opens the door for the assessment of complex, even though useful practices by utilizing computer power instead of traditional statistical theory. This aspect of bootstrap can be considered as its strength, not weakness. They also stress that bootstrap avoids often potentially "harmful" and unnecessary oversimplification of the problem that would make it analytically tractable. As a result, some heuristics is commonly seen in the application of bootstrap approaches and bootstrap hypothesis testing in general, such as testing for the co-movement of two time series in time and scale (Grinsted et al., 2004, Ch. 3.4), or in other cases presented, e.g., in Efron, Tibshirani (1994, Ch. 16) or Davison, Hinkley (2009, Ch. 4).

The stochastic process, which represents our test input, is assumed to contain neither deterministic components nor seasonality. Since the test lacks power against unit root non-stationarity and is not intended to be used as a unit root test, differencing can be applied to remove potential unit roots prior to the test. Thus, procedures such as differencing, detrending and removing seasonality are assumed to have already been applied before the analysis if necessary. Further, let $\{X_t\colon t = 0, \ldots, N-1\}$ be the result of these potential procedures and let a stationary Gaussian ARMA model be assumed to fit $\{X_t\}$ reasonably well. Model-based resampling (see, e.g., Davison, Hinkley, 2009, Ch. 8.2.2) which employs this stationary ARMA model is used to generate B bootstrap time series and B bootstrap replications of the test statistic, s_b* for $b = 1, \ldots, B$. The bootstrap approximation of the hypothesis test p-value is obtained as (see, e.g., Davison, Hinkley, 2009, Ch. 4.2.3)

$$\frac{1+\sum_{b=1}^{B} I(s_b^* \geq s)}{B+1}, \tag{18}$$

where $I(s_b^* \geq s)$ is equal to one if $s_b^* \geq s$, and equal to zero otherwise.

5 SIZE AND POWER OF THE TEST

In this section, the size and power of the proposed test will be examined.

5.1 Size of the test

The following four stationary Gaussian processes are assumed, namely

1. an AR(1) process[18] with the autoregressive parameter equal to 0.9,
2. an AR(1) process with the autoregressive parameter equal to –0.9,
3. an MA(1) process[19] with the parameter equal to 0.8,
4. an MA(1) process with the parameter equal to –0.8.

Two patterns of α_j (for $j = 1, \ldots, J$) of Equation 11 are examined (denoted as *A* and *B*). Namely:

A: $\alpha_1 = 1$ and $\alpha_j = 0$ for $j = 2, \ldots, J$,

B: $\alpha_1 = -1$, $\alpha_3 = 1$ and $\alpha_j = 0$ for $j = 2, 4, \ldots, J$.

Since economic time series are often rather short, their length usually being of order of tens, the following two choices of *N* are assumed:[20] $N = 64$ and $N = 32$. Further, the two possible test statistics are assumed, namely the standard deviation (see Equation 12) and the Spearman's rank correlation coefficient (see Equation 14). 1 000 realizations are generated for each combination of the process type, α_j pattern, *N* and test statistic.

The hypothesis test is performed for each of the realizations and a decision made about the rejection or non-rejection of the null hypothesis – significance levels 0.01, 0.05 and 0.1 being used. Haar filters are employed. A stationary ARMA model which fits the realization is found using the *auto.arima()* function from R *forecast* package (Hyndman, 2015) and employing the following function arguments: max.p = 1,

Table 1 Size of the test estimated for various process types, series lengths (*N*), test statistics and α_j patterns. Each inner cell provides an estimate of the size of the test for significance levels 0.01, 0.05 and 0.1. The results are rounded to two decimal places

Process type, *N* = length of the series	Standard deviation, pattern A	Standard deviation, pattern B	Spearman, pattern A	Spearman, pattern B	Nason (2013b), Bonferroni	Nason (2013b), FDR
1, 32	0.01 0.04 0.09	0.01 0.04 0.11	0.01 0.05 0.09	0.01 0.05 0.10	0.00 0.00 0.00	0.00 0.00 0.00
2, 32	0.01 0.06 0.12	0.01 0.05 0.11	0.01 0.04 0.08	0.01 0.05 0.09	0.00 0.00 0.00	0.00 0.00 0.00
3, 32	0.00 0.04 0.09	0.01 0.05 0.09	0.02 0.05 0.11	0.00 0.04 0.10	0.00 0.00 0.00	0.00 0.00 0.00
4, 32	0.01 0.04 0.09	0.00 0.03 0.09	0.01 0.04 0.10	0.01 0.05 0.10	0.00 0.00 0.00	0.00 0.00 0.00
1, 64	0.01 0.04 0.07	0.01 0.04 0.09	0.01 0.06 0.10	0.01 0.05 0.10	0.00 0.00 0.00	0.00 0.00 0.00
2, 64	0.00 0.04 0.10	0.01 0.05 0.12	0.01 0.03 0.08	0.01 0.03 0.08	0.00 0.00 0.02	0.00 0.00 0.02
3, 64	0.01 0.04 0.09	0.00 0.04 0.10	0.01 0.05 0.10	0.01 0.04 0.09	0.00 0.00 0.00	0.00 0.00 0.00
4, 64	0.00 0.03 0.07	0.01 0.05 0.10	0.01 0.04 0.10	0.01 0.04 0.09	0.00 0.00 0.00	0.00 0.00 0.00

Source: Own construction

[18] $X_t = \phi X_{t-1} + a_t$, where $\{a_t\}$ is a unit-variance Gaussian white noise sequence, ϕ being a parameter.

[19] $X_t = a_t + \theta a_{t-1}$, where $\{a_t\}$ is a unit-variance Gaussian white noise sequence, θ being a parameter.

[20] Powers of two are assumed to allow a comparison with the test by Nason (2013b) (see below in this section for further details).

max.q = 1, max.d = 0, seasonal = FALSE, allowdrift = FALSE. The number of bootstrap replications is[21] $B = 99$. The results are presented in Table 1.

A comparison of the results with those obtained by the test proposed in Nason (2013b) is provided.[22] Nason (2013b) test utilizes, among others, multiple hypothesis testing, suggesting two approaches, namely the Bonferroni method and the false discovery rate (FDR) approach of Benjamini, Hochberg (1995).

The size of our test seems to be reasonably close to the nominal level (0.01, 0.05 or 0.1), no large deviation between the estimated size and the nominal level occurring in any of the simulation combinations. On the other hand, the test proposed in Nason (2013b) seems to be extremely conservative for the settings used in our simulation, the estimated size being far below the nominal level for all the simulation combinations.

5.2 Power of the test

The power of the test will be assessed under the conditions when the true data-generating process is either an AR(1) or MA(1) process with a time-varying coefficient. The values of the coefficient are stored in the sequence $\{\phi_t: t = 0, \ldots, N-1\}$ and are given as

$$\phi_t = 0.95\cos(2\pi\, F\, t / N), \quad t = 0,..., N-1, \tag{19}$$

where $F > 0$ is the frequency of the cosine. Two possible values can be attained in the simulation: $F = 0.5$ and $F = 1$. The process is thus defined either as (AR(1))

$$X_0 = \frac{a_0}{\sqrt{1-\phi_0^2}} \quad \text{and} \quad X_t = \phi_t X_{t-1} + a_t,\ t = 1,..., N-1, \tag{20}$$

or as (MA(1))

$$X_0 = a_0\sqrt{1+\phi_0^2} \quad \text{and} \quad X_t = a_t + \phi_t a_{t-1},\ t = 1,..., N-1, \tag{21}$$

where $\{a_t: t = 0, \ldots, N-1\}$ is unit-variance Gaussian white noise. The transition from high values (+0.95) to low values (–0.95) of the AR(1) (or MA(1)) coefficient is associated with a decrease in the variance of large-scale changes and an increase in the variance of short-scale changes. Both the models of non-stationary time series (AR(1) and MA(1)) are statistical ones, resembling also some of the non-stationary models used in Nason (2013b).

The other simulation settings are the same as in Section 5.1. The estimates of the power of the test are presented in Table 2. Again, a comparison with the test of Nason (2013b) is made.

As expected, higher values of N (ceteris paribus) mostly lead to a higher power of the test. It is also not surprising to often (but not always) find a higher power for the AR process than for the MA one (ceteris paribus). This is due to the fact that the time-varying coefficient is generally accompanied by more pronounced variations in the variance of changes associated with the examined scales in the case of the AR process.

[21] General considerations of Davison, Hinkley (2009, Ch. 4.2.5) on the choice of B in bootstrap hypothesis testing suggest that too small values of B can lead to a loss of the size and power of the test. Davison, Hinkley (2009, Ch. 4.2.5) conclude that 99 bootstrap replications should generally be sufficient provided the significance level is greater or equal to 0.05. A little larger loss of size and power, though not a serious one, can occur if 99 bootstrap replications are used with the 0.01 significance level.

[22] The test proposed in Nason (2013b) is implemented in *hwtos2()* function in R *locits* package (Nason, 2013a). Default settings of the function parameters are used. This function works only with a time series whose length is a power of two.

Table 2 Power of the test estimated for various process types, values of *F*, series lengths (*N*), test statistics and α_j patterns. Each inner cell provides an estimate of the power of the test for significance levels 0.01, 0.05 and 0.1. The results are rounded to two decimal places

Process type, *F*, *N* = length of the series	Standard deviation, pattern A	Standard deviation, pattern B	Spearman, pattern A	Spearman, pattern B	Nason (2013b), Bonferroni	Nason (2013b), FDR
AR, 0.5, 32	0.05 0.22 0.35	0.17 0.44 0.57	0.06 0.18 0.27	0.14 0.35 0.49	0.00 0.00 0.00	0.00 0.00 0.00
MA, 0.5, 32	0.01 0.08 0.17	0.02 0.13 0.24	0.03 0.10 0.17	0.06 0.19 0.30	0.00 0.00 0.00	0.00 0.00 0.00
AR, 1, 32	0.08 0.26 0.41	0.10 0.34 0.47	0.00 0.01 0.02	0.00 0.01 0.02	0.00 0.00 0.00	0.00 0.00 0.00
MA, 1, 32	0.01 0.09 0.18	0.01 0.08 0.16	0.00 0.02 0.06	0.01 0.04 0.07	0.00 0.00 0.00	0.00 0.00 0.00
AR, 0.5, 64	0.13 0.39 0.53	0.35 0.62 0.74	0.12 0.27 0.38	0.25 0.49 0.64	0.00 0.01 0.05	0.00 0.01 0.05
MA, 0.5, 64	0.03 0.12 0.21	0.10 0.29 0.41	0.06 0.14 0.24	0.13 0.30 0.44	0.00 0.00 0.00	0.00 0.00 0.00
AR, 1, 64	0.17 0.45 0.58	0.36 0.61 0.70	0.00 0.00 0.01	0.00 0.00 0.01	0.00 0.04 0.07	0.00 0.04 0.08
MA, 1, 64	0.02 0.09 0.19	0.05 0.18 0.29	0.00 0.02 0.04	0.00 0.01 0.03	0.00 0.00 0.00	0.00 0.00 0.00

Source: Own construction

A change in the pattern of α_j values from A to B, ceteris paribus, has generally resulted in an increase in the power of the test in our simulations. This is due to the fact that if the variance of short-scale changes increases, the variance of large-scale changes decreases in our case, and vice versa. Consequently, a more powerful test can be obtained using pattern B which contrasts the dynamics on short and large scales.

The test employing the Spearman's rank correlation coefficient as the test statistic is much more powerful for $F = 0.5$ than $F = 1$ (ceteris paribus). The reason for this behavior is obvious since $\{q_t^\alpha\}$ (or $\{q_{j,t}\}$) is expected to be close to monotone for $F = 0.5$, while no such monotone behavior can be expected for $F = 1$ because, for $F = 1$, the time-varying coefficient starts to revert back to its original value at a time half-way from the start and so does the variance of changes associated with individual scales. It is also interesting to note that the test using Spearman's rank correlation is more powerful than the one using the standard deviation provided $F = 0.5$, the process type being MA. This can presumably be explained by the fact that $\{q_t^\alpha\}$ (or $\{q_{j,t}\}$) tends to be quite close to monotone in such a situation ($F = 0.5$, process type = MA) and not too variable (in terms of the standard deviation).

For the settings used in our simulation, the power of the test proposed by Nason (2013b) is often inferior to ours. Nason (2013b) also experimented with various non-stationary models and reports very good power characteristics of his test. This can be explained by the fact that much longer time series were used in his simulation, namely the length of 512 was assumed. To further support this claim, we have run additional minor Monte Carlo simulations. More specifically, having assumed an AR(1) model with a time-varying autoregressive coefficient and $F = 0.5$ (see Equations 19 and 20), we studied the power of the test proposed by Nason (2013b) in dependence upon the length of the series (N), using the 0.05 significance level. The following power estimates have been obtained from 1 000 simulations (the first number corresponding to the Bonferroni method, the second to the FDR approach):

0.00 and 0.00 for $N = 32$, 0.01 and 0.01 ($N = 64$), 0.36 and 0.38 ($N = 128$), 0.94 and 0.97 ($N = 256$), 1.00 and 1.00 ($N = 512$). These additional results suggest that a reasonable power of the test proposed by Nason (2013b) can be obtained provided that the time series is long enough. We have not included such long time series in our major simulation since our test aims – by its design, where cross-validation, robust filtering and bootstrapping is utilized – at rather short time series, not being directly applicable to very long time series due to extensive computations that would be required. For short time series, such as those often occurring in economic settings, our test is, however, expected to enjoy reasonably good size and power characteristics, having reasonable computational demands.

6 APPLICATION OF THE TEST TO THE U.S. GROSS DOMESTIC PRODUCT

We illustrate the hypothesis test using the yearly time series of the U.S. gross domestic product (GDP) retrieved from the U.S. Bureau of Economic Analysis, FRED (2015). The time series is given in current prices in billions of dollars, measuring the value of goods and services in each year's prices. GDP measured in constant prices (i.e. adjusted for inflation) is preferable when the focus is on actual productivity growth, GDP in current prices being of potential interest, for instance, for monetary policy objectives (see, e.g., Feldstein, Stock, 1994; Bernanke, Mishkin, 1997). The decision to use current prices instead of constant ones in this paper is due to the fact that the former facilitate a better demonstration of the various aspects and settings of the hypothesis test. Moreover, the yearly nature of the time series also avoids the need to deal with the seasonal pattern.

The time series will be denoted as $\{Y_t: t = -1, 0, \ldots, N - 1\}$, where $N = 85$, time $t = -1$ corresponding to the year 1929 and $t = 84$ to 2014, the total length of the time series being $N + 1 = 86$. The hypothesis test will be performed on $\{X_t: t = 0, \ldots, N - 1\}$ which is defined as the first difference of the natural logarithm of $\{Y_t\}$, i.e.

$$X_t \equiv \log(Y_t) - \log(Y_{t-1}), \quad t = 0,\ldots, N-1. \tag{22}$$

Haar filters are employed. A stationary ARMA model which fits $\{X_t\}$ is found using the *auto.arima()* function from R *forecast* package (Hyndman, 2015) – the function is called with the following arguments: max.p = 4, max.q = 4, max.d = 0, seasonal = FALSE, allowdrift = FALSE. The maximum possible orders of the AR and MA part of the model are chosen to be equal to four (i.e. max.p = 4, max.q = 4) so that the model can be flexible enough if needed. The number of bootstrap replications is $B = 499$.

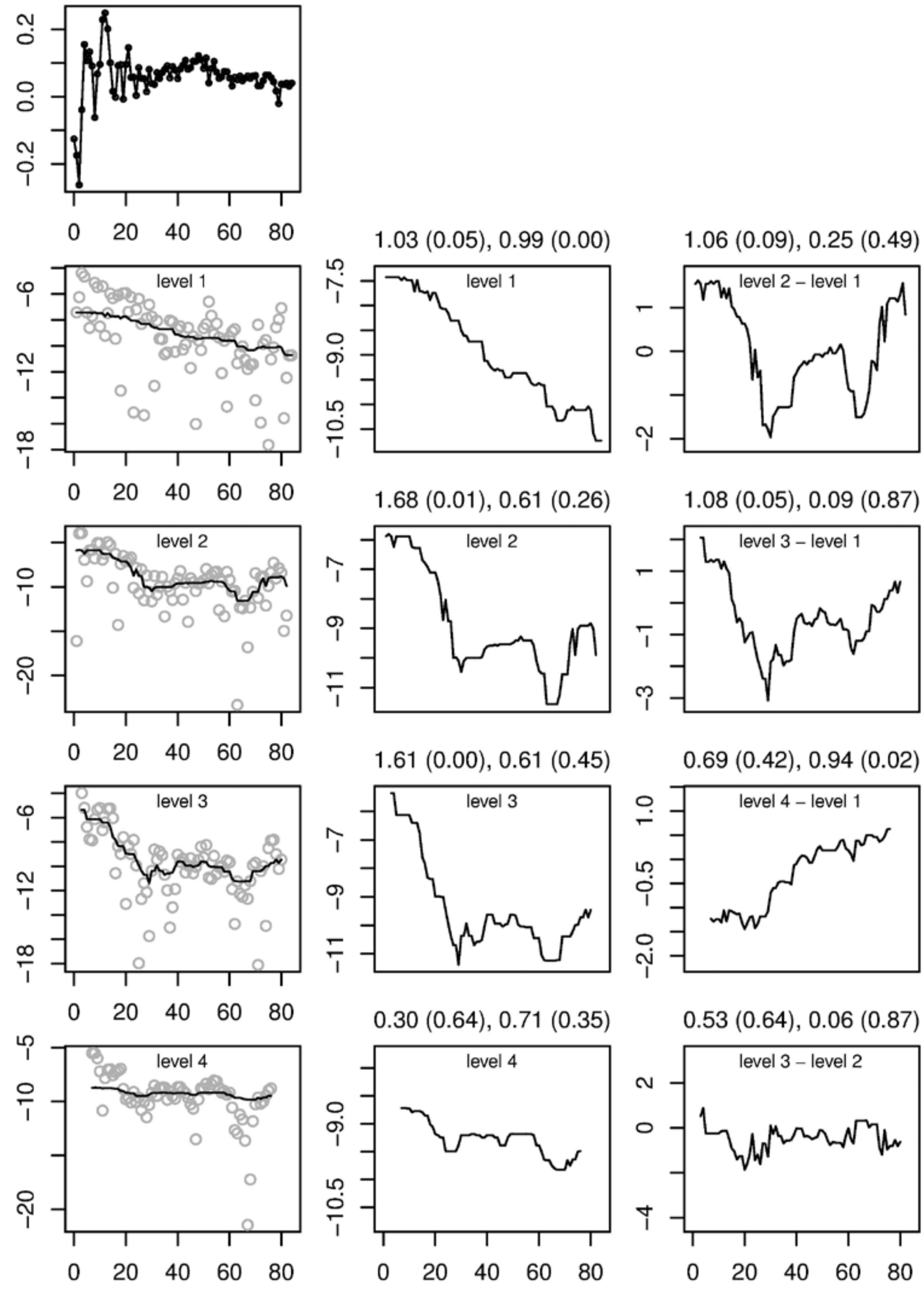
The results are presented in Figure 1. More specifically, the uppermost subfigure in the first column displays the time series $\{X_t\}$. Below this column subfigure, $\{z_{j,t}\}$ ($j = 1, \ldots, 4$) are presented (gray color) in separate subfigures, together with $\{q_{j,t}\}$ ($j = 1, \ldots, 4$) (black color). The asymmetric nature of $z_{j,t}$ with a heavy left tail can be clearly observed.

The subfigures in the second column display $\{q_{j,t}\}$ ($j = 1, \ldots, 4$) again. The y-axis range in these subfigures is set in such a way that the ratio of this range to the range of $\{q_{j,t}\}$ is an increasing function of p-value. This leads to the visual perception of more variable $\{q_{j,t}\}$ in the case of a more significant outcome – a hypothesis test employing the standard deviation of $\{q_{j,t}\}$ as the test statistic is used. The value of the test statistic is presented above each subfigure, the p-value following in parentheses. A test employing the Spearman's rank correlation coefficient between $\{q_{j,t}\}$ and time as the test statistic is also performed. The test statistic is given after the comma above each of the subfigures, the p-value being shown in parentheses.

The subfigures in the third column display $\{q_t^{\alpha}\}$ (see Equation 11) with four patterns of α_j values, namely:

- $\alpha_2 = 1$, $\alpha_1 = -1$ and $\alpha_j = 0$ for j equal neither 2 nor 1,
- $\alpha_3 = 1$, $\alpha_1 = -1$ and $\alpha_j = 0$ for j equal neither 3 nor 1,
- $\alpha_4 = 1$, $\alpha_1 = -1$ and $\alpha_j = 0$ for j equal neither 4 nor 1,
- $\alpha_3 = 1$, $\alpha_2 = -1$ and $\alpha_j = 0$ for j equal neither 3 nor 2.

Figure 1 Illustration of the hypothesis test for U.S. gross domestic product data. See the text for a detailed description



Source: Own construction using data retrieved from U.S. Bureau of Economic Analysis, FRED (2015)

Hypothesis tests employing the standard deviation of $\{q_t^\alpha\}$ and the Spearman's rank correlation coefficient between $\{q_t^\alpha\}$ and time are performed for each of the four patterns. Similarly to the second column, the test statistics and p-values are presented above the subfigures in the third column.

It follows from all the subfigures of Figure 1 that the rejection or non-rejection of the null hypothesis of stationarity is influenced by the choice of α_j values in $\{q_t^\alpha\}$ (or the level j if $\{q_{j,t}\}$ is used) and the chosen test statistic. All these choices have to be made before the analysis. A discussion of the results corresponding to particular choices is presented in the paragraphs to follow.

For example, $\{q_{1,t}\}$ is almost monotonically decreasing over time, which implies that the variance of changes associated with the shortest scale is almost monotonically decreasing over time too. If we decided a priori to use the Spearman's rank correlation coefficient between $\{q_{1,t}\}$ and time as the test statistic, we would "definitely" reject the null hypothesis of stationarity at the 5% significance level (see the subfigure called "level 1" in the second column). On the other hand, if the standard deviation of $\{q_{1,t}\}$ was used as the test statistic, the p-value of the hypothesis test would be close to 0.05 and the decision about stationarity or non-stationarity at the 5% significance level would not be so clear.

There seems to be a decrease in the variance of changes, not only on scales associated with the first level but also on those associated with the second and third level. In contrast to the first level, the decrease of the variance associated with the second and third level is far from being monotone. Consequently, the tests using Spearman's rank correlation between $\{q_{2,t}\}$ and time and $\{q_{3,\mathrm{t}}\}$ and time, respectively, are not significant at 5% levels, whereas those using the standard deviation of $\{q_{2,t}\}$ and $\{q_{3,\mathrm{t}}\}$, respectively, are significant (see the subfigures called "level 2" and "level 3" in the second column).

Despite the significance of the two lastly mentioned tests, the hypothesis test associated with the subfigure called "level 3 – level 2" in the third column is non-significant. This is due to the fact that the time-varying patterns in $\{q_{2,t}\}$ and $\{q_{3,\mathrm{t}}\}$ are rather "synchronous". On the other hand, the test employing Spearman's rank correlation in the subfigure called "level 4 – level 1" is significant. This suggests that non-stationarity manifests itself in different ways on short (j = 1) and large (j = 4) scales.

The tests provide a decision about the rejection or non-rejection of the null hypothesis. Moreover, visual inspection of the plots similar to those of Figure 1 may supply additional information about the character of non-stationarity and the size and importance of effects.

CONCLUSION

We have introduced a new wavelet-based hypothesis test for second-order stationarity which is based on exploring the variability of the smoothed series of squared MODWT wavelet coefficients. Having noted that there are alternative techniques for smoothing available in the wavelet literature, we decided to use robust filtering and modified cross-validation. Even though cross-validation is widely used and generally recognized as a flexible tool in the statistical learning literature, it has not been employed much in the resources on wavelets, not being applied at all in formal wavelet-based tests for stationarity.

Further, we have proposed several test statistics that explicitly answer important questions on whether the variability (or the close-to-monotone behavior) observed in the smoothed series represents a significant effect, or whether the characteristic of non-stationarity is scale-specific.

We have used bootstrap to approximate the distribution of the test statistic under the null hypothesis. In agreement with the literature on bootstrap, we have preferred flexible, well-established smoothing techniques and appealing test statistic to the analytical tractability of the procedure. Although the test is computationally expensive, it enjoys reasonable size and power properties for lengths of time series typical in economics. We consider the properties of the test for these lengths superior to those of the test proposed by Nason (2013b).

Our test was also used to assess the stationarity of the time series of the first difference of the logarithm of the U.S. gross domestic product. The results suggest that the variance of changes associated with various scales alters over time, making the time series non-stationary. In particular, the variance of changes associated with the shortest scales exhibits a significant close-to-monotone variation over time. This pattern is not present on larger scales.

ACKNOWLEDGEMENTS

This paper was written with the support of the Czech Science Foundation (GA CR), grant no. P402/12/P507, Modelling of Financial and Economic Time Series – Application and Comparison of Wavelet and Traditional Methods. We acknowledge the linguistic assistance provided by Karel Helman, Sr. (Private University College of Economic Studies Ltd., Prague). We would also like to thank the anonymous reviewers whose comments helped to improve the paper.

References

ARLOT, S., CELISSE, A. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 2010, 4, pp. 40–79.

BENJAMINI, Y., HOCHBERG, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B*, 1995, 57, pp. 289–300.

BERNANKE, B., MISHKIN, F. *Inflation Targeting: A New Framework for Monetary Policy? NBER Working paper No. 5893.* National Bureau of Economic Research, 1997.

BROCKWELL, P., DAVIS, R. *Introduction to Time Series and Forecasting*. 2nd edition, Springer, 2002.

CHU, C.-K., MARRON, J. S. Comparison of Two Bandwidth Selectors with Dependent Errors. *The Annals of Statistics*, 1991, 19, pp. 1906–1918.

CONSTANTINE, W., PERCIVAL, D. *wmtsa: Wavelet Methods for Time Series Analysis. R package version 2.0-0*, 2013.

DAVISON, A. C., HINKLEY, D. V. *Bootstrap Methods and Their Applications.* 11th printing, Cambridge University Press, 2009.

DICKEY, D. A., FULLER, W. A. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, 1979, 74, pp. 427–431.

EFRON, B., TIBSHIRANI, R. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1994.

FARAWAY, J. On the Cost of Data Analysis. *Journal of Computational and Graphical Statistics*, 1992, 1, pp. 213–229.

FELDSTEIN, M., STOCK, J. The use of a monetary aggregate to target nominal GDP. In *Monetary policy*, The University of Chicago Press, 1994, pp. 7–69.

FRIED, R., EINBECK, J., GATHER, U. Weighted Repeated Median Smoothing And Filtering. *Journal of the American Statistical Association*, 2007, 102, pp. 1300–1308.

FRYZLEWICZ, P. Modelling and Forecasting Financial Log-returns as Locally Stationary Wavelet Processes. *Journal of Applied Statistics*, 2005, 32, pp. 503–528.

GRINSTED, A., MOORE, J. C, JEVREJEVA, S. Application of the Cross-Wavelet Transform and Wavelet Coherence to Geophysical Time Series. *Nonlinear Processes in Geophysics*. 2004, 11, pp. 561–566.

HÄRDLE, W., GASSER, T. Robust Non-parametric Function Fitting. *Journal of the Royal Statistical Society. Series B*, 1984, 46, pp. 42–51.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, 2011.

HYNDMAN, R. J. *forecast: Forecasting Functions For Time Series and Linear Models. R package version 5.8.*, 2015.

JENSEN, M., WHITCHER, B. Measuring the Impact Intradaily Events Have On The Persistent Nature of Volatility. In GALLEGATI, M., SEMMLER, W., eds. *Wavelet Applications in Economics and Finance*. Springer, 2014.

MORELL, O., OTTO, D., FRIED, R. On Robust Cross-validation for Non-parametric Smoothing. *Computational Statistics*, 2013, 28, pp. 1617–1637.

NASON, G. *locits: Tests of Stationarity and Localized Autocovariance. R package version 1.4*, 2013a.

NASON, G. A Test For Second-Order Stationarity and Approximate Confidence Intervals for Localized Autocovariances for Locally Stationary Time Series. *Journal of the Royal Statistical Society, Series B*, 2013b, 75, pp. 879–904.

NASON, G., VON SACHS, R., KROISANDT, G. Wavelet Processes and Adaptive Estimation of the Evolutionary Wavelet Spectrum. *Journal of the Royal Statistical Society, Series B*, 2000, 62, pp. 271–292.

PERCIVAL, D. B., WALDEN, A. T. *Wavelet Methods for Time Series Analysis (reprint)*. Cambridge University Press, 2002.

PHILLIPS, P. C. B., PERRON, P. Testing for a unit root in time series regression. *Biometrika*, 1988, 75, pp. 335–346.

R CORE TEAM R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

RAWLINGS, J., PANTULA, S., DICKEY, D. *Applied Regression Analysis: A Research Tool.* 2nd edition, Springer, 2001.

SAID, S. E., DICKEY, D. A. Hypothesis testing in ARIMA(p,1,q) models. *Journal of the American Statistical Association*, 1985, 80, pp. 369–374.

SAID, S. E., DICKEY, D. A. Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 1984, 71, pp. 599–607.

TORRENCE, C., COMPO, G. A Practical Guide to Wavelet Analysis. *Bulletin of the American Meteorological Society*, 1998, 79, pp. 61–78.

U.S. Bureau of Economic Analysis, *Gross Domestic Product* (GDPA) [online]. Retrieved from FRED, Federal Reserve Bank of St. Louis. [10.8.2015]. <https://research.stlouisfed.org/fred2/series/GDPA>.

VON SACHS, R., NEUMANN, M. H. A Wavelet-Based Test for Stationarity. *Journal of Time Series Analysis*, 2000, 21, pp. 597–613.

WHITCHER, B., GUTTORP, P., PERCIVAL, D. Wavelet Analysis of Covariance with Application to Atmospheric Time Series. *Journal of Geophysical Research: Atmospheres*, 2000, 105, pp. 14941–14962.

WICKERHAUSER, M. *Adapted Wavelet Analysis from Theory to Software,* Wellesley, MA: A. K. Peters, 1994.

Segmented Regression Based on B-Splines with Solved Examples

Miloš Kaňka[1] | *University of Economics, Prague, Czech Republic*

Abstract

The subject of the paper is segmented linear, quadratic, and cubic regression based on B-spline basis functions. In this article we expose the formulas for the computation of B-splines of order one, two, and three that is needed to construct linear, quadratic, and cubic regression. We list some interesting properties of these functions. For a clearer understanding we give the solutions of a couple of elementary exercises regarding these functions.

Keywords

Normalized B-spline basis functions, explicit expression, system of normal equations, Weibull plot, Bairstow's iteration method

JEL code

C63, C65

INTRODUCTION

The introduction of the paper, that constitutes its first part, is dedicated to the basic notation of B-spline functions can be found in detail in the existing literature on splines in general (see e.g. Bézier, 1972; Böhmer, 1974; Meloun, Militký, 1994; Spät, 1996). The main content of the paper lies in the aforementioned segmented regression, the theoretical background of which is given in Section 2. Here the most important part is the least squares method that leads to a system of (so-called normal) equations to compute the estimates for the parameters of the chosen regression model.

In Section 3 we describe the so-called polygonal method of value assignment of the parametric variable t (usually time) to experimentally obtained points in $\mathbb{R}^2$ or $\mathbb{R}^3$. The starting point of this method is an oriented graph with vertices given by the experimentally obtained points with the corresponding oriented edges. We associate to the graph vertices, as the value of the parameter t, the length of the polygonal trail that has its starting point in the first vertex of the graph and end point in that particular vertex. The computation of the so-called knots on the axis of the parametric variable that separate the set of experimentally obtained points into line segments (groups, sections) is automatically provided in this method.

In Section 4 we address the question of the transformation of the parametric variable into a unit-length interval, the purpose of which is to increase the numerical stability of the equations of the resulting regression curves.

In Section 5 we solve two given problems. In Example 5.2 we discuss also the notion of so-called optimal regression with respect to the coefficients of determination.

[1] Ekonomická 957, 148 00 Prague, Czech Republic. E-mail: kanka@vse.cz.

1 B-SPLINE FUNCTIONS

There exists a large literature on B-splines, see e.g. (Bézier, 1972), however, let us fix the basic ideas about these functions that we will prefer.

By the symbol $(t)_+$ we denote the real-valued function

$$(t)_+ = \begin{cases} t, & \text{if } t > 0, \\ 0, & \text{if } t > 0. \end{cases}$$

A B-spline function $B_{Q,r} = B_{Q,r}(t)$ is defined for $Q \geq 1$ an integer, r an integer, and $Q + 2$ knots $T_{r-Q-1} < T_{r-Q} < \ldots < T_{\mathrm{r}}$ as a normalized $(Q+1)$-th divided difference of the function $g(T) = [(T - t)_+)]^Q$ of real variable T. Thus, $g(T)$ is, for a given T, function of the real variable t, which we will denote as $(T - t)_+{}^Q$. Hence,

$$B_{Q,r} = (T_r - T_{r-Q-1})\,[T_{r-Q-1}, T_{r-Q}, \ldots, T_r]g\,. \tag{1.1}$$

The first divided difference of g is defined as

$$[T_{r-1}, T_r]\,g = \frac{g(T_r) - g(T_{r-1})}{T_r - T_{r-1}} = \frac{g(T_{r-1}) - g(T_r)}{T_{r-1} - T_r} = [T_r, T_{r-1}]g\,,$$

while the second and the third are

$$[T_{r-2}, T_{r-1}, T_r]g = \frac{[T_r, T_{r-1}]g - [T_{r-1}, T_{r-2}]g}{T_r - T_{r-2}}\,,$$

$$[T_{r-3}, T_{r-2}, T_{r-1}, T_r]g = \frac{[T_r, T_{r-1}, T_{r-2}]g - [T_{r-1}, T_{r-2}, T_{r-3}]g}{T_r - T_{r-3}}\,,$$

etc. Normalization of a divided difference lies in its multiplication with the corresponding denominator. More on divided differences can be found e.g. in (Schrutka, 1945). For example, for $Q = 1$ we have

$$\begin{aligned} B_{1,r} &= (T_r - T_{r-2})\cdot[T_{r-2},\ T_{r-1},\ T_r]g = (T_r - T_{r-2})\,\frac{[T_r,\ T_{r-1}]g - [T_{r-1},\ T_{r-2}]g}{T_r - T_{r-2}} = \\ &= [T_r,\ T_{r-1}]g - [T_{r-1},\ T_{r-2}]g = \frac{g(T_r) - g(T_{r-1})}{T_r - T_{r-1}} + \frac{g(T_{r-2}) - g(T_{r-1})}{T_{r-1} - T_{r-2}} = \\ &= \frac{(T_r - t)_+^1 - (T_{r-1} - t)_+^1}{T_r - T_{r-1}} + \frac{(T_{r-2} - t)_+^1 - (T_{r-1} - t)_+^1}{T_{r-1} - T_{r-2}}\,, \end{aligned}$$

that is

$$B_{1,r}(t) = \frac{T_r - t}{T_r \text{-}\ T_{r-1}} \qquad \text{for } T_{r-1} \leq t \leq T_r, \tag{1.2}$$

$$B_{1,r}(t) = \frac{t - T_{r-2}}{T_{r-1} - T_{r-2}} \qquad \text{for } T_{r-2} \le t \le T_{r-1}, \tag{1.3}$$

everywhere else $B_{1,r}(t)$ takes the value zero.

For the practical computation of B-spline functions we advise to use the recursive de Boor formula, see (de Boor, 1972),

$$B_{Q+1,r} = \frac{t - T_{r-Q-2}}{T_{r-1} - T_{r-Q-2}} B_{Q,r-1} + \frac{T_r - t}{T_r - T_{r-Q-1}} B_{Q,r}\,. \tag{1.4}$$

Example 1.1. According to (1.4), there is for $Q = 1$

$$B_{2,r}(t) = \frac{t - T_{r-3}}{T_{r-1} - T_{r-3}} B_{1,r-1} + \frac{T_r - t}{T_r - T_{r-2}} B_{1,r}\,. \tag{1.5}$$

We shall find the explicit expression of $B_{2,r}$ in $\langle T_{r-1}, T_r\rangle$, $\langle T_{r-2}, T_{r-1}\rangle$, $\langle T_{r-3}, T_{r-2}\rangle$.

For $t \in \langle T_{r-1}, T_r\rangle$, according to (1.2) there is

$$B_{1,r}(t) = \frac{T_r - t}{T_r - T_{r-1}},$$

while $B_{1,r-1}(t) = 0$ (because $B_{1,r-1}$ is non-zero only in (T_{r-3}, T_{r-2})). Therefore, according to (1.5)

$$B_{2,r}(t) = \frac{(T_r - t)^2}{(T_r - T_{r-1})(T_r - T_{r-2})} \qquad \text{for } T_{r-1} \le t \le T_r. \tag{1.6}$$

For $t \in \langle T_{r-2}, T_{r-1}\rangle$, according to (1.3) there is

$$B_{1,r}(t) = \frac{t - T_{r-2}}{T_{r-1} - T_{r-2}} \quad \text{and} \quad B_{1,r-1}(t) = \frac{T_{r-1} - t}{T_{r-1} - T_{r-2}}$$

(in (1.2) we replaced r by $r - 1$). According to (1.5) there is then

$$B_{2,r}(t) = \frac{(t - T_{r-3})(T_{r-1} - t)}{(T_{r-1} - T_{r-3})(T_{r-1} - T_{r-2})} + \frac{(T_r - t)(t - T_{r-2})}{(T_r - T_{r-2})(T_{r-1} - T_{r-2})} \tag{1.7}$$

for $T_{r-2} \le t \le T_{r-1}$.

For $t \in \langle T_{r-3}, T_{r-2}\rangle$, there is $B_{1,r}(t) = 0$ and

$$B_{1,r-1}(t) = \frac{(t - T_{r-3})}{T_{r-2} - T_{r-3}}$$

(in (1.3) we replaced r by $r-1$). Thus, according to (1.5),

$$B_{2,r}(t) = \frac{(t - T_{r-3})^2}{(T_{r-1} - T_{r-3})(T_{r-2} - T_{r-3})} \tag{1.8}$$

for $T_{r-3} \le t \le T_{r-2}$. Everywhere else is $B_{2,r}(t)$ zero.

Example 1.2. For $Q = 2$, according to (1.4) we have

$$B_{3,r}(t) = \frac{t - T_{r-4}}{T_{r-1} - T_{r-4}} B_{2,r-1} + \frac{T_r - t}{T_r - T_{r-3}} B_{2,r} .$$

Analogously as in Example 1.1 we find that

$$B_{3,r}(t) = \frac{(T_r - t)^3}{(T_r - T_{r-1})(T_r - T_{r-2})(T_r - T_{r-3})}, \tag{1.9}$$

for $T_{r-1} \le t \le T_r$,

$$\begin{aligned} B_{3,r}(t) = {} & \frac{(t - T_{r-4})(T_{r-1} - t)^2}{(T_{r-1} - T_{r-2})(T_{r-1} - T_{r-3})(T_{r-1} - T_{r-4})} + \\ & + \frac{(T_r - t)(t - T_{r-3})(T_{r-1} - t)}{(T_r - T_{r-3})(T_{r-1} - T_{r-2})(T_{r-1} - T_{r-3})} + \\ & + \frac{(T_r - t)^2 (t - T_{r-2})}{(T_r - T_{r-2})(T_r - T_{r-3})(T_{r-1} - T_{r-2})}, \end{aligned} \tag{1.10}$$

for $T_{r-2} \le t \le T_{r-1}$,

$$\begin{aligned} B_{3,r}(t) = {} & \frac{(t - T_{r-4})^2 (T_{r-2} - t)}{(T_{r-1} - T_{r-4})(T_{r-2} - T_{r-4})(T_{r-2} - T_{r-3})} + \\ & + \frac{(t - T_{r-3})(t - T_{r-4})(T_{r-1} - t)}{(T_{r-1} - T_{r-3})(T_{r-1} - T_{r-4})(T_{r-2} - T_{r-3})} + \\ & + \frac{(t - T_{r-3})^2 (T_r - t)}{(T_r - T_{r-3})(T_{r-1} - T_{r-3})(T_{r-2} - T_{r-3})}, \end{aligned} \tag{1.11}$$

for $T_{r-3} \le t \le T_{r-2}$, and lastly

$$B_{3,r}(t) = \frac{(t - T_{r-4})^3}{(T_{r-1} - T_{r-4})(T_{r-2} - T_{r-4})(T_{r-3} - T_{r-4})}, \tag{1.12}$$

for $T_{r-4} \le t \le T_{r-3}$. Everywhere else is $B_{3,r}(t)$ zero.

For $Q \geq 1$ whole and r whole, the functions $B_{Q,r}$ have interesting properties, see for example (Meloun, Militký, 1994):

a) They are positive only in the intervals $T_{r-Q-1} < t < T_r$ and are zero everywhere else.

b) They are normalized, i.e. for $k \geq 1$

$$\sum_{r=1}^{k+Q+1} B_{Q,r}(t) = 1 \tag{1.13}$$

in $\langle T_0,\, T_{k+1}\rangle$; for a complete definition of B-splines in the sum (1.13) we need to set on every side of that interval another Q so-called complementary knots

$$T_{-Q} \leq T_{-Q+1} \leq \ldots \leq T_{-1} \leq T_0, \qquad T_{k+1} \leq T_{k+2} \leq \ldots \leq T_{k+Q+1},$$

in the simplest case they merge with T_0 and T_{k+1}, respectively, on the left or right side, respectively. We call $T_1 < T_2 < ... < T_k$, where $T_0 < T_1$ and $T_k < T_{k+1}$, the main knots.

c) In every interval $\langle T_{s-1}, T_s\rangle$, $s = 1, 2, \ldots, k + 1$, exactly $B_{Q,s}$, $B_{Q,s+1}$, … , $B_{Q,s+Q}$ are non-zero, altogether $Q + 1$ in number.

d) $B_{Q,r}$ is in $\langle T_{r-Q-1},\, T_r\rangle$ polynomial spline of order Q with knots $T_{r-Q-1} < T_{r-Q} < ... < T_r$, i.e., in every closed interval defined by two neighbouring points $B_{Q,r}$ is a polynomial of order Q that belongs to the class C^{Q-1} $(T_{r-Q-1},\, T_r)$.

We show the latter properties on the following examples.

Example 1.3. For $Q = 1$, $k = 2$, let us consider main knots $T_1 = 1$, $T_2 = 3$, complementary knots $T_{-1} = T_0 = -\,3{,}6 = T_3 = T_4$. According to (1.2), (1.3), we easily verify that

$$B_{1,1}(t) = \begin{cases} -\frac{1}{4}(t-1) & \text{for } -3 \leq t \leq 1, \\ \quad 0 & \text{otherwise,} \end{cases} \tag{1.14}$$

$$B_{1,2}(t) = \begin{cases} \frac{1}{4}(t+3) & \text{for } -3 \leq t \leq 1, \\ -\frac{1}{2}(t-3) & \text{for } 1 \leq t \leq 1, \\ \quad 0 & \text{otherwise,} \end{cases}$$

$$B_{1,3}(t) = \begin{cases} \frac{1}{2}(t-1) & \text{for } 1 \leq t \leq 3, \\ -\frac{1}{3}(t-6) & \text{for } 3 \leq t \leq 6, \\ \quad 0 & \text{otherwise,} \end{cases}$$

$$B_{1,4}(t) = \begin{cases} \frac{1}{3}(t-3) & \text{for } 3 \leq t \leq 6, \\ \quad 0 & \text{otherwise.} \end{cases}$$

For example, $B_{1,3}$ is positive only in $(T_{3-1-1},\, T_3) = (T_1,\, T_3) = (1{,}6)$, everywhere else is zero; see a). For $s = 2$, in $\langle T_{s-1}, T_s\rangle = \langle T_1, T_2\rangle = \langle 1{,}3\rangle$ only $B_{1,2}$ and $B_{1,3}$ non zero; see c).

For example, for $t_0 = \frac{7}{2} \in \langle T_2,\, T_3\rangle$, where $\langle T_2,\, T_3\rangle = \langle T_{s-1},\, T_s\rangle$ for $s = 3$, only $B_{1,3}$ and $B_{1,4}$ are non-zero, while

$$B_{1,s}\left(\frac{7}{2}\right) = B_{1,3}\left(\frac{7}{2}\right) = -\frac{1}{3}\left(\frac{7}{2} - 6\right) = \frac{5}{6},$$

$$B_{1,s+1}\left(\frac{7}{2}\right) = B_{1,4}\left(\frac{7}{2}\right) = \frac{1}{3}\left(\frac{7}{2} - 3\right) = \frac{1}{6}.$$

Hence,

$$\sum_{r=1}^{k+Q+1} B_{Q,r}(t_0) = \sum_{r=1}^{4} B_{1,r}\left(\frac{7}{2}\right) = B_{1,3}\left(\frac{7}{2}\right) + B_{1,4}\left(\frac{7}{2}\right) = \frac{5}{6} + \frac{1}{6} = 1;$$

see b). For $r = 3$, the B-spline $B_{1,3}$ is in the interval $\langle T_{r-Q-1}, T_r\rangle = \langle T_1, T_3\rangle = \langle 1,6\rangle$ of class C^{Q-1} $(T_1, \mathrm{T}_3) = C^0$ (T_1, T_3), i.e., continuous in this interval. For example, for $t_0 = 3 \in \langle 1,6\rangle$ we have $B_{1,3}(3-) = 1 = B_{1,3}(3+)$; see d).

Example 1.4. For $Q = 2$, $k = 3$ let us consider main knots $T_1 = 3$, $T_2 = 6$, $T_3 = 9$ together with complementary knots $T_{-2} = T_{-1} = T_0 = 0$ and $12 = T_4 = T_5 = T_6$. According to (1.6), (1.7), (1.8), we easily find that

$$B_{2,1}(t) = \begin{cases} \frac{1}{9}(t^2 - 6t + 9) & \text{for } 0 \le t \le 3, \\ 0 & \text{otherwise,} \end{cases} \tag{1.15}$$

$$B_{2,2}(t) = \begin{cases} -\frac{1}{18}(3t^2 - 12t) & \text{for } 0 \le t \le 3, \\ \frac{1}{18}(t^2 - 12t + 36) & \text{for } 3 \le t \le 6, \\ 0 & \text{otherwise,} \end{cases}$$

$$B_{2,3}(t) = \begin{cases} \frac{1}{18}t^2 & \text{for } 0 \le t \le 3, \\ -\frac{1}{18}(2t^2 - 18t + 27) & \text{for } 3 \le t \le 6, \\ \frac{1}{18}(t^2 - 18t + 81) & \text{for } 6 \le t \le 9, \\ 0 & \text{otherwise,} \end{cases}$$

$$B_{2,4}(t) = \begin{cases} \frac{1}{18}(t^2 - 6t + 9) & \text{for } 3 \le t \le 6, \\ -\frac{1}{18}(2t^2 - 30t + 99) & \text{for } 6 \le t \le 9, \\ \frac{1}{18}(t^2 - 24t + 144) & \text{for } 9 \le t \le 12, \\ 0 & \text{otherwise,} \end{cases}$$

$$B_{2,5}(t) = \begin{cases} \frac{1}{18}(t^2 - 12t + 36) & \text{for } 6 \le t \le 9, \\ -\frac{1}{18}(3t^2 - 60t + 288) & \text{for } 9 \le t \le 12, \\ 0 & \text{otherwise,} \end{cases}$$

$$B_{2,6}(t) = \begin{cases} \frac{1}{9}(t^2 - 18t + 81) & \text{for } 9 \le t \le 12, \\ 0 & \text{otherwise.} \end{cases}$$

For example, for $t_0 = \frac{13}{3} \in \langle T_1, T_2\rangle$, where $\langle T_1, T_2\rangle = \langle T_{s-1}, T_s\rangle$ for $s = 2$, only $B_{2,2}$, $B_{2,3}$ and $B_{2,4}$ are non-zero in this interval, and

$$B_{2,2}\left(\frac{13}{3}\right) = \left(\frac{25}{162}\right), \quad B_{2,3}\left(\frac{13}{3}\right) = \frac{121}{162}, \quad B_{2,4}\left(\frac{13}{3}\right) = \frac{16}{162}.$$

Thus,

$$\sum_{r=1}^{k+Q+1} B_{Q,r}(t_0) = \sum_{r=1}^{6} B_{2,r}\left(\frac{13}{3}\right) = B_{2,2}\left(\frac{13}{3}\right) + B_{2,3}\left(\frac{13}{3}\right) + B_{2,4}\left(\frac{13}{3}\right) = \frac{25}{162} + \frac{121}{162} + \frac{16}{162} = 1;$$

see b).

For $r = 4$, in the interval $\langle T_{r-Q-1}, T_r\rangle = \langle T_1, T_4\rangle = \langle 3, 12\rangle$ the function $B_{2,4}$ belongs to C^1 (T_1, T_4), i.e., it has a continuous derivative in this interval. E.g., at $t_0 = 9 \in \langle 3, 12\rangle$ the left derivative of this function is $-\frac{1}{3}$, while its right derivative is also $-\frac{1}{3}$; see d).

Example 1.5. For $Q = 3$, $k = 4$, let us consider main knots $T_1 = 1$, $T_2 = 2$, $T_3 = 3$, $T_4 = 4$ together with complementary nodes $T_{-3} = T_{-2} = T_{-1} = T_0 = -1$, $6 = T_5 = T_6 = T_7 = T_8$. As before, we get

$$B_{3,1}(t) = \begin{cases} -\frac{1}{8}(t^3 - 3t^2 + 3t - 1) & \text{for } -1 \le t \le 1, \\ 0 & \text{otherwise,} \end{cases} \tag{1.16}$$

$$B_{3,2}(t) = \begin{cases} \frac{1}{72}(19t^3 - 33t^2 - 15t) + 37) & \text{for } -1 \le t \le 1, \\ -\frac{1}{9}(t^3 - 6t + 12t - 8) & \text{for } 1 \le t \le 2, \\ 0 & \text{otherwise,} \end{cases}$$

$$B_{3,3}(t) = \begin{cases} -\frac{1}{72}(13t^3 - 3t^2 + 33t - 23) & \text{for } -1 \le t \le 1, \\ \frac{1}{72}(23t^3 - 111t^2 + 141t - 13) & \text{for } 1 \le t \le 2, \\ -\frac{1}{8}(t^3 - 9t^2 + 27t - 27) & \text{for } 2 \le t \le 3, \\ 0 & \text{otherwise,} \end{cases}$$

$$B_{3,4}(t) = \begin{cases} \frac{1}{24}(t^3 - 3t^2 + 3t + 1) & \text{for } -1 \le t \le 1, \\ -\frac{1}{72}(27t^3 - 99t^2 + 81t - 33) & \text{for } 1 \le t \le 2, \\ \frac{1}{24}(11t^3 - 87t^2 + 213t - 149) & \text{for } 2 \le t \le 3, \\ -\frac{1}{6}(t^3 - 12t^2 + 48t - 64) & \text{for } 3 \le t \le 4, \\ 0 & \text{otherwise.} \end{cases}$$

For capacity reasons, for $B_{3,5}, \dots, B_{3,8}$ we do not provide here their expression.

2 SEGMENTED LINEAR, QUADRATIC, AND CUBIC REGRESSION

Let $n \ge 2$ be an integer. In the Euclidean space $\mathbb{R}^m$ (for integer $m > 1$) let us consider n points $P_i = (x_1{}^{(i)}, \dots, x_m{}^{(i)}) = x_j{}^{(i)}$, $i = 1, \dots, n$ (to save space, here and in what follows j will represent the numbers $1, 2, \dots, m$) where at least two are different, obtained during a specific experiment.

Besides these points, $x_j{}^{(i)}$, $j = 1, \dots, m$, are assumed to be real random variables, consider further knots $T_1 < T_2 < \dots < T_k$, $k \ge 1$ an integer, and $T_0 < T_1$, $T_{k+1} > T_k$ complementary knots. As in Section 1, we call $T_1 < T_2 < \dots < T_k$ main knots.

In the interval $\langle T_{l-1}, T_l\rangle$ for $l = 1, \dots, k+1$ where the variable t changes, let us consider and increasing sequence $t_{l1} < t_{l2} < \dots < t_{l,n(l)}$, $n(l) \ge 1$ an integer, while each its member corresponds to one point $x_j{}^{(\mathrm{w})}$, $w = 1, \dots, n(l)$. It holds that $n = \sum_{l=1}^{k+1} n(l)$. The knots form the interval boundaries, in the union of which we will consider depending on the number $Q = 1, 2, 3$ a real function of variable t

$$g_j(t) = \sum_{r=1}^{k+Q+1} \gamma_j^{(r)} B_{Q,r}(t), \tag{2.1}$$

for real parameters $\gamma_j^{(r)}$ and $B_{Q,r}$ B-splines, $r = 1, \dots, k + Q + 1$; see Section 1. For $Q = 1$ we say that (2.1) is a linear spline in the form of B-splines, for $Q = 2$ quadratic, and for $Q = 3$ cubic spline in the aforementioned form.

We will assume that the model of the monitored process is additive, that is, for all values of j, l, w under consideration, it holds that

$$x_j^{(lw)} = g_j(t_{lw}) + \varepsilon_j^{(lw)},$$

where $\varepsilon_j^{(lw)}$ are independent and identically distributed random variables with constant variance. So the estimates $c_j^{(1)}, c_j^{(2)}, \dots, c_j^{(k+Q+1)}$ of the parameters $\gamma_j^{(1)}, \gamma_j^{(2)}, \dots, \gamma_j^{(k+Q+1)}$ can be obtained by the least squares method:

$$U_j = \sum_{l=1}^{k+1} \sum_{w=1}^{n(l)} [x_j^{(lw)} - g_j(t_{lw})]^2 = \sum_{l=1}^{k+1} \sum_{w=1}^{n(l)} \left[x_j^{(lw)} - \sum_{r=1}^{k+Q+1} \gamma_j^{(r)} B_{Q,r}(t_{lw}) \right]^2 . \tag{2.2}$$

Differentiating (2.2) partially with respect to the parameters, for $1 \le p \le k + Q + 1$, we get

$$\frac{\partial U_j}{\partial \gamma_j^{(p)}} = -2 \sum_{l=1}^{k+1} \sum_{w=1}^{n(l)} [x_j^{(lw)} - g_j(t_{lw})] \cdot B_{Q,p}(t_{lw}). \tag{2.3}$$

It is known from mathematical analysis that the necessary condition for U_j, as a function of the parameters, $c_j^{(1)}, c_j^{(2)}, \dots, c_j^{(k+Q+1)}$, to attain its minimum is given by the system of equations

$$\frac{\partial U_j}{\partial \gamma_j^{(p)}} = 0 \quad \text{for } p = 1, \dots, k + Q + 1.$$

This yields through nullification of (2.3) a system of $k + Q + 1$ linear equations for the estimates $c_j^{(1)}, c_j^{(2)}, \dots, c_j^{(k+Q+1)}$ of the parameters $\gamma_j^{(1)}, \gamma_j^{(2)}, \dots, \gamma_j^{(k+Q+1)}$:

$$\boldsymbol{Mc}_j = \boldsymbol{Z}_j , \tag{2.4}$$

where $M = (m_{pq})_{1 \le p,q \le k+Q+1}$ is a $(k+Q+1) \times (k+Q+1)$ matrix, $Z_j = (z_{pj})_{1 \le p \le k+Q+1}$ and $c_j = (c_j^{(p)})_{1 \le p \le k+Q+1}$ are p-dimensional vectors.

The structure of $\boldsymbol{M}$ and $\boldsymbol{Z}_j$ depends on the type of $g_j(t)$, see (2.1). If we put, for brevity, $N_r = B_{Q,r}$ then after nullification of (2.3) we arrive to the expression of the components of $\boldsymbol{M}$ and $\boldsymbol{Z}_j$:

$$m_{pq} = \sum_{l=1}^{k+1} \sum_{w=1}^{n(l)} N_p(t_{lw}) \cdot N_q(t_{lw}), \tag{2.5}$$

$$z_{pj} = \sum_{l=1}^{k+1} \sum_{w=1}^{n(l)} x_j^{(lw)} N_q(t_{lw}). \tag{2.6}$$

From (2.5) it follows that M is symmetric. For $Q = 1$ it is always tridiagonal, five-diagonal for $Q = 2$, and seven-diagonal for $Q = 3$. Such systems of equations can be solved by a recursive procedure which is stable in the sense of error accumulation, see (Makarov, Chlobystov, 1983); the existence of a main diagonal for $\boldsymbol{M}$ means according to definition that

$$\min_{p} \left\{ |m_{pp}| - \sum_{q \neq p} |m_{pq}| \right\} > 0.$$

After solving the system (2.4), we acquire the sought estimates $c_j^{(1)}, c_j^{(2)}, \dots, c_j^{(k+Q+1)}$ of parameters $\gamma_j^{(1)}, \gamma_j^{(2)}, \dots, \gamma_j^{(k+Q+1)}$ in the linear combination $g_j(t)$, $t \in \langle T_0, T_{k+1} \rangle$, see (2.1). The corresponding regression spline to these estimates (linear for $Q = 1$, quadratic for $Q = 2$, and cubic for $Q = 3$), for $t \in \langle T_0, T_{k+1} \rangle$, admits the equations

$$x_j = G_j(t) = \sum_{r=1}^{k+Q+1} c_j^{(r)} B_{Q,r}(t). \tag{2.7}$$

To summarize (for $j = 1, \dots, m$), these equations represent the parametric expression of a curve in $\mathbb{R}^m = (0; x_1, x_2, \dots, x_m)$, which is the output of the regression model of the monitored process; we will call it, in short, a regression curve (linear for $Q = 1$, quadratic for $Q = 2$, and cubic for $Q = 3$).

Due to the special structure of the matrix of the (normalized) system of equations (2.4), that is three-diagonal for $Q = 1$, five-diagonal for $Q = 2$, seven-diagonal for $Q = 3$, the author of the article decided for segmented regression based on B-spline basis functions; such a matrix, the elements of which are all zero except for the given diagonals, enables for an easier and faster computation of the sought solution.

Example 2.1. Let us assume that there were values of the following two parameters: temperature and air pressure detected during 24 hours, every two hours beginning at 6 am, at a given place. Table 1 states the results of this measurement.

Table 1 Fictitious temperature and pressure measurement over a 24h period that could represent a real-world experiment

Time [*h*]		Temperature	Pressure
real	**fictitious**	**[°C]**	**[hPa]**
6:00	0	15	800
8:00	1	16	850
10:00	2	17	900
12:00	3	22	1 000
14:00	4	28	1 050
16:00	5	26	1 020
18:00	6	20	950
20:00	7	19	900
22:00	8	18	890
24:00	9	16	840
2:00	10	15	820
4:00	11	13	810

Source: Own construction

In $\mathbb{R}^2$ (hence, $j = 1, 2$) we have 12 experimentally obtained points, split in four groups (hence, $k = 3$) by three points (we may call them morning, noon, evening and night group):

$$x_j^{(11)} = (15,\ 800),\quad x_j^{(12)} = (16{,}850),\quad x_j^{(13)} = (17{,}900),$$

$$x_j^{(21)} = (22{,}1000),\quad x_j^{(22)} = (28{,}1050),\quad x_j^{(23)} = (26{,}1020),$$

$$x_j^{(31)} = (20{,}950),\quad x_j^{(32)} = (19{,}900),\quad x_j^{(33)} = (18{,}890),$$

$$x_j^{(41)} = (16{,}840),\quad x_j^{(42)} = (15{,}820),\quad x_j^{(43)} = (13{,}810),$$

to which we assign the following (increasing) time values

$$t_{11} = 0,\quad t_{12} = 1,\quad t_{13} = 2,\quad \text{thus}\quad n(1) = 3,$$

$$t_{21} = 3,\quad t_{22} = 4,\quad t_{23} = 5,\quad \text{thus}\quad n(2) = 3,$$

$$t_{31} = 6,\quad t_{32} = 7,\quad t_{33} = 8,\quad \text{thus}\quad n(3) = 3,$$

$$t_{41} = 9,\quad t_{42} = 10,\quad t_{43} = 11,\quad \text{thus}\quad n(4) = 3,$$

see Table 1. It is the case of three main knots, their values can be $T_1 = 3$, $T_2 = 6$, $T_3 = 9$, together, for example, with additional time moments $T_0 = 0$ and $T_4 = 12$.

For example, for $Q = 2$, the matrix $\boldsymbol{M}$ of the system (2.4) is a 6×6 matrix, note that $k + Q + 1 = 3 + 2 + 1 = 6$. To save space, we neither give its full expression, nor for $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$. This computationally intensive work was conducted by the computer program TRIO, that the author of this article created for the purposes of segmented regression based on B-splines.

Nevertheless, for demonstration purposes, let us compute the element m_{56} of $\boldsymbol{M}$ in accord with (2.5). There will be

$$\begin{aligned} m_{56} &= \sum_{l=1}^{4} \sum_{w=1}^{3} N_5(t_{lw}) \cdot N_6(t_{lw}) = \sum_{s=0}^{11} N_5(s) \cdot N_6(s) \\ &= N_5(9) \cdot N_6(9) + N_5(10) \cdot N_6(10) + N_5(11) \cdot N_6(11), \end{aligned} \tag{2.8}$$

as all the other terms of the sum vanish, see (1.15). According to (1.15), there are

$$N_5(9) \cdot N_6(9) = \frac{1}{2} \cdot 0, \qquad N_5(10) \cdot N_6(10) = \frac{2}{3} \cdot \frac{1}{9}, \qquad N_5(11) \cdot N_6(11) = \frac{1}{2} \cdot \frac{4}{9},$$

hence

$$m_{56} = 0 + \frac{2}{27} + \frac{2}{9} = \frac{2 + 6}{27} = \frac{8}{27} = 0.2963E + 00.$$

Let us also compute the component z_{62} of the vector $\boldsymbol{Z}_2$. According to (2.6), there will be

$$z_{62} = \sum_{l=1}^{4}\sum_{w=1}^{3} x_2^{(lw)} N_6\,(t_{lw}) = x_2^{(41)} N_6\,(9) + x_2^{(42)} N_6\,(10) + x_2^{(43)} N_6\,(11), \tag{2.9}$$

as all the other terms of the sum are zero. Thus,

$$z_{62} = 840 \cdot 0 + 820 \cdot \frac{1}{9} + 810 \cdot \frac{4}{9} = \frac{820 + 3240}{9} = \frac{4060}{9} = 4.5111E + 02.$$

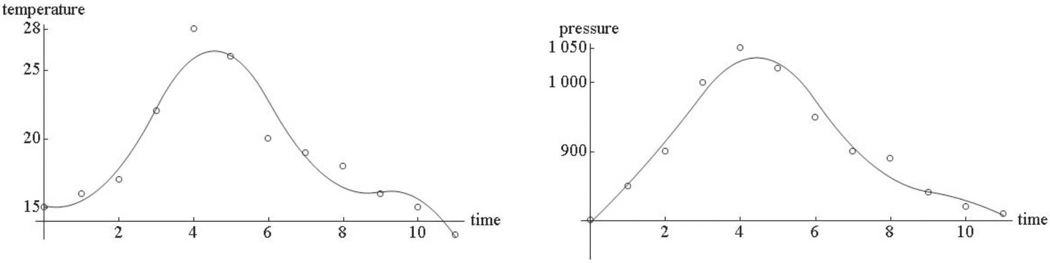
The parametric equations of the regression curve, compare with (2.7), which were obtained with the aid of the computer program TRIO, are the following:

$$x_1 = G_1\,(t) = \begin{cases} 15.1238 - 0.6221t + 09999t^2 & \text{for } 0 \le t < 3, \\ -\,9.5371 + 15.8185t - 1.7402t^2 & \text{for } 3 \le t < 6, \\ 87.3552 - 16.4789t + 0.9521t^2 & \text{for } 6 \le t < 9, \\ -\,81.5195 + 21.0488t - 1.1336t^2 & \text{for } 9 \le t \le 12, \end{cases}$$

and

$$x_2 = G_2\,(t) = \begin{cases} 797.0927 + 52.2798t + 3.3781t^2 & \text{for } 0 \le t < 3, \\ 538.6311 + 224.5875t - 25.3398t^2 & \text{for } 3 \le t < 6, \\ 1876.4612 - 221.3558t + 11.8221t^2 & \text{for } 6 \le t < 9, \\ 574.7062 + 67.9230t - 4.2489t^2 & \text{for } 9 \le t \le 12. \end{cases}$$

Figure 1 B-spline approximation for the temperature and pressure



Source: Own computation

For example, to the value $t = t_{23} = 5$ corresponds on the regression curve (in the plane $(0; x_1\ x_2)$) the point (26.0504,1028.0736), which lies "near" the point (26,1020) of the experiment. Or to $t = 8.5$ corresponds on the regression curve the point (16.0088, 849.0836). We can infer that one hour before midnight the air temperature was approximately 16°C and the air pressure was approximately 850 hPa.

3 THE POLYGONAL METHOD

By polygonal method we shall call in short the following procedure of assigning values of t, our "operating" variable (usually time), to the experimentally obtained points.

In $\mathbb{R}^2$ let us consider the planar connected oriented graph $\vec{G} = [A, \vec{B}]$ with the set of vertices $A = \{1, 2, \dots, n\}$, $n \geq 2$, and $\vec{B} = \{(1, 2),(2, 3), \dots, (n-1, n)\}$, the set of (oriented) edges. We could imagine that the planar polygonal trail obtained in this way, with starting point in 1 and end point in n, is an idealized route of a car moving by constant speed, which started from point 1 and ended the journey at n. Each vertex of the graph $\vec{G}$ can be thought of as trial points, the position of which in the map we find by measuring its distance (for example in km) from the left and bottom edges of the map. We divide the vertices of the graph into $k + 1$ groups, for $k \geq 1$, by $n(l) \geq 1$ points $x_j^{(lw)}$ ($l = 1, \dots, k + 1$; $w = 1, \dots, n(l)$; $j = 1, 2$) in such a way that

$$n = \sum_{l=1}^{k+1} n(l)$$

(this division of the vertices might be caused, e.g., by the difficulty of the corresponding road terrain), and we assign to them an (increasing) sequence of values t_{lw} (in km), where t_{lw} indicates the length of the accomplished route from the start at 1 to the place at $x_j^{(lw)}$, that can be thought of as a resting place during the drive.

We include the values $t_{l1} < t_{l2} < \dots < t_{l,n(l)}$, for $l = 1, \dots, k + 1$, into intervals $\langle T_{l-1}, T_l)$. We further demand that $T_0, T_1, \dots, T_k, T_{k+1}$ is an increasing sequence such that $T_{l-1} \leq t_{l1}$, for $l = 1, \dots, k + 1$ (we shall call $T_1, \dots, T_k$ main knots, while $T_0 < T_1$, $T_{k+1} > T_k$ complementary knots for the observed drive; compare with Section 2).

It is meaningful to set $T_0 = 0$, further, from $T_l \leq t_{l+1,1}$ it follows after substituting for $T_l = \lfloor t_{l,n(l)} \rfloor + p_l \leq t_{l+1,1}$ that $p_l \leq t_{l+1,1} - \lfloor t_{l,n(l)} \rfloor$ (where $\lfloor x \rfloor$ denotes the integer part of the real number x). Let

$$P = \min_{l = 1,\dots,k} \{t_{l+1,1} - \lfloor t_{l,n(l)} \rfloor\} \tag{3.1}$$

and $p = \lfloor P \rfloor$. If $p \geq 1$, then we set $p_l = p$, for $l = 1, \dots, k + 1$; we shall return to the case when $p = 0$.

Putting aside the drive route, we may say that the polygonal method presents a certain automatization in the assignment of operating-variable values to experimental points, divided by a given procedure into groups, that includes the computation of knots defining the range of assigned values to groups (in the aforementioned car drive example the operating variable is the length of the passed track). This polygonal method is implemented in the program TRIO and is capable of solving segmented regression problems in $\mathbb{R}^2$ and $\mathbb{R}^3$, as well.

Example 3.1. In $\mathbb{R}^3$ let us consider the following points $x_j^{(lw)}$ ($l = 1, 2, 3$; $w = 1, \dots, n(l)$, where $n(1) = 2$, $n(2) = 3$, $n(3) = 2$) divided into three groups:

$$\begin{array}{lll}
x_j^{(11)} = (1, 1, 1), & x_j^{(12)} = (1.1, 1.2, 1.3), & \\
x_j^{(21)} = (1.5, 1, 1.4), & x_j^{(22)} = (2, 3, 4), & x_j^{(23)} = (3, 3, 5), \\
x_j^{(31)} = (3.1, 3.2, 5.05), & x_j^{(32)} = (4, 4, 6). &
\end{array}$$

Through the polygonal method we assign (increasing) operating-variable values to them (that can be, for example, time):

$t_{11} = 0.0000,\qquad t_{12} = 0.3742,$

$t_{21} = 0.8325,\qquad t_{22} = 4.1506,\qquad t_{23} = 5.5648,$

$t_{31} = 5.7939,\qquad t_{32} = 7.3277,$

According to (3.1), there is $P = \min\{0.8325, 0.7939\} = 0.7939$, hence $p = \lfloor P \rfloor = 0$.

We will proceed further as follows. We replace the points $x_j^{(lw)}$ considered above by $\tilde{x}_j^{(lw)} = L \cdot x_j^{(lw)}$, where $L > 1$ is a sufficiently large number, and through the polygonal method we assign to them (increasing) operating-variable values $\tilde{t}_{lw} = L \cdot t_{lw}$. For example, for $L = 10$ we obtain now $P = \min\{5.3242, 2.9390\} = 2.9390$, hence $p = \lfloor P \rfloor = 2$. We get the following knots that will we applied to the desired segmented regression (for which the program TRIO is ready): $\tilde{T}_0 = 0$, $\tilde{T}_1 = 5$, $\tilde{T}_2 = 57$, $\tilde{T}_3 = 75$, to which in the initial situation correspond the knots $T_0 = 0$, $T_l = \frac{\tilde{T}_l}{10}$, $l = 1, 2, 3$, i.e. $T_0 = 0$, $T_l = 0.5$, $T_2 = 5.7$, $T_3 = 7.5$.

It is worth to note one more remark. It might happen that the computed knot T_{k+1} will be too far to the right from the length of the entire polygonal trail processed by the computer. The program TRIO enables in this case its reduction to the demanded size.

4 THE TRANSFORMATION OF THE PARAMETRIC (OPERATING) VARIABLE

The elements of $\boldsymbol{M}$ and $\boldsymbol{Z}_j$ in the system of equations (2.4) are structured by the fact that we are working with B-splines. For the improvement of numerical stability of the parametric equations of the regression curve (compare with (2.7)) that is the result of the used regression model, it is recommended in the literature to transform the respective parameter into a unit-length interval (if the length of interval for the initial parameter is much larger than 1; see for example (Meloun, Militký, 1994)). Let us remind that, vaguely speaking, numerical stability of a computational process means “reasonable” or “unreasonable” loss of decimals during the computation.

We transform $t \in \langle T_0, T_{k+1} \rangle$ into

$$t' = M + \frac{M - N}{T_{k+1} - T_0}\,(t - T_{k+1}) = M + K\,(t - T_{k+1}) = f(t), \tag{4.1}$$

where $0 \leq N < M$ are real numbers and $K = \frac{M - N}{T_{k+1} - T_0} > 0$.We can easily see that for any two values $t_1 < t_2$ from this interval it holds that

$$f(t_2) - f(t_1) = K(t_2 - t_1). \tag{4.2}$$

For $l = 1, \dots, k + 1$, the interval $\langle T_{l-1}, T_l \rangle$, where t is changing, transforms onto the interval $\langle T'_{l-1}, T'_l \rangle$, where the variable to change will be t'.

In general, for $Q = 1, 2, 3$ and integer $k \geq 1$, for B-splines $N'_{l+s-1}(t') = B_{Q,l+s-1}(t')$, where $s = 1,\dots,Q + 1$, which are non-zero in $\langle T'_{l-1}, T'_l \rangle$, it holds that

$$N'_{l+s-1}(t') = N_{l+s-1}(t) = N_{l+s-1}(f^{-1}(t')). \tag{4.3}$$

Indeed, for example, for $Q = 3$, $k = 2$, $s = 4$, $l = 3$ we get, according to (1.12) and (4.2), that

$$N_6'(t') = B_{3,6}'(t') = \frac{(t'-T_2')^3}{(T_5'-T_2')(T_4'-T_2')(T_3'-T_2')}$$

$$= \frac{[f(t)-f(T_2)]^3}{[f(T_5)-f(T_2)][f(T_4)-f(T_2)][f(T_3)-f(T_2)]} =$$

$$= \frac{K^3(t-T_2)^3}{K^3(T_5-T_2)(T_4-T_2)(T_3-T_2)} = N_6(t) = N_6(f^{-1}(t')),$$

for $T'_2 \le t' \le T'_3$, that is, for $T_2 \le t = f^{-1}(t') \le T_3$. According to (4.3), the transformation $t' = f(t)$ of the interval $\langle T_1, T_{k+1}\rangle$ onto $\langle N, M\rangle$ does not change the system of normal equations (2.4) (for $j = 1, \dots, m$, $Q = 1, 2, 3$ and integer $k \ge 1$), it provides, therefore, the same estimates $c_j^{(1)}, c_j^{(1)}, \dots, c_j^{(k+Q+1)}$ of the parameters $\gamma_j^{(1)}, \gamma_j^{(2)}, \dots, \gamma_j^{(k+Q+1)}$ in the linear combination of B-splines

$$g_j'(t') = \sum_{r=1}^{k+Q+1} \gamma_j^{(r)} B_{Q,r}'(t'),$$

as in the untransformed case (2.1). The regression spline corresponding to these estimates (linear for $Q = 1$, quadratic for $Q = 2$, cubic for $Q = 3$) admits, for $t' \in \langle T'_0 = N, T'_{k+1} = M\rangle$ the equation

$$x_j = G_j'(t') = \sum_{r=1}^{k+Q+1} c_j^{(r)} B_{Q,r}(f^{-1}(t')). \tag{4.4}$$

To summarize, the equations (4.4) represent, for $j = 1, \dots, m$, the parametric expression of the same regression curve (linear for $Q = 1$, quadratic for $Q = 2$, and cubic for $Q = 3$) as equations (2.7).

5 TWO EXAMPLES

Example 5.1. Let us provide, using a Weibull plot, the failure analysis of lining pads of front disc brakes of cars based on real values observed for cars in Federal Republic of Germany (see (VDA3, 1995)). The goal is to determine the characteristic lifetime defined as the lifetime until which 63.2121% of monitored units is broken ($63.2121 = (1 - e^{-1}) \cdot 100$).

The starting point is Table 2. For the mean order number, median order there are in (VDA3, 1995) available the corresponding formulas. We display the points (t_q km · 1000, H_q%), for $q = 1, \dots, 30$, in a Weibull plot, divided e.g. into three groups of ten (in accord with previous notations, $k = 2$, see Section 2):

$$x_j^{(11)} = (8.8, 1.74), \quad x_j^{(12)} = (10.3, 6.02), \quad \dots, \quad x_j^{(1,10)} = (16.4, 31.54),$$

$$x_j^{(21)} = (17.7, 33.79), \quad x_j^{(22)} = (19.3, 36.13), \quad \dots, \quad x_j^{(2,10)} = (29.9, 59.09),$$

$$x_j^{(31)} = (30.4, 62.03), \quad x_j^{(32)} = (32.1, 64.98), \quad \dots, \quad x_j^{(3,10)} = (55.7, 97.48),$$

And we assign to them values of an (operating) variable t through the polygonal method, see Section 3:

Table 2 Lining pads of front disc brakes of cars in Federal Republic of Germany

Order num. q	Increasing sequence of t_q (km · 10^3)	Num. of broken parts $n_f\,(t_q)$	Num. of good parts $n_s\,(t_q)$	Middle order num. $j(t_q)$	Median order $r(t_q) = H_q$ (%)
1	8.8	2	5	2.10	1.74
2	10.3	4	5	6.53	6.02
3	10.7	3		9.85	9.24
4	11.8	1		10.96	10.31
5	12.9	2	2	13.23	12.50
6	13.4	2		15.50	14.70
7	14.4	4	1	20.09	19.14
8	15.4	4	1	24.75	23.65
9	15.6	2		27.08	25.90
10	16.4	5		32.91	31.54
11	17.7	2		35.24	33.79
12	19.3	2	2	37.65	36.13
13	21.1	6	1	45.03	43.25
14	21.6	2		47.48	45.63
15	22.4	1	1	48.74	46.85
16	23.9	1	4	50.12	48.18
17	25.3	1		51.50	49.52
18	27.7	1		52.88	50.85
19	29.1	1	1	54.30	52.23
20	29.9	5		61.40	59.09
21	30.4	2	2	64.44	62.03
22	32.1	2		67.49	64.98
23	38.4	2	2	70.81	68.19
24	39.7	3		75.78	73.00
25	40.2	3		80.76	77.82
26	40.6	3		85.74	82.63
27	41.8	2		89.06	85.84
28	45.7	2		92.38	89.05
29	50.0	3	1	98.19	94.67
30	55.7	1	1	101.09	97.48

Source: Czech Society for Quality

$$\begin{array}{llll}
t_{11} = 0.0000, & t_{12} = 4.5352, & \ldots, & t_{1,10} = 31.1475, \\
t_{21} = 33.7460, & t_{22} = 36.5807, & \ldots, & t_{2,10} = 63.3745, \\
t_{31} = 66.3567, & t_{32} = 69.7614, & \ldots, & t_{3,10} = 113.3966.
\end{array}$$

For example, $\boldsymbol{t_{32}}$ expresses the length of the polygonal trail measured from the initial point 1 = (8.8, 1.74) to 22 = (32.1, 64.98). In the first group there are, in accord with previous notations, see Section

2, $n(1) = 10$ points, in the second group there are also $n(2)=10$ points, and the same $n(3)=10$ holds for the number of points in the third group. It holds that $\sum_{l=1}^{k+1} n(l) = \sum_{l=1}^{3} n(l) = 10 + 10 + 10 = 30 = n$ (the total amount of observed points).

For t we set the following knots (main and complementary, see Section 1): $T_0 = 0$, further $P = \min\{2.7460, 3.3567\} = 2.7460$, according to (3.1), hence $p = \lfloor P \rfloor = 2$. Therefore, the additional knots are

$$T_1 = \lfloor t_{1,10} \rfloor + 2 = 33,$$

$$T_2 = \lfloor t_{2,10} \rfloor + 2 = 65,$$

$$T_3 = \lfloor t_{3,10} \rfloor + 2 = 115.$$

It holds that $T_0 < T_1 < T_2 < T_3$ and $T_{l-1} \leq t_{l1}$, for $l = 1, 2, 3 = k + 1$.

We choose the transformation of $t \in \langle T_0, T_{k+1}\rangle = \langle T_0, T_3\rangle = \langle 0, 115\rangle$ onto the interval $\langle KT_0, KT_{k+1}\rangle = \langle T'_0, T'_3\rangle = \langle K \cdot 0, K \cdot 115\rangle$ for the factor $K = \frac{1}{T_{k+1} - T_0} = \frac{1}{T_3 - T_0} = \frac{1}{115}$, thus onto the interval $\langle 0, 1\rangle$. The new knots with respect to the new variable t' will then be $T'_0 = 0$, $T'_1 = 0.29$, $T'_2 = 0.57$, $T'_3 = 1$.

The program TRIO is constructed in such a way that it solves the given regression problem for a chosen $Q \in \{1, 2, 3\}$. Thus, for example, $Q = 2$ it presents the following output for the equations of the regression curve

$$x_1 = G'_1(t') = \begin{cases} 9.4844 + 21.5927t' + 30.3873(t')^2 & \text{for } 0 \leq t' < 0.29, \\ 8.4484 + 28.8135t' + 17.8056(t')^2 & \text{for } 0.29 \leq t' < 0.57, \\ 7.7046 + 31.4453t' + 15.4774(t')^2 & \text{for } 0.57 \leq t' \leq 1, \end{cases} \tag{5.1}$$

$$x_2 = G'_2(t') = \begin{cases} 1.4805 + 112.7863t' - 11.9353(t')^2 & \text{for } 0 \leq t' < 0.29, \\ 0.5542 + 119.2418t' - 23.1135(t')^2 & \text{for } 0.29 \leq t' < 0.57, \\ 6.5237 + 98.1193t' - 4.4982(t')^2 & \text{for } 0.57 \leq t' \leq 1. \end{cases} \tag{5.2}$$

It remains to determine an approximate value of the characteristic lifetime with the help of the obtained equations (5.1), (5.2). According to (5.2), there is $G'_2(0.592) = 63.0339$, $G'_2(0.595) = 63.3122$, what in turn means that $x_2 = 63.2121(\%)$ lies between these two values. In the interval (0.592, 0.595) we will search for the solution of the equation

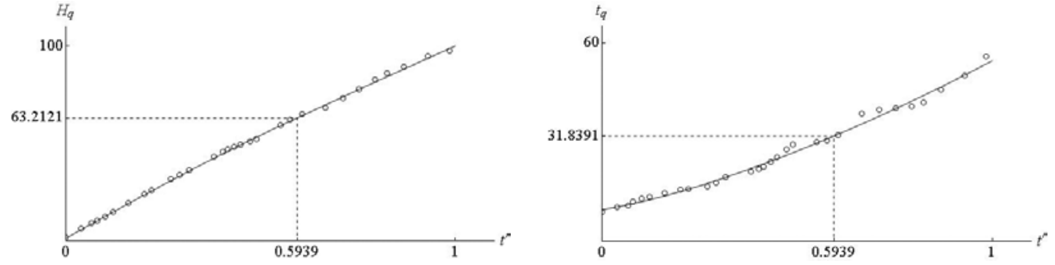
$$63.2121 = 6.5237 + 98.1193t' - 4.4982(t')^2,$$

i.e., after the rearrangement, of the quadratic equation

$$4.4982(t')^2 - 98.1193t' + 56.6884 = 0.$$

The desired solution, gained e.g. by the Bairstow iteration method, with a precision of four decimals is $t' = 0.5939$, after the substitution of which into (5.1), we get that $x_1 = 31.8391$ (km · 1 000). Therefore, the desired characteristic lifetime is $T \doteq 30\,000$ km . Figure 2 depicts the obtained solution.

Figure 2 B-spline approximation of H_q, t_q and solution for specific life expectancy



Source: Own computation

Example 5.2. The following values, in CZK · 10 000, have been obtained from the Czech Statistical Office's table Expenditures of households by level of net money income per person (ZUR0050UU) between 2006 and 2014:

Table 3 Expenditures of households by level of net money income per person 2006–2014

Year		Gross money expenditure x_1	Net money expenditure x_2	Consumption expenditure x_3
real	**fictitious**			
2006	0	11.5839	10.2462	9.4711
2007	1	12.9480	11.5200	10.1399
2008	2	13.3191	11.8367	10.9177
2009	3	13.5882	12.3118	11.2723
2010	4	13.6671	12.3176	11.3464
2011	5	14.3507	13.1116	11.8728
2012	6	14.1125	12.8124	11.8150
2013	7	14.3533	13.0129	12.1921
2014	8	14.7604	13.1873	12.2578

Source: Czech Statistical Office

Hypothetically, let us assume that the data for the year 2009 are missing in Table 3. For further inquiries, however, at least the probable values of x_1, x_2, x_3 are needed for that year. We try to obtain them by regression.

In $\mathbb{R}^3$ we shall, therefore, consider 8 points divided for example into 4 groups (hence $k = 3$) by 2 points:

$$x_j^{(11)} = (11.5839, 10.2462, 9.4711), \qquad x_j^{(12)} = (12.9480, 11.5200, 10.1399),$$

$$x_j^{(21)} = (13.3191, 11.8367, 10.9177), \qquad x_j^{(22)} = (13.6671, 12.3176, 11.3464),$$

$$x_j^{(31)} = (14.3507, 13.1116, 11.8728), \qquad x_j^{(32)} = (14.1125, 12.8124, 11.8150),$$

$$x_j^{(41)} = (14.3533, 13.0129, 12.1921), \qquad x_j^{(42)} = (14.7604, 13.1873, 12.2578),$$

and assign to them (increasing) values of the parametric variable t:

$$
\begin{array}{lll}
t_{11} = 0, & t_{12} = 1, & n(1) = 2, \\
t_{21} = 2, & t_{22} = 4, & n(2) = 2, \\
t_{31} = 5, & t_{32} = 6, & n(3) = 2, \\
t_{41} = 7, & t_{42} = 8, & n(4) = 2 \text{ (see Table 3).}
\end{array}
$$

This is a case with 3 main knots, e.g. $T_1 = 2$, $T_2 = 5$, $T_3 = 7$, together with the complementary knots $T_0 = 0$, $T_4 = 8$.

For $Q = 3 = Q_3$ (cubic regression), the program TRIO provides the equations of the resulting regression curve, also the coefficients of determination $I_{x_1}^2 = 0.9859$, $I_{x_2}^2 = 0.9829$, $I_{x_3}^2 = 0.9894$, according to which 98.59% of the observed values x_1, 98.29% of the observed values x_2, and 98.94% of the observed values x_3 can be explained by this regression model. The program tells us also that the matrix $\boldsymbol{M}$ of the system of equations (2.4) does not have a dominant main diagonal.

The table of the coefficients of determination is the following:

Table 4 Coefficients of determination for linear, quadratic, and cubic regression

Regression	$I_{x_1}^2$	$I_{x_2}^2$	$I_{x_3}^2$
linear	0.9613	0.9605	0.9930
quadratic	0.9816	0.9746	0.9883
cubic	0.9859	0.9829	0.9894

Source: Own construction

It can be seen from Table 4 that the coefficients $I_{x_1}^2$, $I_{x_2}^2$ are maximal for $Q = 3 = Q_3$, while $I_{x_3}^2$ is the highest for $Q = 1 = Q_1$. Having this in mind, one can consider a kind of "optimal" regression curve for the given problem with respect to the coefficients of determination, the construction of which we in turn describe.

Generally, we shall deal with a given problem in $\mathbb{R}^m$, $m > 1$, with observed values x_j, $(j = 1, \dots, m)$ by gradual application of segmented regression for $Q = 1 = Q_1$, $Q = 2 = Q_2$, and $Q = 3 = Q_3$. For a fixed $j \in \{1, \dots, m\}$, let the coefficients of determination $I_{x_j}^2$ attain their highest value for Q_r, $r \in \{1, 2, 3\}$ (the program TRIO chooses r as the lowest possible). If in (2.7) substitute x_j, for that fixed j, with the equation obtained by the particular method Q_r, in the end (for $j = 1, \dots, m$) we can comprehend (2.7) as the parametric expression of the "optimal" regression curve with respect to the coefficients of determination.

In our case, the equation of the "optimal" regression curve is

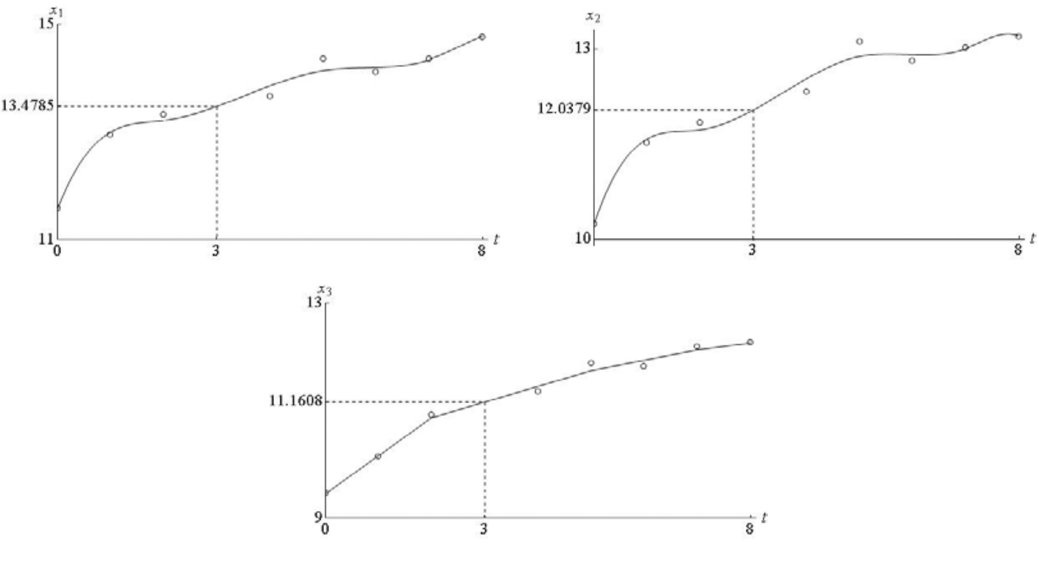
$$
x_1 = \begin{cases}
11.5773 + 2.5421t - 1.3763t^2 + 0.2576t^3 & \text{for } 0 \le t < 2, \\
13.9321 - 0.9900t + 0.3897t^2 - 0.0367t^3 & \text{for } 2 \le t < 5, \\
0.9432 + 6.8034t - 1.1689t^2 + 0.0672t^3 & \text{for } 5 \le t < 7, \\
72.9874 - 24.0727t + 3.2420t^2 - 0.1428t^3 & \text{for } 7 \le t \le 8,
\end{cases}
$$

$$
x_2 = \begin{cases}
10.2390 + 2.4986t - 1.4416t^2 + 0.2819t^3 & \text{for } 0 \le t < 2, \\
12.9217 - 1.5255t + 0.5705t^2 - 0.0534t^3 & \text{for } 2 \le t < 5, \\
-4.7048 + 9.0504t - 1.5447t^2 + 0.0876t^3 & \text{for } 5 \le t < 7, \\
165.5307 - 63.9077t + 8.8779t^2 - 0.4087t^3 & \text{for } 7 \le t \le 8,
\end{cases}
$$

$$x_3 = \begin{cases} 9.4588 + 0.7058t & \text{for } 0 \le t < 2, \\ 10.2893 + 0.2905t & \text{for } 2 \le t < 5, \\ 10.7687 + 0.1947t & \text{for } 5 \le t < 7, \\ 11.2457 + 0.1265t & \text{for } 7 \le t \le 8. \end{cases}$$

For $t = 3$ we get the point (13.4785, 12.0379, 11.1608) on the optimal regression curve from these equations that we can use to substitute the hypothetically missing point in Table 3 for the year 2009, see Figure 3 also. It can be seen that this point obtained through regression lies “nearby” the actual point given in Table 3 for the year 2009.

Figure 3 Optimal regression for household expenditures (cubic for x_1 and x_2, linear for x_3)



Source: Own computation

CONCLUSION

Segmented linear, quadratic, cubic regression can be built also on cut-off splines, see (Meloun, Militký, 1994). We prefer B-splines $B_{Q,r}$, as the matrix of the system of normal equations is three-diagonal (for $Q = 1$), five-diagonal (for $Q = 2$), and seven-diagonal (for $Q = 3$), that is, its structure is much simpler than in the case of cut-off polynomials; such systems can then be solved by fast recursive methods, see (Makarov, Chlobystov, 1983). For the solution of particular exercises (see e.g. Examples 5.1 and 5.2) the computer program TRIO plays an irreplaceable role that handles every procedure leading to the final result, that is, to the equations of the regression curves.

References

BÉZIER, P. *Numerical Control: Mathematics and Applications.* New York: John Wiley & Sons Ltd, 1972.

DE BOOR, C. On calculating with B-splines. *Journal of Approximation Theory*, 1972, 6, pp. 50–62.

BÖHMER, K. *Spline-Funktionen.* Stuttgart: Teubner, 1974.

MAKAROV, V. L., CHLOBYSTOV, V. V. *Splajn-approximacija funkcij.* Moscow: Vysšaja škola, 1983.

MELOUN, M., MILITKÝ, J. *Statistické zpracování experimentálních dat.* Prague: PLUS, s.r.o., 1994.

SEGER, J. *Statistické metody pro ekonomy průmyslu.* Prague: SNTL/ALFA, 1988.

SPÄT, H. *Spline Algorithmen zur Konstruktion glatter Kurven und Flächen.* 4th Ed., München-Wien: Oldenbourg, 1966.

SCHRUTKA, L. *Leitfaden der Interpolation.* Wien: Springer-Verlag, 1945.

VASILENKO, V. A. *Splajn-funkcii: těorija, algoritmy, programmy,* Novosibirsk: Nauka, 1983.

VDA3. *Management jakosti v automobilovém průmyslu.* Prague, Česká společnost pro jakost, 1995. Translation of the German original *Zuverlässigkeitssicherung bei Automobilherstellern und Lieferanten.* Frankfurt am Main: Verband der Automobilindustrie e.V., 1984.

Tracking Users for a Targeted Dissemination

Philippe Bautier[1] | *Eurostat, Luxembourg*
Chris Laevaert | *Eurostat, Luxembourg*
Bernard Le Goff[2] | *Eurostat, Luxembourg*

Abstract

How to build a dissemination and communication strategy in a world where users have easy access to a deluge of data and information from various origins and where IT tools and design standards change so quickly that users behaviour and their expectations are continuously modified? The first challenge of Eurostat is clearly to know what users want: we know our different types of users but we have to identify how they get our data, what they do with our data, how they react to our outputs and which sort of new service they would like us to propose. Translating these needs into a visual dissemination is a new challenge undertaken by Eurostat through a new portal, new mobile apps and new info graphs and basic application as well as increasing the visibility on Google. The objective of this paper is to share Eurostat's experience in identifying user needs and to show how concretely this information has been visually disseminated.

Keywords

Visual dissemination, user behaviour, user needs, user feedback

INTRODUCTION

Today, each national statistical office is confronted with the same challenge: how to build a dissemination and communication strategy in a world where users have easy access to a deluge of data and information from various origins and where IT tools and design standards change so quickly that users behaviour and their expectations are continuously modified?

Eurostat is also facing this challenge. In a document recently adopted by the European Statistical System (ESS),[3] it is said that "the ESS Vision 2020 aims for a future-proof dissemination and communication strategy that satisfies divergent and ever-changing user needs at both national and European level…". The first challenge is clearly to know what users want: we know our different types of users (decision makers, media, researchers, businesses, students, public at large…) but we have to identify how they get our data, what they do with our data, how they react to our outputs and which sort of new services they would like us to propose. In our changing world, this information cannot be obtained only through an annual user survey, but would require continuous and "real time" feedback from our users.

Since a few years, Eurostat has been developing a number of different and complementary tools which give an interesting and up-to date representation of our user needs. The objective of this paper is to share

[1] E-mail: Philippe.Bautier@ec.europa.eu.

[2] E-mail: Bernard.Le-Goff@ec.europa.eu.

[3] See: <http://ec.europa.eu/eurostat/web/european-statistical-system/legislation-in-force>.

Eurostat's experience in identifying user needs and to show how concretely this information has been taken into account and integrated, in particular in the functionalities and services offered by our new website.

1 FINDING THE WAY IN THE LABYRINTH OF USER NEEDS

Since the start of our free dissemination policy in October 2004, the Eurostat website[4] – with all of its associated tools and services – has been identified as the cornerstone of Eurostat's interaction with all kinds of users. It has become the single gateway for Internet users to have on-line access to all Eurostat data and metadata, news releases and publications, or general information about Eurostat. The website is heavily visited. On a monthly basis, the website records more than 3 million visits, over 4 million page views, 700 000 pdf downloads and more than 1 million extractions of data, which ranks the site amongst the top 5 websites of the European Commission. Increasingly, data is being downloaded in bulk, with monthly downloads from the Bulk Download facility reaching 1.2 million files for a volume of 450 Gigabytes.

In order to better understand the needs of our web users, Eurostat has progressively put in place a set of tools. Each of them helps to assemble a more global picture of what modern users expect from suppliers of statistical data.

1.1 Measuring satisfaction

To get an overview of the general level of satisfaction of users, Eurostat conducts an annual on-line user satisfaction survey. This classical method still provides valuable information and feedback on the most consulted statistical domains, the purpose and the frequency of the consultation, as well as an assessment of the quality of our data, publications, and dissemination practices. The 2014 survey had 5 000 replies, the highest response rate in 5 years. Students, academics and private users accounted for the largest proportion of respondents (44%), followed by commercial business (25%) and governments (19%). Replies from international organisations, including EU institutions, and from other users both accounted for more than 5%. As regards the media, a specific survey is also organised every year.

The survey questionnaire has remained similar through the years, allowing for a comparative analysis over time. Overall the results of the survey change only marginally from year to year. Globally, results are positive. Trust remains overwhelmingly positive with 95% of the respondents stating they greatly trust European statistics or tend to trust them. On the dissemination aspects, all user groups are rather satisfied with dissemination practices and support services provided by Eurostat. However, when asked to assess the easiness of access to European statistics, 45% of respondents said it was easy, 40% partly easy and 12% not easy. Improvements are mainly suggested in the area of the search facilities along with the navigation.

1.2 Detecting user behaviour

Website log files provide a wealth of information which is exploited through a detailed and extensive web analytics effort. Each month a 30 page monitoring report on Eurostat electronic dissemination is published on the intranet. Besides figures on the performance and availability of the website, this document compiles all relevant quantitative and qualitative information on what users consult and download; just to name a few: number of consultations for each page, number of publications downloaded and precise timing of the downloads (particularly interesting when you want to monitor the respect of a system of embargo for news releases), navigation and origin of the consultation (Eurostat website, Google, apps,…), average time spent on each visualisation tool, number of consultations of each dataset, etc. This web analytics effort provides a very good picture of what users are interested in and which visualisation tools are used to their full potential. Also, information on usage of Eurostat's mobile apps is available with the number of downloads, giving an indication of the total number of users of such tools, and the number of data updates, providing information on real usage of the mobile app. Although

[4] See: <http://ec.europa.eu/eurostat/web/main>.

it is clear that web analytics is not an absolute science, it allows Eurostat to identify trends and have a more precise view on what and how users consult the on-line statistical information and data.

1.3 Getting feedback in real time

Besides these more traditional methods of measuring user satisfaction and behaviour, it becomes more and more important to measure the impact of dissemination in an on-line world. Indeed, successful dissemination cannot be measured by means of web analytics and usage figures alone, but it needs to take into account new ways of information. For instance, the monitoring of social media brings further insight into who is using our information, how they use it, what they say and think about it and how Eurostat is perceived on the internet in general. Furthermore, statistics are increasingly used by a variety of websites and blogs which target specific peer audience(s). These redistributors serve as a quality vector by adding value to the statistical information supplied by Eurostat. Consumers of such websites will find the relevant statistical information presented in a way which is tailored to their specific needs or context. This enables Eurostat to reach more audiences than it would achieve solely through its own dissemination products.

To measure the impact of dissemination, Eurostat uses a tool to analyse its e-reputation in real time. The tool provides a better knowledge of our users (the ones who are on blogs or social media) and of our impact in the media, and gives a quantitative but also qualitative feedback on our work. In 2014, Eurostat was mentioned nearly 102 000 times (+13% compared to 2013) on the English, French and German speaking web, from 33 500 different identified sources in the media, blogs, forum and social networks. A detailed daily, weekly and monthly analysis of our impact on the web is published internally. In addition, Eurostat disposes of the direct feedback provided by the 58 000 followers of its twitter account. Altogether, this information leads to a much better knowledge of our audience and gives us, in real time, a good idea of our impact on the web.

1.4 Communicating with users

Apart from measuring usage, Eurostat also communicates with users via a permanent user support network, ad-hoc focus groups and benchmarking exercises. For ten years, Eurostat has managed a system of national user support centres[5] offering assistance in nearly all EU languages. Their role is to provide free-of-charge help to users who encounter difficulties in finding or understanding European statistics. In 2014, the whole support network treated more than 15 000 requests. Consumers of statistical information are getting more and more demanding which is confirmed by a clear trend of increasingly complex questions. The valuable feedback collected via this permanent structure enables Eurostat to identify concrete user requirements and helps us to improve the quality of our services.

During the preparation phase of its new website, Eurostat organised ad-hoc focus groups to allow an exchange of views on the current website's strengths and weaknesses. These focus groups were interactive sessions with internal Commission and Eurostat staff, as well as with representative external users (journalists, academics, members of European Parliament, members of European Statistical Advisory Committee). The outcome of the focus groups was integrated in the design and structure of the new website. In particular, more attention has been given to facilitate access to statistical information for non-expert users and to improve the search functionality, in particular by limiting the need to master the statistical jargon.

Furthermore, a more formalised interaction with users is done via the European Statistical Advisory Committee (ESAC)[6] representing users, respondents and other stakeholders of European Statistics (including the scientific community, social partners and civil society) as well as institutional users (e.g. the Council and the European Parliament). The Committee plays an important role in ensuring that

[5] See: <http://ec.europa.eu/eurostat/help/support>.

[6] See: <http://ec.europa.eu/eurostat/web/european-statistical-system/ess-governance-bodies/esac>.

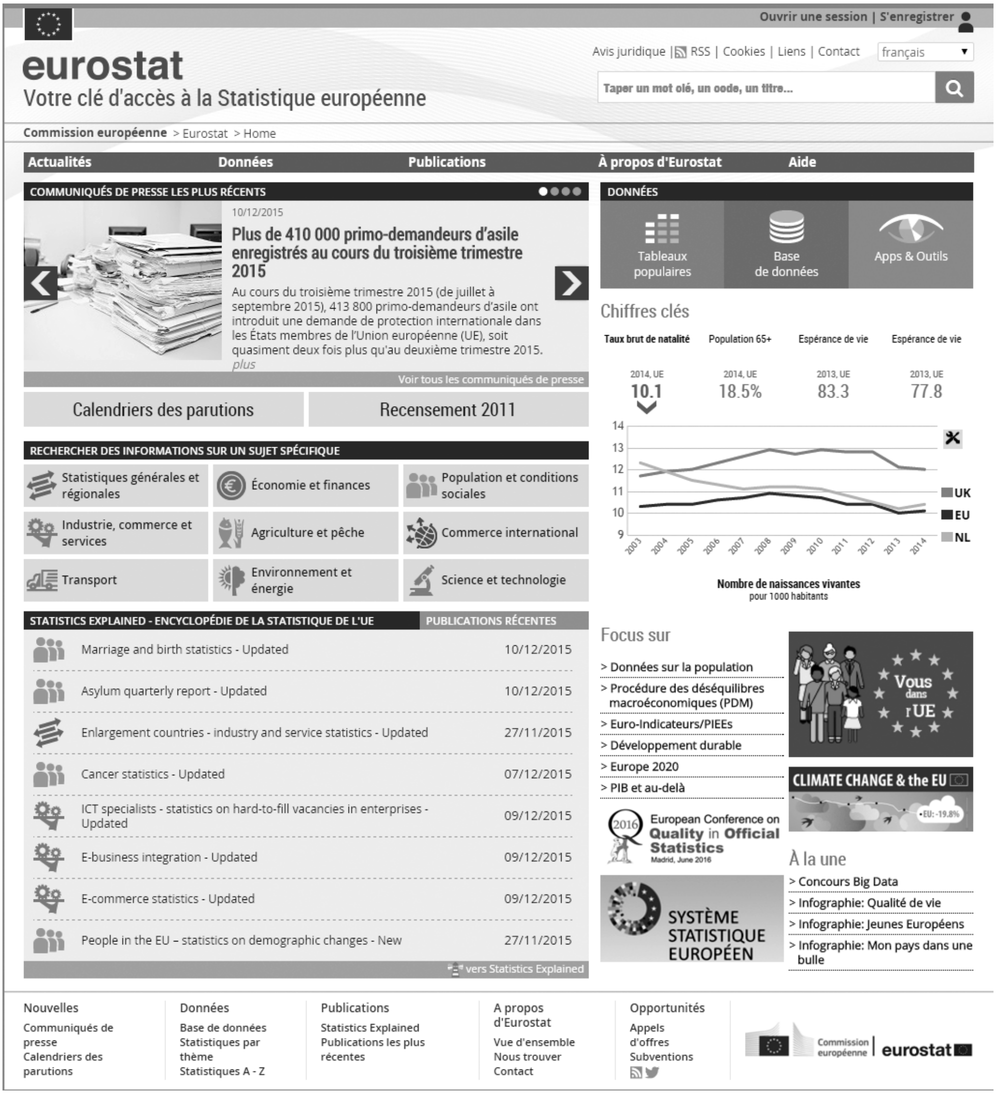
user requirements, as well as the response burden on information providers and producers are taken into account in developing the Statistical Programmes.

As a puzzle, put together, all these different elements provide a relatively good picture of our users and their needs.

2 TRANSLATING USER NEEDS INTO A DISSEMINATION STRATEGY

In order to integrate user needs in Eurostat's dissemination strategy, these needs are translated into concrete objectives and actions and we try to continuously adapt our policy to the new requests of the users. The objective of this paper is not to present a complete list of all actions implemented

Figure 1


Ouvrir une session | S'enregistrer
Avis juridique | RSS | Cookies | Liens | Contact
français
eurostat
Votre clé d'accès à la Statistique européenne
Taper un mot clé, un code, un titre...
Commission européenne > Eurostat > Home
Actualités
Données
Publications
À propos d'Eurostat
Aide
COMMUNIQUÉS DE PRESSE LES PLUS RÉCENTS
DONNÉES
Tableaux populaires
Base de données
Apps & Outils
Chiffres clés
Taux brut de natalité
Population 65+
Espérance de vie
Espérance de vie
2014, UE
10.1
2014, UE
18.5%
2013, UE
83.3
2013, UE
77.8
14
13
12
11
10
9
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
UK
EU
NL
Nombre de naissances vivantes
pour 1000 habitants
Voir tous les communiqués de presse
Calendriers des parutions
Recensement 2011
RECHERCHER DES INFORMATIONS SUR UN SUJET SPÉCIFIQUE
Statistiques générales et régionales
Économie et finances
Population et conditions sociales
Industrie, commerce et services
Agriculture et pêche
Commerce international
Transport
Environnement et énergie
Science et technologie
STATISTICS EXPLAINED - ENCYCLOPÉDIE DE LA STATISTIQUE DE L'UE
PUBLICATIONS RÉCENTES
vers Statistics Explained
Focus sur
Vous dans l'UE
CLIMATE CHANGE & the EU
EU: -19.8%
European Conference on Quality in Official Statistics
Madrid, June 2016
2016
SYSTÈME STATISTIQUE EUROPÉEN
À la une
Commission européenne
eurostat

10/12/2015

Plus de 410 000 primo-demandeurs d'asile enregistrés au cours du troisième trimestre 2015

Au cours du troisième trimestre 2015 (de juillet à septembre 2015), 413 800 primo-demandeurs d'asile ont introduit une demande de protection internationale dans les États membres de l'Union européenne (UE), soit quasiment deux fois plus qu'au deuxième trimestre 2015. *plus*

Marriage and birth statistics - Updated	10/12/2015
Asylum quarterly report - Updated	10/12/2015
Enlargement countries - industry and service statistics - Updated	27/11/2015
Cancer statistics - Updated	07/12/2015
ICT specialists - statistics on hard-to-fill vacancies in enterprises - Updated	09/12/2015
E-business integration - Updated	09/12/2015
E-commerce statistics - Updated	09/12/2015
People in the EU – statistics on demographic changes - New	27/11/2015

> Données sur la population
> Procédure des déséquilibres macroéconomiques (PDM)
> Euro-Indicateurs/PIEEs
> Développement durable
> Europe 2020
> PIB et au-delà

> Concours Big Data
> Infographie: Qualité de vie
> Infographie: Jeunes Européens
> Infographie: Mon pays dans une bulle

Nouvelles	Données	Publications	A propos d'Eurostat	Opportunités
Communiqués de presse Calendriers des parutions	Base de données Statistiques par thème Statistiques A - Z	Statistics Explained Publications les plus récentes	Vue d'ensemble Nous trouver Contact	Appels d'offres Subventions

Source: Eurostat

to be closer to our users but to illustrate our approach through a few examples. Internet is of course at the heart of the efforts made to strengthen our user-orientation.

2.1 A more attractive website, an easier access and a better understanding of European statistics

The launch of the new Eurostat website has been perceived as a good opportunity to better reply to user needs. In the consultation phase of the new website, user's comments often went in the same direction: the most important improvements to the website should focus on its attractiveness and on the access to data, while the information published should be made better understandable. Users also asked for more flexibility in ways to access the data, but did not request important changes to the structure of the website.

Concretely, the main improvements aim to make Eurostat site more attractive and lively, to ease the access and the understanding of our statistics and to offer users a range of visualisation tools.

The previous version of the Eurostat website was created in 2009. Since then, new IT tools, design standards and ways of presenting information on the internet have appeared with, for example, less text and more space for visual information. In consequence, the layout and the design of the web site have undergone a major overhaul to make it more appealing and attractive for both basic and experienced users. This includes, for example, a more colourful design, the possibility to insert photos or videos, and a daily management of the editorial content of the homepage to make it more lively.

Of course, presenting statistical information in a more modern way is not enough. Users, in particular non-specialists, complained about the difficulty to quickly find the information that they were looking for. For that reason, the new website offers several "entries" to ease access to our data, depending on the type of requests or the level of knowledge of users. A quick reply to the simplest requests (on population, GDP, inflation,…) is proposed through our "most popular tables", which include a list of around twenty most downloaded tables. For the more experienced users, a direct access to the full database is proposed where they will find their way to the datasets they need through a simplified navigation tree.

However, the most difficult requests are the ones which are "statistically speaking" less precise and for which users have a more thematic approach. A student, a teacher or a journalist may be interested to know which information is available on women, or on education, climate change, globalisation or tourism. For this type of request, a list of around 60 topics is proposed to users where they can find all datasets and publications relating to their research.

Finally, a new search engine has been developed which provides, on the basis of keywords, the most relevant datasets and articles/publications available, in a similar way to how Google works. To facilitate the search, bridges have been created to enlarge the user request written in current vocabulary (such as profits or family for example) to the associated statistical terminology (gross operating surplus or household).

Data visualisation tools are another possibility to help users better to understand our statistics. Their aim is to communicate clear information or a story through graphs, maps or charts. In recent years, several tools have been implemented by Eurostat, such as country profiles, inflation dashboard, statistical atlas, regional statistics illustrated and widgets. However, the use of these tools requires sometimes the user to have already a good understanding of statistics.

For that reason, Eurostat decided to complement its offer by presenting regular info-graphics on the homepage of its new website, in order to also provide some assistance to less experienced users. For example, new info-graphs are associated with the publication of a selection of euro-indicator news releases, where we try to give to "basic users" a better understanding of the most recent economic trends in the EU, the euro-area and the Member States.

2.2 Simple infographics and visualisation tools

Data visualisation tools are another possibility to help users to better understand our statistics. Their aim is to communicate clear information or a story through graphs, maps or charts. In recent years, several tools have been implemented by Eurostat, such as country profiles, inflation dashboard, statistical atlas, regional statistics illustrated and widgets. However, the use of these tools requires sometimes the user to have already a good understanding of statistics.

For that reason, Eurostat decided to complement its offer by presenting regular infographics on the homepage of its new website, in order to arouse the interest and provide assistance to less experienced users.

"Economic Trends"
A new infograph is associated with the publication of a selection of euro-indicator news releases, where we try to give to non-specialists a better understanding of the most recent economic trends in the EU, the euro-area and the Member States: <http://ec.europa.eu/eurostat/cache/infographs/economy/desktop/index.html>.

"Young Europeans"
In connection with a new Eurostat publication on youth, “Young Europeans” is a new tool which provides the possibility to compare the way of living of a young people aged 15–29 with those of any other young Europeans of the same age and sex. This tool is also intended for parents, decision-makers, politicians or teachers who want to know more about the young generation in Europe.

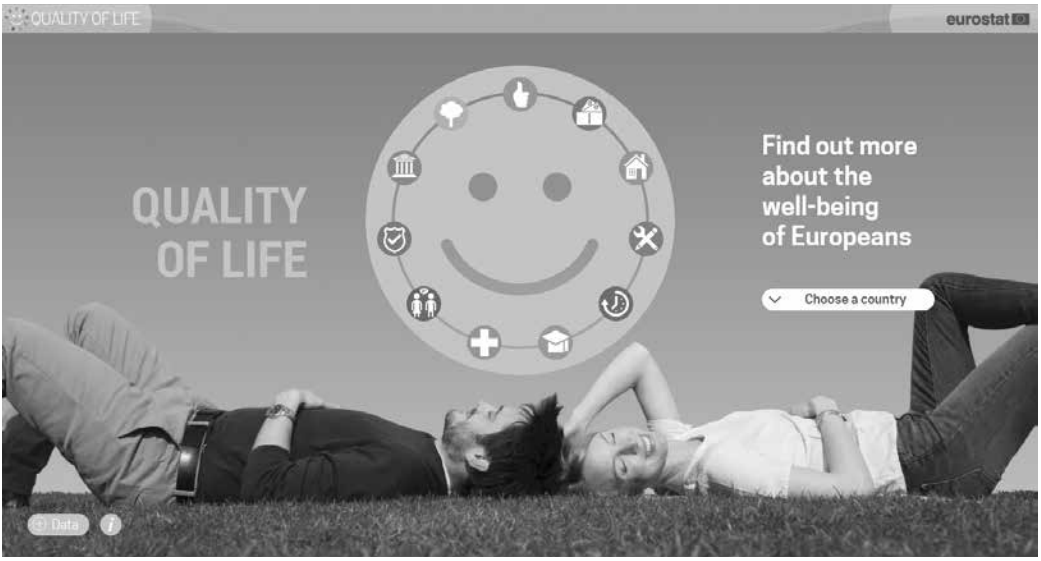
"Young Europeans" consists of quiz like questions about the life of young Europeans on 4 different themes: family, work, free time and studies, and internet. Before starting, users have to define their profile: gender, country and age.

Figure 2

Source: <http://ec.europa.eu/eurostat/cache/infographs/youth/index_en.html>

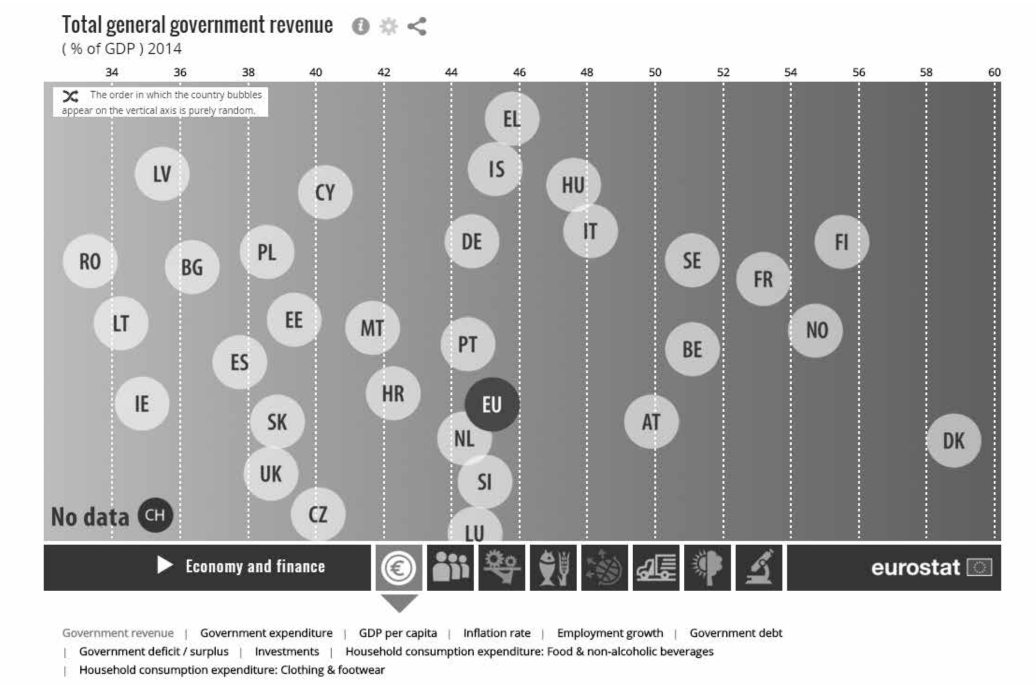
"Quality of life"
Linked to the release of a Eurostat publication on quality of life, this infograph shows both objective and subjective indicators covering 9 themes. It proposes a combination of photos and graphics to display the information in an attractive and innovative way. A new easy recognizable logo for quality of life statistics has also been created.

Figure 3



Source: <http://ec.europa.eu/eurostat/cache/infographs/qol/index_en.html>

Figure 4



Source: <http://ec.europa.eu/eurostat/cache/BubbleChart/?lg=en>

"My country in a bubble"

This simple visualization tool allows users to see in one image the situation in Europe for more than 140 statistical indicators covering all economic, social and environmental domains.

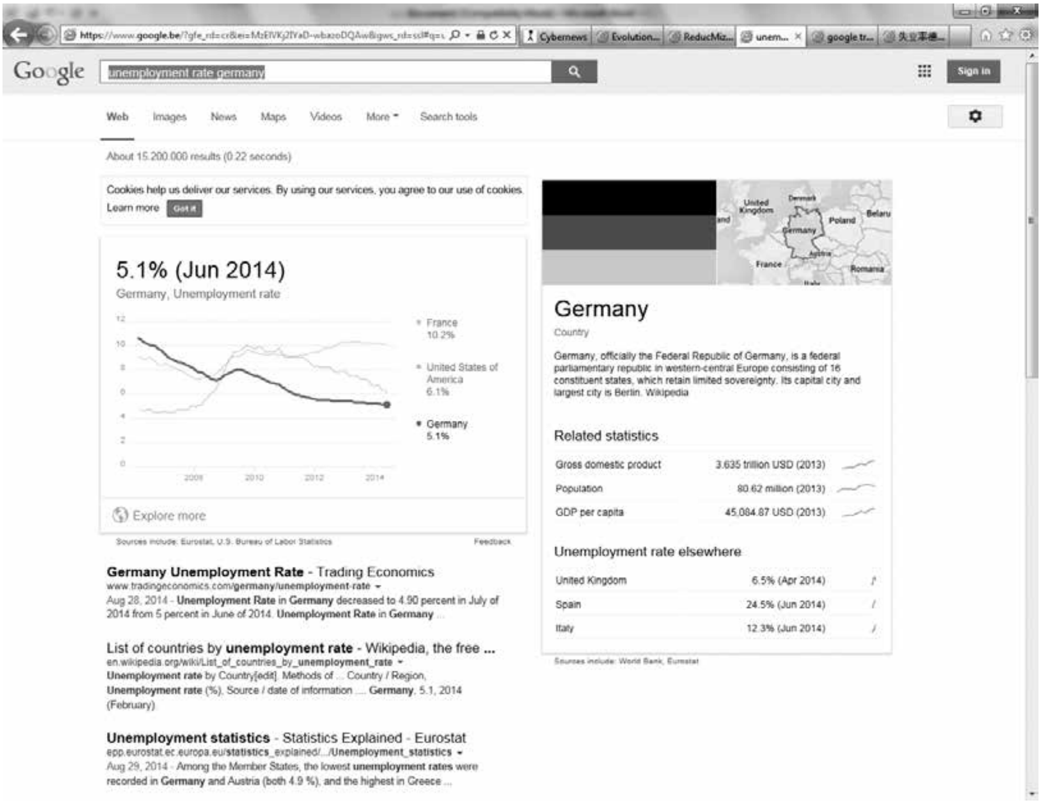
This tool is not really proposed to fix some precise numbers in users' memory but more to give him in one image the perception of the place of his country compared with other EU countries and to encourage him to know more.

2.3 Be present where users are

In the last years, the behavior of internet users has strongly changed. Users are going systematically on Google, Yahoo, etc… Every day the same users are browsing and playing on smartphones and tablets for leisure and/or professional purposes. As a consequence, the market of mobile devices is strongly expanding, and users expect that organisations such as Eurostat offer at least some dedicated information and functionalities for mobile devices. It is then expected that mobile applications (apps) will attract a growing number of users and are therefore increasingly important for Eurostat's image.

In the last years, Eurostat has tried to increase its visibility on Google in different ways. Cooperation with Google started in 2009. In a first step, Eurostat provided a dozen of datasets as well as information about meta information in order to make them directly available via Google search. In doing so, Google translated table titles, definitions, footnotes and labels in 34 different languages. Google also made changes

Figure 5



About 15.200.000 results (0.22 seconds)

Cookies help us deliver our services. By using our services, you agree to our use of cookies. Learn more

5.1% (Jun 2014)

Germany, Unemployment rate

Sources include: Eurostat, U.S. Bureau of Labor Statistics

Germany Unemployment Rate - Trading Economics
www.tradingeconomics.com/**germany/unemployment-rate**
Aug 28, 2014 - **Unemployment Rate** in **Germany** decreased to 4.90 percent in July of 2014 from 5 percent in June of 2014. **Unemployment Rate** in **Germany** ...

List of countries by **unemployment rate** - Wikipedia, the free ...
en.wikipedia.org/wiki/List_of_countries_by_**unemployment_rate**
Unemployment rate by Country[edit]. Methods of ... Country / Region, **Unemployment rate** (%), Source / date of information .... **Germany**, 5.1, 2014 (February)

Unemployment statistics - Statistics Explained - Eurostat
epp.eurostat.ec.europa.eu/**statistics**_explained/.../**Unemployment_statistics**
Aug 29, 2014 - Among the Member States, the lowest **unemployment rates** were recorded in **Germany** and Austria (both 4.9 %), and the highest in Greece ...

Germany

Country

Germany, officially the Federal Republic of Germany, is a federal parliamentary republic in western-central Europe consisting of 16 constituent states, which retain limited sovereignty. Its capital city and largest city is Berlin. Wikipedia

Related statistics

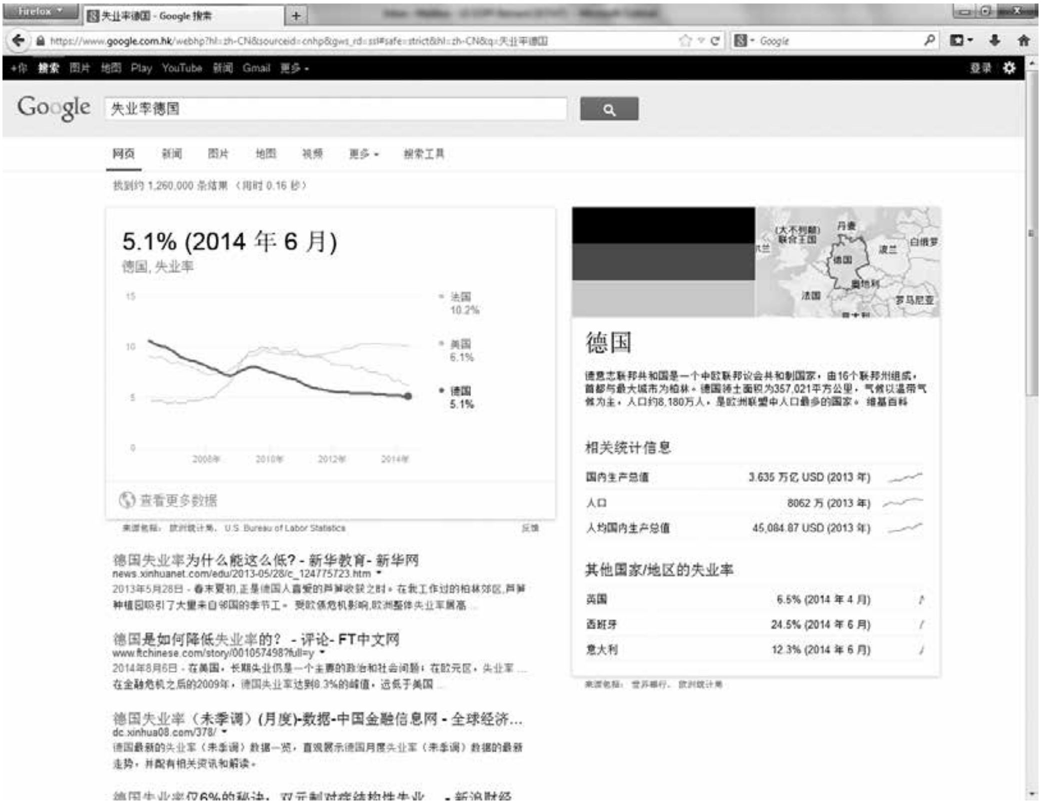
Gross domestic product	3.635 trillion USD (2013)
Population	80.62 million (2013)
GDP per capita	45,084.87 USD (2013)

Unemployment rate elsewhere

United Kingdom	6.5% (Apr 2014)
Spain	24.5% (Jun 2014)
Italy	12.3% (Jun 2014)

Sources include: World Bank, Eurostat

Figure 5 continuation



Firefox
失业率德国 - Google 搜索
https://www.google.com.hk/webhp?hl=zh-CN&sourceid=cnhp&gws_rd=ssl#safe=strict&hl=zh-CN&q=失业率德国
Google
+你 搜索 图片 地图 Play YouTube 新闻 Gmail 更多
登录
Google
失业率德国
网页 新闻 图片 地图 视频 更多 搜索工具

找到约 1,260,000 条结果（用时 0.16 秒）

5.1% (2014 年 6 月)
德国, 失业率

15
10
5
0
2008年
2010年
2012年
2014年
法国 10.2%
美国 6.1%
德国 5.1%

查看更多数据

来源包括：欧洲统计局，U.S. Bureau of Labor Statistics 反馈

德国失业率为什么能这么低? - 新华教育- 新华网
news.xinhuanet.com/edu/2013-05/28/c_124775723.htm
2013年5月28日 - 春末夏初，正是德国人喜爱的芦笋收获之时。在我工作过的柏林郊区，芦笋种植园吸引了大量来自邻国的季节工。受欧债危机影响，欧洲整体失业率居高 ...

德国是如何降低失业率的？ - 评论- FT中文网
www.ftchinese.com/story/001057498?full=y
2014年8月6日 - 在美国，长期失业仍是一个主要的政治和社会问题；在欧元区，失业率 ... 在金融危机之后的2009年，德国失业率达到8.3%的峰值，远低于美国 ...

德国失业率（未季调）(月度)-数据-中国金融信息网 - 全球经济...
dc.xinhua08.com/378/
德国最新的失业率（未季调）数据一览，直观展示德国月度失业率（未季调）数据的最新走势，并配有相关资讯和解读。

德国失业率仅6%的秘诀：双元制对症结构性失业 - 新浪财经

德国

德意志联邦共和国是一个中欧联邦议会共和制国家，由16个联邦州组成，首都与最大城市为柏林。德国领土面积为357,021平方公里，气候以温带气候为主，人口约8,180万人，是欧洲联盟中人口最多的国家。 维基百科

相关统计信息

国内生产总值	3.635 万亿 USD (2013 年)
人口	8062 万 (2013 年)
人均国内生产总值	45,084.87 USD (2013 年)

其他国家/地区的失业率

英国	6.5% (2014 年 4 月)
西班牙	24.5% (2014 年 6 月)
意大利	12.3% (2014 年 6 月)

来源包括：世界银行，欧洲统计局

Source: Eurostat

to its search algorithm to ensure that appropriate searches led directly to these datasets. In a second step, Eurostat worked with Google for the Public Data Explorer. As an example, the results of a Google search on "Unemployment rate in Germany" in English and Chinese show that, in both cases, Eurostat data appear at the first place of Google indexation.

The importance of Google is also noticeable in Statistics Explained, the Eurostat on-line encyclopedia on European statistics and the most consulted collection of Eurostat publications. Here also, important efforts have been invested to obtain a high Google indexation of the articles published in Statistics Explained. We have had very recently the confirmation of the power of Google for our own dissemination when, due to an IT problem at Eurostat, Statistics Explained articles were not indexed on Google during several weeks and the total number of pages viewed on Statistics Explained fell by nearly 70%.

As regards Eurostat's presence on mobile devices, Eurostat has so far released three apps (Country Profiles app at the beginning of 2012, EU economy app at the end of 2013, a quiz on European statistics just released in autumn 2014).

The Country Profiles app shows the latest data for a set of about 160 key indicators. It also allows for displaying the data in the form of dynamic graphs and maps for each indicator. EU Economy app gives mobile access to the most important short-term macroeconomic indicators (Principal European Economic Indicators-PEEIs)[7] for the euro area, the EU and its Member States. The app is available in three languages: English, French and German. It is mainly designed for professionals who need a quick overview on the most recent economic information. The Eurostat Quiz app allows users to test their knowledge about European statistics classified by themes. In answering the questions, users can compete and learn interactively about the European countries. The quiz and the questions are available in 25 languages.

CONCLUSION

All these actions are part of Eurostat's efforts to better respond to user needs but we could have also mentioned a number of other actions, such as the ones more directly related to the content of our publications. Today, statistical institutes are confronted with the same challenge: to continuously adapt their digital and visual dissemination strategy in parallel with the rapid evolution of user needs and IT developments. In a period where human and budgetary resources are limited, this challenge can only be faced if a reinforced cooperation among ESS members is put in place.

References

EUROSTAT. <http://ec.europa.eu/eurostat/web/european-statistical-system/legislation-in-force>.
EUROSTAT. <http://ec.europa.eu/eurostat/web/main>.
EUROSTAT. <http://ec.europa.eu/eurostat/help/support>.
EUROSTAT. <http://ec.europa.eu/eurostat/web/european-statistical-system/ess-governance-bodies/esac>.
EUROSTAT. <http://ec.europa.eu/eurostat/web/euro-indicators/peeis>.

[7] See: <http://ec.europa.eu/eurostat/web/euro-indicators/peeis>.

Recent Publications and Events

New Publications of the Czech Statistical Office

Demographic Yearbook of the Czech Republic 2014. Prague: CZSO, 2015.
External Trade of the Czech Republic in 2014. Prague: CZSO, 2015.
Generation, Recovery and Disposal of Waste for the period 2014. Prague: CZSO, 2015.
Statistical Yearbook of the Czech Republic 2015. Prague: CZSO, 2015.
Vývoj obyvatelstva České republiky (2014). Prague: CZSO, 2015.

Other Selected Publications

Being young in Europe today. Luxembourg: Eurostat, 2015.
CIPRA, T. *Finanční ekonometrie.* 2nd Ed., Prague: Ekopress, 2013.
Competitiveness in the European Economy. New York: Rotledge, 2014.
Development of the basic living standard indicators in the Czech Republic 1993–2014. Prague: MoLSA, 2015.
Eurostat regional yearbook 2015. Luxembourg: Eurostat, 2015.
Financial Production, Flows and Stocks in the System of National Accounts. New York: UN, ECB, 2015.
HEBÁK, P. et al. *Statistické myšlení a nástroje analýzy dat.* 2nd Ed., Prague: Informatorium, 2013.
OECD Labour Force Statistics 2004–2013. Paris: OECD, 2014.
Postavení a vztahy Evropské unie a USA v měnící se globální ekonomice. Brno: Newton College, 2015.
POŠTA, V., MACÁKOVÁ, L., PAVELKA, T. *Strukturální míra nezaměstnanosti v ČR.* Prague: Management Press, 2015.
Quality of life. Facts and views. Luxembourg: Eurostat, 2015.
SOUKUP, J. et al. *Zdroje a perspektivy evropských ekonomik na počátku 21. století v kontextu soudobé globalizace.* Prague: Management Press, 2015.
Sustainable development in the European Union. Luxembourg: Eurostat, 2015.
The role of trade in ending poverty. Geneva: WTO, 2015.
URBAN, J. *Teorie národního hospodářství.* 4th Ed., Prague: Wolters Kluwer, 2015.

Conferences

The European Conference on Quality in Official Statistics (Q2016) will take place in **"Círculo de Bellas Artes" in Madrid, Spain, from 1st to 3rd June 2016.** The conference is organized by The National Statistical Institute of Spain (INE) and Eurostat and aims to cover relevant and innovative topics on quality ranging from the challenges and the new paradigm of quality in an information and knowledge-driven society including big data and multi-source statistics, to governance and management aspects like the ones linked to the ESS Vision 2020 or the lessons learned from 2013–2015 peer reviews in the European Statistical System. More information available at: *http://www.q2016.es*.

The 22nd International Conference on Computational Statistics (COMPSTAT 2016) will take place at the Conference Centre of **Oviedo, Spain, during 23–26 August 2016**. The conference aims at bringing together researchers and practitioners to discuss recent developments in computational methods, methodology for data analysis and applications in statistics. The conference is organized by the University of Oviedo. More information available at: *http://www.compstat2016.org*.

The 19th International Scientific Conference "Applications of Mathematics and Statistics in Economics" (AMSE 2016) will be held in **Banska Stiavnica, Slovakia from 1st to 4th September 2016**. These conferences *"Applications of Mathematics and Statistics in Economics"* are organized each year by three Faculties of three Universities from three countries – University of Economics, Prague (Czech Republic), Matej Bel University in Banska Bystrica (Slovakia) and Wrocław University of Economics (Poland).

Papers
We publish articles focused at theoretical and applied statistics, mathematical and statistical methods, conception of official (state) statistics, statistical education, applied economics and econometrics, economic, social and environmental analyses, economic indicators, social and environmental issues in terms of statistics or economics, and regional development issues.

The journal of *Statistika* has the following sections:
The *Analyses* section publishes high quality, complex, and advanced analyses based on the official statistics data focused on economic, environmental, and social spheres. Papers shall have up to 12,000 words or up to twenty (20) 1.5-spaced pages.

The *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.

In the *Information* section we publish informative (descriptive) texts. The maximum range of information is 6 000 words or up to 10 1.5-spaced pages.

Language
The submission language is English only. Authors are expected to refer to a native language speaker in case they are not sure of language quality of their papers.

Recommended Paper Structure
Title (e.g. On Laconic and Informative Titles) — Authors and Contacts — Abstract (max. 160 words) — Keywords (max. 6 words / phrases) — JEL classification code — Introduction — … — Conclusion — Annex — Acknowledgments — References — Tables and Figures

Authors and Contacts
Rudolf Novak*, Institution Name, Street, City, Country
Jonathan Davis, Institution Name, Street, City, Country
* Corresponding author: e-mail: rudolf.novak@domainname.cz, phone: (+420) 111 222 333

Main Text Format
Times 12 (main text), 1.5 spacing between lines. Page numbers in the lower right-hand corner. *Italics* can be used in the text if necessary. *Do not* use **bold** or underline in the text. Paper parts numbering: 1, 1.1, 1.2, etc.

Headings
1 FIRST-LEVEL HEADING (Times New Roman 12, bold)
1.1 Second-level heading (Times New Roman 12, bold)
1.1.1 Third-level heading (Times New Roman 12, bold italic)

Footnotes
Footnotes should be used sparingly. Do not use endnotes. Do not use footnotes for citing references (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

All views expressed in the journal of Statistika are those of the authors only and do not necessarily represent the views of the Czech Statistical Office, the Editorial Board, the staff, or any associates of the journal of Statistika.

95th 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