HUNGARIAN STATISTICAL REVIEW

JOURNAL OF THE
HUNGARIAN CENTRAL
STATISTICAL OFFICE

VOLUME 90 2012. SPECIAL NUMBER 16

CONTENTS

ISSN 0039 0690

Published by the Hungarian Central Statistical Office
Responsible for publishing: dr. Gabriella Vukovich
Editor in Chief: dr. László Hunyadi
Printed by the Xerox Magyarország Kft.
2012.153 – Budapest, 2012

Managing Editor: Orsolya Dobokay-Szabó
Editor: dr. Cosette Kondora
Technical Editors: Ágnes Simon-Káli

Editorial Office: H-1024 Budapest II., Keleti Károly u. 5–7.
Mailing address: H-1525 Budapest, P.O. Box 51. Phone: +36-(1)345-6546,
Internet: www.ksh.hu/statszemle E-mail: statszemle@ksh.hu
Publishing Office: Central Statistical Office, H-1024 Budapest II., Keleti Károly u. 5–7.
Mailing address: H-1525 Budapest, P.O. Box 51. Phone: +36-(1)345-6000
The publication can be purchased at the Statistical Bookshop:
H-1024 Budapest II., Fényes Elek u. 14–18. Phone: +36-(1)345-6789

Examination of Income Inequalities of Hungarian Households in 2012 Using a Microsimulation Model*

Ilona Cserháti
Head of Department
Office of Public Administration and Justice
ECOSTAT Division for Impact Assessment
E-mail: ilona.cserhati@kih.gov.hu

Tibor Keresztély
Senior Councillor
Office of Public Administration and Justice
ECOSTAT Division for Impact Assessment
E-mail: tibor.keresztely@kih.gov.hu

Tibor Takács
Senior Councillor
Office of Public Administration and Justice
ECOSTAT Division for Impact Assessment
E-mail: tibor.takacs@kih.gov.hu

The paper estimates the distribution of income of the Hungarian households in 2012. Since the first statistical data concerning the incomes of 2012 will be available only in 2014, the figures have been determined by a microsimulation model. It ages the income data of the official Household Budget Survey (HBS) of 2010 published in 2012 by the Hungarian Central Statistical Office. The early information on income distribution among different social strata is useful for both policy makers and for researchers. The HBS is a representative sample of households; therefore their weights have to be adjusted in every step of the microsimulation, assuring the best fit to already known macroeconomic data. This means that the model is dynamic although statically aged. The income inequalities are presented by income deciles, by the number of dependent children, by age structure of households, by activity status and by region and settlement types.

KEYWORDS:
Income inequalities.
Microsimulation.
Household budget survey.

* The authors would like to express their gratitude to *László Mihályffy,* Statistical Key Advisor of the Hungarian Central Statistical Office, who carried out the reweighting and also to *István Molnár*, Professor of Bloomsburg University (USA) for his very useful comments and suggestions. All errors are the authors' alone.

The financial crisis that erupted at the end of 2008 had serious immediate effects on the Hungarian economy. The growth potential was relatively low in that year (about 2 percent) and the main indicators of the economic equilibrium were extremely unfavourable. The debt of the general government had been already continuously increasing for the seven previous years, and by 2008 exceeded 70 percent of the GDP. It was then the fourth highest rate in the EU. In addition, foreign exchange reserves were relatively low, while the deficit of the current account balance exceeded 7 percent of the GDP. Suddenly, the government could not finance itself from the market. The International Monetary Fund (IMF), the World Bank and the EU had to assure the necessary credits, the condition of which was the introduction of fiscal consolidation measures. Although these measures were necessary, they proved to be procyclical, leading to a negative growth of 6.7 percent in 2009. Meanwhile the Hungarian currency depreciated resulting in the decrease of disposable income of the household sector which had a high rate of foreign currency based mortgage loans. All these developments increased the inequalities of the different strata that were analysed in *Cserháti–Takács* [2010] by a microsimulation model. Further measures affecting the income distribution were taken in 2011–2012. A flat tax on the personal incomes has been introduced combined with a family tax benefit. At the same time, tax credit was decreased, and it will disappear starting from 2013. The new tax system may further increase income disparities. This was supported by the results of *Cserháti* and *Takács* [2011], which presented preliminary results for 2011. Especially the reduction of the tax credit may have a negative effect in the lower income deciles. Employees in the government sector received compensation in 2011–2012 in order to prevent the decrease of nominal wages: this compensation had to be provided for 56 percent of the state employees. Meanwhile, on the one hand, several measures have been taken to ease the indebtedness of the household sector. On the other hand, further austerity measures have been introduced to keep the deficit of the general government under 3 percent, since the flat rate could not boost the economy as it was expected. These developments have had also significant effect on the households. The main purpose of the present paper is to assess the distribution of the disposable income among different strata in 2012 using microsimulation.

There are several papers analysing the austerity measures dealing with this problem. *Agnello* and *Sousa* [2011] examine the economic policy of 18 countries in the 1980–2010 period. The conclusion of the paper is that austerity measures increase social inequalities during the time of fiscal consolidation and the inequalities further grow after the adjustment period. *Callan et al*. [2011] examined six countries by mi-

crosimulation, and found that the austerity measures after the financial crisis may have different effects on income distribution, depending on the specific adjustment measures applied in the particular countries. Their research was based on the European Union Statistics on Income and Living Conditions (EU-SILC) database, which contains only income data, that is, they could quantify only the effect of direct tax increases. Authors found that the distribution of burden shows a different picture, but if the effects of indirect taxes are taken into consideration, the adjustment measures put an onus especially on the low-income households. Similar results have been presented in *Callan et al*. [2012], which simulates incomes for different types of families for 2012 by another microsimulation model. *Matsaganis* and *Leventi* [2011] found that the austerity measures introduced recently in Greece have only slightly increased inequalities. *Belyó* and *Molnár* [2006] investigated the opportunities of simulating the capital income for Hungarian data.

The ECOS-TAX to be presented is a microsimulation model, which was basically developed to analyse the impacts of the changes of direct taxes. Its preliminary version was developed in co-operation with the experts of the HCSO (*Cserháti et al*. [2007]). Later, the model was further developed and applied for impact analyses (see for example *Cserháti–Takács* [2010], [2011]).

The structure of the paper is as follows. First, the methodology of research will be outlined in Section 1. The results of the model runs will be presented in Section 2, and finally the conclusions will be drawn in Section 3. The contribution of the paper to current scientific achievements is the quantification of expected income inequalities in Hungary in 2012. Income distribution processes which are one of most hotly debated economic policy issues in Hungary can be analyzed in real time by the model.

1. Methodology

The distribution of household incomes can be based on the information of the Household Budget Survey (HBS) published by the Hungarian Central Statistical Office (HCSO). However, this survey has several disadvantages. On the one hand, the data normally underestimates the incomes, and on the other hand, it does not cover all elements of disposable incomes (that is, only the actually withdrawn interests are reported). Another problem is that the data are published only with more than a year delay. If decision makers wish to get a picture about the actual income distribution, the microsimulation seems to be a good tool. The HCSO and the National Employment Service (NES) provide actual statistics on wages, but it is only a part of dispos-

able incomes and this information is not suitable for analysis of income distribution among different strata. However, with microsimulation, changes can be modelled on the level of the households, starting from the last available HBS data. These data contain very detailed information on the households and on the observed persons, therefore it is possible to determine special response functions for the particular households (practically for different groups of households). Then the actual effects of economic policy measures can be determined for different groups of households. In order to eliminate the former problems of the HBS information, the actual data have to be adjusted and/or completed by using external data sources. The method is suitable for assessing the results of the previously introduced measures by simulating the case of the "no-measures" scenario, and the same is true for the evaluation of planned measures (*Callan et al*. [2011], [2012]).

1.1. Adjustment of weights

The ECOS-TAX is a static ageing dynamic microsimulation model, adjusting the weights of the observations. However, the weights of the households are adjusted in each year of the examination in order to assure the optimal fit to the aggregated data published by the HCSO or by other authorities like NES. Such adjustments are needed also for the starting year of the examination. At present, the HBS data of 2010 are already available. We note that integrated weights are applied to the HBS data, that is, the weights of persons belonging to a certain household equal to the weight of that household. There are certain incomes, which characterize the whole household and there are ones belonging to persons. The wage is a typical example of this latter type: it is followed therefore on the level of persons. On the contrary, the taxable income is determined on the level of the family in order to take the family tax benefit into account. The ECOS-TAX follows the development of wages in a sophisticated way. The original weights of the HBS data were calculated primarily on the basis of demographic data, although some economic conditions were also taken into consideration. In our investigations eight groups of employees have been determined by the appropriate adjustment of weights. Employees in the government sphere, in the corporate sector and in nonprofit institutions are distinguished, and also the fostered workers are treated separately. For each of these four groups, employees covered by the institutional wage statistics of the HCSO are separated from others. The weights have been recalculated in such a way that the number of these groups should optimally fit the actual values of the labour statistics. Also the number of the unemployed is taken into account. The recalculation of weights requires the solution of a nonlinear optimization problem (*Éltető–Mihályffy* [2002]).

1.2. Adjustment of data

Since the income data in the HBS are underestimated, it was necessary to correct the data, in particular the wages, on the basis of an external data source. Three types of data sources are available. Although the personal income tax database contains the data of all employees, it does not contain information on the duration of employment and does not distinguish employees according to institutional sectors. The institutional wage statistics of the HCSO is based on average wages (only one record characterizes every corporation), therefore it is less suitable for the examination of distribution. The wage survey of the NES contains information on full-time and part-time employment and also on the (government or corporate) sector. However, its main drawback is that it gives a picture for only about one month of the year. We have therefore chosen a special two-step method of wage adjustment. In the case of employees covered by the HCSO institutional wage statistics (full-time employees and corporations with at least 5 employees), the adjustment was based on the NES wage tariff data. In this first step it was ensured that the distribution among the deciles corresponds to that of the NES data. In the second step, both the HBS and the personal income tax data were grouped according to regions, age categories and income deciles. Then all the HBS wage data were multiplied by the multiplier of its particular group.

The pension is also underreported in the HBS, although it is a decisive part of the aggregate disposable income of the household sector. We started from the aggregated value of pension in the starting year (in 2010), which is published by the HCSO (STADAT system). The HBS data were adjusted in such a way that the pension per capita value corresponded to the value of the STADAT.

One of the most serious drawbacks of the HBS is that it has very little information on property incomes. Even in the case of interest incomes, only the actually withdrawn interests are reported, which is probably underestimated. Nor does the personal income tax data provide any information about it, since the tax on interest is paid by the financial institution directly. Therefore, we started from the aggregated value of the National Accounts 2010 (D.4. according to ESA95). Since we assumed that interest incomes might vary according to region, this amount has been broken down firstly among regions proportionally to the regional GDP (the STADAT contains information on regional GDP). Secondly, these regional volumes have been further analyzed by deciles. We assumed that there is no interest income in the lowest three deciles at all. If S_i denotes the interest income on the regional level and dec_{ji} is the aggregated net income of decile j in region i ($i = 1,...,7$, $j = 4,...,10$), the following condition is required:

$$S_i = \alpha_i \cdot dec_{4i} + \alpha_i \cdot dec_{5i} + \alpha_i \cdot dec_{6i} + \alpha_i \cdot dec_{7i} + 2 \cdot \alpha_i \cdot dec_{8i} + 2 \cdot \alpha_i \cdot dec_{9i} + \\ + 3 \cdot \alpha_i \cdot dec_{10i} ,$$

where α_i can be determined. This formula assures that the interest income is not assumed to be simply proportional to the overall income; some progressivity is forced by the coefficients assuming that the saving ratio in the richer deciles tends to be higher.

1.3. Ageing of incomes and simulation of taxation

In the case of certain income categories there were unambiguous rules according to which the new values of 2011 and 2012 could be determined, that is, the increase of pensions must equal to the annual inflation rate. In other cases there were macrostatistical data that were taken into consideration. For example, gross earning of a particular person was multiplied by the wage index of his/her corresponding category (government or corporate sector, non-profit institutions, fostered workers). There were incomes that were frozen, for example child-care allowance. Certainly, for certain categories, we could rely only on experts' judgements. Table 1 presents the applied aging rules of the main types of incomes. After aging the gross incomes, the net incomes have been calculated by the actual rules of personal taxations. The application of integrated weights provides an opportunity to determine the family tax benefits. Finally, real incomes are determined by taking into account the actual inflation rate.

Table 1

Aging rules of incomes in the ECOS-TAX model

Incomes	Aging rules
Personal incomes	
Earnings from main activity	
Gross income from main activity	Average increase of wages in the correspondent sector/group
Dismission pay	Average increase of wages with the maxmimum of 3.5 million forints
Income from second job	Average increase of wages
Supplementary compensations	
Voucher for meal	Supposed to be unchanged
Voucher for holiday	100 percent of minimum wage
Voucher for lodgings	Supposed to be unchanged
Voucher for clothing	200 percent of the remuneration base of public officials

(Continued on the next page.)

(Continuation.)

Incomes	Aging rules
Voucher for transport	Experts' judgement
Support for starting school	Changed by the growth rate of minimum wage
Internet access	Supposed to be unchanged
Other in-kind support	Changed by the inflation rate
Company car provided for private use	Supposed to be unchanged
Mobile phone provided for private use	Supposed to be unchanged
Income from self-employment, total	
Annual revenue in the case of simplified corporate tax	Experts' judgement
Income from self-employment	Experts' judgement
Income from corporate enterprises (dividend and wage)	Average increase of wages
Dividend payment from self-employment	Average increase of wages
Other income from work, total	
Income from single commission	Average increase of wages
Income from authorship	Average increase of wages
Income from casual work	Average increase of wages
Tip, gratuity	Changed by the inflation rate
Pensions, supplementary pension	
Pension	Changed by the inflation rate
Pension for disabled persons	Changed by the inflation rate
Pension for widows	Changed by the inflation rate
Supplementary pension for widows	Changed by the inflation rate
Disability benefit	Changed by the growth rate of the minimum pension
Old age benefit	Changed by the growth rate of the minimum pension
Unemployment benefits	
Unemployment allowance	According to the law
Unemployment aid	Changed by the growth rate of minimum wage
Regular social allowances	Changed by the growth rate of the minimum pension
Other support	Changed by the growth rate of the minimum pension

(Continued on the next page.)

(Continuation.)

Incomes	Aging rules
Support related to children	
Child-care fee	Average increase of wages not exceeding the maximum fixed by the law (or changed by the growth rate of minimum wage if there is no gross income from main activitiy)
Child-care allowance	Changed by the growth rate of the minimum pension
Child-care support	Changed by the growth rate of the minimum pension
Pregnancy aid	Average increase of wages
Maternity aid	Changed by the growth rate of the minimum pension
Other social income, total	
Attendance fee	Changed by the growth rate of the minimum pension
Scholarship	Experts' judgement
Regular allowances	Changed by the growth rate of the minimum pension
Non-regular allowances	Changed by the growth rate of the minimum pension
Property income	
Income from leasing of movables and immovables	Experts' judgement
Perpetuity for indemnification	Experts' judgement
Income from abroad	
Wages and salaries from abroad	Average increase of wages
Income from self-employment abroad	Experts' judgement
Social income from abroad	Average increase of pensions
Income from abroad, other	Mean of the average increases of wage/pension
Other	
Transport contribution for disabled persons	Experts' judgement
Sick pay	Average increase of wages not exceeding 200 percent of minimum wage in 2012
Household incomes	
Agricultural income	Experts' judgement
Agricultural expenditures	Experts' judgement
Consumption from own production	Experts' judgement
Family allowance	According to the law
Orphan's allowance	Average increase of pensions not less than the prescribed minimum

(Continued on the next page.)

(Continuation.)

Incomes	Aging rules
Income of children younger than 16 years	Average increase of wages
Support on housing	Experts' judgement not exceeding 30 000 forints/month
Other income, total (income in kind included)	Experts' judgement
Income from interest and dividend payment	Experts' judgement
Reimbursement from insurance companies	Experts' judgement

As a result of the structure of the model, it is not possible to test and to validate the equations using the traditional methodology. In the absence of actual income data, the only option to validate the results is to check the goodness of fit of gross and net earnings for 2011. It was found that the model estimations satisfactorily approximated the actual macroeconomic data regarding gross earnings both for the corporate and the government sectors. Also, the calculated net earnings proved to be plausible, since a remarkable growth rate surplus was experienced, caused by the family tax benefit introduced in 2011. However, it would not have been correct from the methodological point of view to validate the net earnings data, since the official family tax benefit adjusted net earnings calculations were based also on our ECOS-TAX model (*Cserháti–Dobszayné–Takács* [2012]). The complete ex post validation of data will be possible only in 2014 when the final national accounts will be available.

2. Analysis of simulation results

A great number of new laws and changes affected household incomes and the income distribution during the past one and a half years in Hungary. The most important change was obviously the introduction of the flat rate personal income tax combined with the family tax benefit. Meanwhile, the tax credit was removed from the system. This change itself led to the increase of income inequalities (*Cserháti–Takács* [2011]). However, the linear taxes could not boost economic growth in short term, but decreased the revenues of the state budget. The government compensated this loss partly by introducing special taxes for certain sectors and partly by cutting social benefits. It also affected the unemployment benefit; therefore, the government supported the increase of fostered work. We wanted to quantify the result of all these changes in the income inequalities among the different social strata.

Inequality can be characterized by standard indices, like the Lorenz-curve based Gini-index[1] or the income ratio of the highest and lowest quintiles. Eurostat also provides these indices for the EU countries; the last data refer to 2010. It shows that inequality in Hungary was relatively low then. (We note that these indices are calculated on the basis of EU SILC data, which may be biased since households with extremely low income are typically missing.) The purpose of our investigation was to examine how the incomes developed in the last one and a half year.

We analysed the income distribution from different viewpoints. First, all the income deciles were examined, which basically shows the income inequality of the whole society. The number of dependent children was an important view of the analysis because of the introduction of the new family tax benefit. Also, the generation structures of households have been studied. We have calculated the differences according to the activity status of the household head and to the number of economically active family members. Since there are relevant differences among regions, the geographical distribution has also been analysed. Finally, the income distribution according to various types of settlements has been examined. We note that income have been determined for the so-called consumption unit according to the recent concept of the OECD. This means that the weight of the first adult (aged 18 and over) is 1, while that of subsequent adults is 0.5, and 0.3 is given to persons under 18. If the income is determined simply per person, it would unrealistically underestimate the income.

According to the simulation results, the overall increase of disposable net real income was 3 percent in 2011, while it may decrease in 2012 by 2.1 percent as a result of the fiscal consolidation program and the weak growth. (The forecasted inflation rate for 2012 is 5.2 percent.) Table 2 shows the real income distribution per consumption unit among the deciles. The results demonstrate that the income polarization has slightly increased, the loss of the lowest deciles are definitely higher than those of the upper three deciles.

The Gini-index has grown moderately in 2010–2012 from 0.299 to 0.312. This latter value is already higher than the average of the EU27. The ratio of the highest and lowest quintiles (S80/S20) shows a similar picture: it has grown from 4.45 to 4.73. Eurostat reported 5.0 for the EU27 for 2010. *Éltető* [1997] calculated a special index for Hungary (see also *Éltető–Frigyes* [1968]), this is the ratio of the average income of those above the mean to the average income of those below the mean. This index was around 2 in the 1980s, while it is 2.25 in 2012 according to our model. The index is slightly increasing year by year. According to a widely accepted definition, the poor are those with incomes below 60 percent of the median income. After a slight decrease in 2011, the ratio of the poor (poverty headcount

[1] It may vary between 0 and 1, where 0 means perfect equality.

ratio) may reach 13.6 percent in 2012 according to the model results. The income gap ratio – showing the average relative distance from the poverty threshold among the poor – is around 23-24 percent; it has also increased in 2012 according to our computation.

Table 2

Distribution of incomes per consumption unit by decile

Income deciles	Real change 2012/2011 (percent)
Decile 1 (the lowest incomes)	97,1
Decile 2	97,5
Decile 3	97,7
Decile 4	97,1
Decile 5	97,6
Decile 6	97,5
Decile 7	97,9
Decile 8	98,5
Decile 9	98,6
Decile 10 (the highest incomes)	99,0
Total	97,9

Source: Here and hereinafter ECOS-TAX model results.

Table 3

Income polarization per consumption unit by the number of dependent children

Household category according to the number of children	Consumption units	Real change 2012/2011 (percent)
No children	3 773 482	98,0
One child	1 262 260	97,4
Two children	961 773	97,9
Three or more children	496 879	98,3
Total	*6 494 394*	97,9

If the distribution by dependent children is considered, one can see in Table 3 that families with at least three children have the smallest loss. The main cause is obviously the realizable family tax benefit. Although this type of benefit already existed in 2011,

the parents may be able to realize more benefits as a result of the increase of gross salaries (primarily in the corporate sector). Considering the age structure of the families, the dynamics is different compared to our earlier examinations for 2010 (*Cserháti–Takács* [2010]). The relatively small loss of families with only aged members can be explained by the fact that pensions are still being increased in line with inflation. The increase of taxes burdened primarily families with middle-aged members, who cannot realize the tax benefit (see Table 4). This may explain also the figures of Table 5, where households only with inactive members have the smallest loss.

Table 4

Income polarization per consumption unit by the age structure of the household

Age structure of the household members	Consumption units	Real change 2012/2011 (percent)
Only young	235 511	98,3
Only middle-aged	844 046	97,7
Only aged	1 146 256	99,3
Young and middle-aged	3 015 729	97,6
Young and aged	99 728	98,0
Middle-aged and aged	596 640	97,8
Three generations	556 484	97,0
Total	*6 494 394*	97,9

Table 5

Income polarization per consumption unit by activity status of the household head

Activity status	Consumption units	Real change 2012/2011 (percent)
Households only with non-active members	1 959 132	98,6
Wage earner, active household head	3 648 492	97,8
Pensioner, inactive household head	423 004	97,4
Other inactive household head	463 766	96,4
Total	6 494 394	97,9

Regarding the geographic aspects, the capital Budapest has the most favourable position (we have examined it separately from Pest County to which it belongs). This means that the relative position of the traditionally developed parts of the country may increase their advantage, in other words, the regional polarization may become

greater. Similarly, Budapest and the towns increase their advantage if the different types of settlements are compared. The unemployment rate is high in small settlements, which may explain the worsening of their position.

Table 6

Income per household by region

Region	Number of households	Real change 2012/2011 (percent)
Budapest	756 843	99,0
Pest County	424 702	98,3
Central Transdanubia	412 474	98,0
Western Transdanubia	354 799	97,5
Southern Transdanubia	342 844	97,3
Northern Hungary	431 240	97,7
Northern Great Plain	529 747	96,2
Southern Great Plain	523 732	97,5
Total	3 776 381	97,9

Table 7

Income polarization per consumption unit by type of settlements

Type of settlements	Consumption units	Real change 2012/2011 (percent)
Budapest (capital)	1 185 682	99,0
Towns with county rights	1 313 234	98,0
Other towns	1 871 392	97,5
Villages	2 124 086	97,2
Total	6 494 394	97,9

3. Conclusions

This paper has analysed the effects of fiscal consolidation measures on income distribution by the ECOS-TAX microsimulation model. The general experience in the different countries confirms that such austerity measures increase income inequalities,

although the effects of the changes of direct taxes – ceteris paribus – may show a different picture. Our examination was based on the income data of households, so it is suitable primarily for quantifying the effect of direct taxes and that of transformations in social policy. The results suggest that these changes themselves increase income polarization: the Gini-index grew by 1.3 percentage points during the last two years. The basic indices of poverty have not changed significantly in this period, but the head count poverty ratio has remained near 14 percent, according to the median income criterion. The increase of minimum wage and the relevant growth of the number of fostered workers could contribute to the stabilization of these indices. Families of at least three dependent children have a relatively smaller loss as a consequence of the increase in the realizable family tax benefit. Although the fiscal consolidation required the reduction of social benefits, there has been no change in the rule of growing pensions: they increase by the annual inflation. As a result, families with only aged members will have a relatively small loss in 2012. Income polarization can also be examined from a geographical point of view; the advantage of the traditionally developed parts of the country has grown in the examined period. Correspondingly, the position of the capital and of the larger settlements has improved relatively to others.

The model ECOS-TAX proved to be a very useful tool to follow the actual income processes. The timeliness is very important for both fiscal and monetary decision-makers; hence the official data on the income distribution among different social strata are published two years later. The model is also suitable for ex ante policy analyses as well.

References

AGNELLO, L. – SOUSA, R. M. [2011]: *Fiscal Consolidation and Income Inequality*. NIPE Working Paper 34. Universidade do Minho. Braga.

BELYÓ, P. – MOLNÁR, I. [2006]: Use of Microsimulation Models for Political Decision Making. *Public Finance Quarterly*. Vol. LI. No. 3. pp. 353–365.

CALLAN, T. – LEVENTI, C. – LEVY, H. – MATSAGANIS, M. – PAULUS, A. – SUTHERLAND, H. [2011]: *The Distributional Effects of Austerity Measures*: *A Comparison of Six EU Countries*. EUROMOD Working Paper. No. EM6/11. University of Essex. Essex.

CALLAN, T. – KEANE, C. – SAVAGE, M. – WALSH, J. R. [2012]: *Distributional Impact of Tax*, *Welfare and Public Sector Pay Policies*: *2009–2012*. Quarterly Economic Commentary. The Economic and Social Research Institute. Dublin. http://hdl.handle.net/2262/63908

CSERHÁTI, I. – DOBSZAYNÉ HENNEL, J. – HAVASI, É. – KERESZTÉLY, T. – KŐVÁRI, ZS. – SZÉP, K. – TAKÁCS, T. – TALLÉR, A. – TAMÁSI, B. – VARGA, ZS. [2007]: *A háztartások jövedelemalakulásának elemzése mikroszimulációs modellel*. ECOSTAT-KSH. Budapest.

CSERHÁTI, I. – DOBSZAYNÉ HENNEL, J. –TAKÁCS, T. [2012]: Mikroszimuláció alkalmazása a munkaügyi statisztikában. *Statisztikai Szemle*. Vol. 90. No. 9. pp. 844–861.

CSERHÁTI, I. – TAKÁCS, T. [2010]: Analysis of Income Disparities by Microsimulation. *Hungarian Statistical Review*. Special No. 14. pp. 110–124.

CSERHÁTI, I. – TAKÁCS, T. [2011]: Flat Rate Tax in Hungary. *Journal of International Scientific Publication*: *Economy and Business*. Vol. 5. Part 3. http://www.science-journals.eu, pp. 489–497.

ÉLTETŐ, Ö. [1997]: Disparities in the Economic Well-Being of Hungarian Society from the Late 1970s to the 1980s. In: *Gottschalk*, *P*. – *Gustaffson, B*. – *Palmer, E*. (eds.): *Changing Patterns in the Distribution of Economic Welfare*. Cambridge University Press. Cambridge.

ÉLTETŐ, Ö – FRIGYES, E. [1968]: New Income Inequality Measures as Efficient Tools for Causal Analysis and Planning. *Econometrica.* Vol. 36. No. 2. pp. 383-396.

ÉLTETŐ, Ö. – MIHÁLYFFY, L. [2002]: Household Surveys in Hungary. *Statistics in Transition*. Vol. 5. No. 4. pp. 521–540.

MATSAGANIS, M. – LEVENTI, C. [2011]: The Greek Crisis in Focus: Austerity, Recession and Paths to Recovery. In: *Monastiriotis*, *V*. (ed.): *The Greek Crisis in Focus*. Hellenic Observatory Papers on Greece and Southeast Europe. Special Issue. pp. 1–43. The London School of Economics and Political Science. London.

Ethnic Segregation Between Hungarian Schools: Long-run Trends and Geographic Distribution*

Gábor Kertesi
Senior Research Fellow
Institute of Economics of the Hungarian Academy of Sciences, RCERS
E-mail: kertesi@econ.core.hu

Gábor Kézdi
Associate Professor
Central European University, Research Fellow
Institute of Economics of the Hungarian Academy of Sciences, RCERS
E-mail: kezdig@ceu.hu

Using all of the available data on the ethnic composition of Hungarian primary schools, this paper documents the degree of between-school segregation of Roma versus non-Roma students in the 1980–2011 period. We calculate the measures of segregation within school catchment areas as well as within micro-regions and the larger municipalities (towns and cities). Catchment areas are clusters of villages, towns and cities that are closed in terms of student commuting, and they are defined by us using the observed commuting patterns. Our results show that ethnic segregation between Hungarian schools strengthened substantially between 1980 and 2011. Segregation appears to have decreased between 2006 and 2008 and increased again afterwards, but the noise in the data prevents us from drawing firm conclusions. In the cross section, school segregation is positively associated with the size of the educational market and the share of Roma students, similar to the results from US metropolitan areas. These relationships strengthened over time in Hungary, and the change in segregation is associated with changes in the number of schools and the share of Roma students.

KEYWORDS:
School segregation.
Roma minority.

* We thank *Melinda Tir* for her assistance with data management, *Tímea Molnár*, *Péter Dívós* and *Ágnes Szabó-Morvai* for their earlier work on programming and *László Göndör* for his help with the maps. We thank *Gábor Bernáth*, *János Zolnay*, as well as the editor and the referee for their thoughtful comments. All the remaining errors are ours. Individual research grants from the Institute of Economics of the Hungarian Academy of Sciences are gratefully acknowledged.

Over ten percent of the Hungarian students in primary schools are Roma. The typical Roma students come from substantially poorer families and have lower achievement than the typical non-Roma students (*Kertesi–Kézdi* [2011]). The extent to which Roma and non-Roma students study in the same schools can have serious consequences for ethnic differences in accomplishment and other outcomes as well as for the integrity of Hungarian society.

Using all of the available comprehensive data on the ethnic composition of Hungarian primary schools, this paper documents the degree of between-school segregation of Roma versus non-Roma students between 1980 and 2011. We show the long-run trends and the geographic distribution, and we estimate regressions to uncover the associations between segregation and other characteristics of the areas, which are identified from the cross-section and from the long-differenced panel of the areas for which school segregation is defined.

It is necessary to have some institutional knowledge of the Hungarian school system to understand school segregation. We are interested in the primary schools that cover grades 1 through 8 (these include some secondary schools that cover grades 5 through 8). Importantly, and similar to other countries in the region, Hungary is characterized by the dominance of state-owned primary schools, and parents are free to choose schools for their children. On top of the enrolment from within their own district, which is defined by the municipality, schools can admit children living outside of the district. The total enrolment in schools is determined by their capacity, the level of demand from within and from outside of their district and the allocation decision by the municipality.

We estimate the degree of segregation within three types of geographic area: the 174 micro-regions, the larger school catchment areas (clusters of villages, towns and cities that are closed in terms of student commuting in the 2000s and have two schools or more) and the larger municipalities (towns and cities with two or more schools). Our preferred unit of measurement is the catchment area because it represents the territory that is the most relevant for school choice. In a sense, micro-regions are too large: school segregation within micro-regions is likely to be heavily influenced by the residential patterns across towns and villages. The towns and cities are too small: measuring segregation within their administrative boundaries misses potentially important commuting from and to villages in their agglomeration. The school catchment areas are not administratively registered units; they are defined by commuting possibilities. A contribution of our paper is to define the boundaries of those areas using the actual commuting patterns of all sixth graders observed in three different years.

Our preferred measure of segregation is the index of segregation (also known as the isolation index, see *Clotfelter* [2004]), but we also show results using the more traditional index of dissimilarity. There is no data from between 1992 and 2006, and the missing data decreases the reliability of the post-2006 figures. Aside from our best estimates, we also present conservative lower and upper bounds. We introduce time series of the average level of segregation and maps for its geographic distribution. Finally, we show cross-sectional and log-differenced regressions for partial correlations of the between-school segregation with the size of the educational market, the average school size and the fraction of Roma students.

Our results indicate that school segregation, on average, is moderate in Hungary. The mean of the index of segregation is approximately 0.2 in the geographic areas covered by our analysis and is approximately 0.3 in the areas around the three largest cities. Note that Hungarian schools are characterized by fixed assignment to groups within schools ("classes"). Within-school between-class segregation may therefore be as important for inter-ethnic contact as between-school segregation. Unfortunately, our data does not make calculating indices within-school ethnic segregation possible. But it allows for looking at the segregation of students whose mother has eight grades of education or less, both between schools and within schools. On average, the level of their within-school segregation is about 40 percent on top of the level of their between-school segregation (details of the calculations are available from the authors upon request). This suggests that the level of ethnic segregation, if measured across classes instead of schools, is likely to be about 40 percent higher than the level of ethnic segregation across schools (0.28 instead of 0.20 on average, and 0.4 instead of 0.3 in the areas around the largest cities).

The data also show that, on average, the level of school segregation within Hungarian towns strengthened substantially between 1980 and 2011. According to our benchmark estimates, between-school segregation appears to have decreased between 2006 and 2008 and increased again afterwards. However, the trends after 2006 cannot be robustly identified due to severe data limitations. In the cross-sectional regressions, school segregation is positively associated with the size of the educational market and the share of the ethnic minority, similar to results from US metropolitan areas, and these relationships strengthened over time. In the regressions estimated in long term differences, the change in segregation is also linked with these factors, but the associations are weaker except for the change in the size of the Roma minority

The rest of the paper is organized as follows. Section 2 introduces the data, Section 3 defines the effective catchment areas of schools, and Section 4 presents the measures of segregation. Section 5 shows the average levels of segregation and its times series, and Section 6 introduces its geographic distribution. Section 7 details the regression results, and Section 8 concludes the paper.

1. Data and methods

The level of school segregation for a particular area is measured using the total number of students and the fraction of Roma students in each school within the area. We use two sources that cover the population of Hungarian primary schools. Before 1992, all schools filled out a compulsory questionnaire that contained, among other things, the total number of students and the number of Roma students in the school. The latter was based on counts by classes, carried out by teachers. We have data from the years 1980, 1989 and 1992. The reporting on Roma students was discontinued after 1992.

The data on the fraction of Roma students are available from 2006 in the Hungarian National Assessment of Basic Competences (NABC). It is a standards-based assessment, with tests on reading and mathematical literacy in grades 6 and 8 in primary schools (grades 4 and 8 in 2006 and 2007). The NABC became standardized in 2006, and we use data from 2006 through 2011 for our analysis. Aside from testing the students, it collects additional data on students and schools. School-level data are provided by the school principals in May of each year, when the testing takes place. Among other things, these contain information on the number of students and the school principal's estimate of the fraction of Roma students in the school. These estimates are likely to contain significantly more noise than the figures from 1992 and before, but we have no reason to believe that they are biased (they were not used for targeting any policy measure and they were not published, either).

The information is collected from each school site, that is, from each unit of the school with a separate address. This level of data collection is important because in some towns, the schools as administrative units comprise units at multiple locations, sometimes far from each other. Throughout the entire study, we use the word "school" to denote the school site and "institution" for the level of administrative organization that can contain more than one school site.[1]

Our analysis contains data on the population of Hungarian schools that teach primary school students, in other words, students in grades 1 through 8.[2] Of these schools, the NABC covers all that had students in grade 4 or 8 in 2006 and 2007, and all schools that had students in grades 6 and 8 from 2008 onwards. Coverage by the NABC is limited because it misses the institutions that teach students with special educational needs (S.E.N. students) except in 2006. Another source of bias is that the

[1] With very few exceptions, institutions were single-address schools before the early 1990s, so the data from between 1980 and 1992 are at the school and the institutional level at the same time.

[2] Traditionally, secondary schools would start with grade 9. In the early 1990s, some secondary schools began to recruit students in the lower grades and have incoming classes in grade 7 or as early as grade 5. These secondary schools are concentrated in the largest cities, most of them in Budapest. See *Horn* [2012] for a more detailed discussion. Our data cover all students in grades 1 through 8 including those enrolled in secondary schools. For simplicity, we call these institutions primary schools as well.

information on the fraction of Roma students is missing in some schools that do participate in the assessment. In addition to the problem of S.E.N. students, therefore, nonresponse is an additional cause of missing data.

Missing data can bias the segregation indices. Suppose, for example, that the schools in which the principal fails to provide information have no Roma students at all. In that case, our measures overestimate exposure and therefore underestimate segregation because the missing schools have exposure levels below the average. In theory, it is also possible that the schools with missing data have an ethnic composition that is very close to the town-level average. In that case, our measure of segregation would be biased upwards. Similarly, missing data can bias the estimates of the size of the Roma student population. If the schools with no information all have zero Roma students, the true share of Roma students among all students is lower than the estimate. If, instead, all of the schools with missing information are all-Roma schools, the true fraction of Roma students is higher than the estimates. Note that the bias is different for the segregation measures (a measure of dispersion) and the overall share of Roma students (a mean).

Table 1

Number of institutions and schools in Hungary in the administrative and NABC data, 2006–2011

Year	Number of institutions		Number of school sites	
	all (from KIR-STAT)	in the NABC data	in the NABC data	in the NABC data with non-missing fraction of Roma students
2006	3334	3267	3966	3444
2007	3247	3048	3420	2883
2008	2693	2465	3130	2885
2009	2541	2371	3097	2858
2010	2481	2307	3060	2792
2011	2454	2278	2925	2763

Note. "Schools" are defined by their physical location (address); "institutions" can contain more than one school. We consider primary schools (and their institutions) to be the schools that teach students from grade 1 through grade 8. KIR-STAT (statistical data collection part of the central Hungarian educational information system) is the administrative register for all educational institutions in Hungary. NABC (the National Assessment of Basic Competences) is the national standard-based assessment, with tests on reading and mathematics for grades 6 and 8 (grades 4 and 8 in 2006 and 2007). Students with special educational needs do not participate in the assessment, except in 2006. The school-level data in NABC cover all schools with at least one student who took part in the assessment.

Table 1 shows the prevalence of missing data. The table shows the number of institutions from the administrative files (KIR-STAT), the number of institutions in the NABC data, the number of schools in the NABC data (recall that we define a school as a facility

with a separate mailing address; some institutions have more than one school), and the number of schools with valid data. Administrative sources (KIR-STAT) have information on the number of students at the institution level but not at the school level as we define it. KIR-STAT has no information on the ethnic composition of schools.

Table 1 shows that both of the missing schools in the NABC data (and thus the missing information on all students) and the missing information on the Roma students in the NABC data are potentially important. We address the first problem by linking the schools through time and imputing student numbers from KIR-STAT. We address the problem of the missing Roma data in three alternative ways. The benchmark imputation is our best estimate. We complement the benchmark with an imputation that leads to the lowest possible value for the segregation index and one that leads to the highest possible one. Similarly, we compute the lower and upper bound estimates for the fraction of Roma students.[3] In most of the analysis, we focus on the results using the benchmark imputation, but we show the results with the alternative missing data treatments as well when they are important.

2. Defining catchment areas

School choice results in the extensive commuting of students between their residence and school. In this setting, the natural geographic unit for studying school segregation is the smallest area that covers all of the schools available to the students living in the area. In other words, it is the smallest area that is closed in terms of potential commuting. School segregation measured within larger units is influenced by residential patterns that commuting cannot overcome; school segregation measured within smaller units misses schools that should be considered.

In this section, we define the effective catchment areas of primary schools. Our smallest geographic units of observation are the municipalities (villages, towns, and cities; there are over 3000 in Hungary). A catchment area can consist of a single municipality and a single school, more than one municipality and a single school, or

[3] The benchmark procedure uses the data from previous and subsequent years for the schools that do not experience large changes in total student numbers. Approximately 30 schools are still missing data in each year after this procedure. The imputation that results in the lowest possible value of the segregation index uses the area-level average fraction of Roma students for the missing data (all initially missing data, including those that were filled in with our best estimate in the benchmark procedure). The imputation that leads to the highest value of the index of segregation imputes zero or one for the missing fraction of Roma students in a way that leaves the overall fraction of Roma students unchanged, up to indivisibility issues (it assigns the value of one to the smaller schools and zero to the larger schools following the observed relationship in the non-missing data). The imputation that leads to the lowest (highest) fraction of Roma students is simply zero (100 percent).

multiple municipalities and/or multiple schools. Ideally, all students who live in a catchment area go to a school within the area, and nobody from outside the area goes to the schools within the area. The goal is to partition Hungary into a complete collection of disjoint areas. Ideally, they should not be too large. Areas that are too large would not only work against the purpose of the exercise (by making area-level analysis difficult) but would go against spirit of the definition (very few schools would be available for any particular student within the area).

We used individual data collected from the NABC data for the students' residence and the location of their schools for three years. We created a directed and weighted graph using the individual data on commuting connections. Municipalities are the nodes (vertices) and the numbers of students commuting between the nodes are the links (edges). The direction of the link is from the node of residence to the node of the school, and the weights are the number of commuters. The largest weights in this graph are on the links that connect the nodes to themselves (loops): these are the students whose school and residence is within the same municipality.

Catchment areas are a partition of the set of all municipalities: every municipality belongs to one and only one catchment area. In the language of graph theory, catchment areas are the connected components in the entire graph. Connected components are defined for undirected (symmetric) and unweighted graphs: graphs that indicate whether two nodes are connected or not without any further information. For this problem, the original graph can be transformed into an undirected and unweighted graph with the help of a threshold value: two nodes are connected if and only if the number of students commuting between them exceeds a threshold level in any direction. Given the undirected and unweighted graph, the breadth-first-search algorithm finds all of the connected components in the graph and thus creates a partition of the set of all municipalities.[4]

The data on students' residence come from administrative records of all sixth-graders from three years, 2008, 2009, and 2010. The overall number of observations is 304 125. Simple coding errors or administrative mistakes could create apparent links between two municipalities with no links. The probability of such events is never zero, but the same event is unlikely to happen twice. For this reason, we have chosen two for the threshold value used to transform the weighted into the unweighted graph: nodes are considered to be connected if the data imply that more than one student is commuting in any direction between them.[5]

[4] See, for example, http://en.wikipedia.org/wiki/Breadth-first_search for a detailed description of the algorithm.

[5] If a municipality has no school and it sends one student only in these three years to any other municipality, that link is preserved. Similarly, the links that were below the threshold value of one student were preserved if they represented over 20 percent of all students from the sending municipality. Municipalities without schools that are not connected to any other municipality in the data were linked to the nearest neighbouring municipality that has a school (using geographic coordinates).

It turns out, however, that this benchmark graph has one giant component and many tiny ones. The graph contains 99 components; out of these, 96 have 13 or fewer nodes (the distribution is, of course, very skewed). Of the remaining three components, one has 44 nodes, one has 229 nodes, and the largest has 2 669 nodes.[6] The giant component contains Budapest and most cities from all regions of Hungary. This partition is clearly useless for any practical analysis. Therefore, we created an alternative partition: we broke the largest three components into smaller clusters by increasing the threshold value for links to 5 students per year on average (a total of 15 students for the three years) or at least 20 percent of the originating node (the municipality of residence).[7] The resulting partition contains 1 055 catchment areas. The largest area contains 71 municipalities, and it covers the Budapest agglomeration. The other large areas contain large cities and their agglomerations.[8]

Table 2 shows the most important summary statistics on the catchment areas.[9] Not surprisingly, the size distribution is skewed, and the areas with the highest number of municipalities are even larger in terms of student population because they contain the largest cities.

Table 2

Number of municipalities, primary schools and students

Size of catchment area (number of municipalities)	Number of catchment areas	Average number of municipalities	Average number of primary schools	Average number of primary school students
1	624	1.0	1.2	232
2 to 4	297	2.7	2.1	408
5 to 9	74	6.3	7.1	1 885
10 to 19	37	12.9	9.4	2 192
20 to 49	20	30.5	25.4	6 102
50 to 71	3	60.0	224.9	65 045
Total	*1055*	3.0	3.3	782

Note. Information from schools is averaged over 2006 through 2011.

[6] The emergence of a giant component is a classic result in graph theory: if links are created randomly, almost all nodes are connected with a high probability when the number of links exceeds a threshold value.

[7] Similarly to the previous step, links were preserved even if they were below the threshold when a municipality has no school and it sends its students to one and only one other municipality. Municipalities without schools that are not connected to any other municipality in the data were linked to the nearest neighboring municipality that has a school (using geographic coordinates).

[8] The threshold values used in the new partition are obviously ad-hoc, but the results represent an intuitively compelling partition and any attempt to break the giant component would require assumptions of this kind.

[9] Additional data on the catchment areas, including the set of municipalities in them and further data on students, are available from the authors upon request.

3. Measuring school segregation

Following the literature (for example, *Clotfelter* [2004]), we measure segregation with the help of the following three indices: exposure of non-Roma students to Roma students (*ENR*), exposure of Roma students to non-Roma students (*ERN*), and the standardized version of these indices, referred to here as the segregation index (*S*). For completeness, we also look at the more traditional but theoretically less attractive index of dissimilarity (*D*). When we calculate the extent of exposure or segregation, we look at schools within a catchment area (or, alternatively, a micro-region, a town, or a city). To define and interpret these indices, we work with the following notation. Index i denotes the schools, and index j denotes the areas (these are the areas that contain the schools; students may reside outside the areas, see our discussion later). I_j is the number of schools in area j, N_{ij} is the number of students in school i in area j, N_j is the number of students in area j, R_{ij} is the number of Roma students in school i in area j, R_j is the number of Roma students in area j, r_{ij} is the fraction of the Roma students among all students in school i in area j, r_j is the fraction of the Roma students among all students in area j, $\left(1-r_{ij}\right)$ is the fraction of the non-Roma students among all students in school i in area j, $\left(1-r_j\right)$ is the fraction of the non-Roma students among all students in area j. Index ENR_j measures the exposure of an average (a randomly chosen) non-Roma student in area j to the possibility of meeting Roma students. ENR_j is equal to the fraction of Roma students in each school averaged over schools, where the average is taken with weights that are equal to the share of non-Roma students in the school in all non-Roma students in the area. Formally,

$$ENR_j = \sum_{i=1}^{I_j} r_{ij} \frac{N_{ij} - R_{ij}}{N_j - R_j}, \quad \text{so that} \quad 0 \le ENR_j \le r_j .$$

The minimum value of the exposure index is zero: in this case, no contact is possible between Roma and non-Roma students within the schools because the schools are either all-non-Roma (when $r_{ij} = 0$) or all-Roma (when $N_{ij} - R_{ij} = 0$). The maximum value of exposure is when the fraction of minority students in each school is equal to the fraction in the area: $r_{ij} = r_j$ for all i in j. For ENR_j to make sense, we need $0 < r_j < 1$, that is, there must be both Roma and non-Roma students in area j. This condition is satisfied in all of the areas that we consider.

The exposure of Roma students to non-Roma students (ERN_j) is analogous: it measures the exposure of an average (a randomly chosen) Roma student in area j to the possibility of meeting non-Roma students. ERN_j is equal to the fraction of non-Roma students in each school averaged over schools, where the average is taken with weights that are equal to the share of the school in the Roma student population of the area. Formally,

$$ERN_j = \sum_{i=1}^{I_j} \left(1 - r_{ij}\right) \frac{R_{ij}}{R_j}, \quad \text{so that} \quad 0 \le ERN_j \le 1 - r_j .$$

The minimum value of this exposure index is zero, also, and $ERN_j = 0$ exactly when $ENR_j = 0$. This value indicates that no contact is possible between Roma and non-Roma students within the schools because the schools are either all-Roma $\left(1 - r_{ij} = 0\right)$ or all-non-Roma $r_{ij} = 0$. The maximum value of Roma exposure occurs when the fraction of non-Roma students in each school is equal to the fraction in the area: $1 - r_{ij} = 1 - r_j$ for all *i* in *j*. The two indices are intimately related:

$$ERN_j = \frac{1 - r_j}{r_j} ENR_j .$$

Despite their intuitive content, the exposure indices are rarely used. Their values depend on the overall fraction of minority students in the area, which poses a severe constraint on their use in comparing segregation across time or areas. The segregation index is intended to solve this problem. It is a normalized version of the exposure indices, and thus it retains their information content, albeit in a less intuitive way. The normalization amounts to comparing exposure to its attainable maximum; there is also a reversal of sign so that the higher levels of the index indicate higher levels of segregation (less exposure). Intuitively, the segregation index shows the fraction of contact possibilities that are made impossible by segregation. Formally,

$$S_j = \frac{r_j - ENR_j}{r_j} = \frac{\left(1 - r_j\right) - ERN_j}{1 - r_j}, \quad \text{so that} \quad 0 \le S_j \le 1 .$$

The maximum value of the index is one: segregation is at its maximum when the exposure is zero. The minimum value is zero: it is attained at maximum exposure, which is when the fraction of Roma students is the same in every school.

An alternative measure of segregation is the index of dissimilarity. Defined from the viewpoint of Roma students, and with many schools in mind, this index can be interpreted as the percentage of non-Roma students that would have to move to different schools to have schools with the same fraction of Roma students within the area. Formally, the index of dissimilarity is defined as

$$D_j = \frac{1}{2}\sum_{i=1}^{I_j}\left|\frac{R_{ij}}{R_j} - \frac{N_{ij}-R_{ij}}{N_j - R_j}\right|, \quad \text{so that} \quad 0 \le D_j \le 1.$$

Similar to the index of segregation defined formerly, a value of 1 would denote complete segregation, and a value of 0 would denote equal distribution across schools. In any other case, the index of dissimilarity is, in general, not equal to the index of segregation. The index of dissimilarity is a more traditional measure than the index of segregation, but it lacks the latter's theoretical relationship to exposure. For that reason, the index of segregation is a more useful measure that is used in the new literature on school segregation (*Clotfelter* [1999]).

4. Trends in school segregation in Hungary between 1980 and 2011

We measure the ethnic composition of primary schools and segregation between schools in years 1980, 1989, 1992 and yearly between 2006 and 2011. Recall that the data in 1980, 1989 and 1992 are high quality, that there are no data from between 1992 and 2006, and that the data starting with 2006 are of lower quality, characterized by many schools without information on the fraction of Roma students. For that reason, from 2006 onwards, we show the conservative lower and upper bound estimates of both the overall share of Roma students and the index of segregation in addition to our best estimates. We define segregation within three geographic areas: catchment areas, micro-regions, and municipalities. Naturally, between-school segregation is defined for the areas with two schools or more. We restricted the analysis to areas that had two schools or more in each year of observation. This criterion was fulfilled by 175 out of the 1 055 catchment areas and 140 towns or cities of the over 3 000 municipalities.

Table 3 shows the averages of the segregation indices in 1980 and 2011 in the three geographic areas. The averages shown in the table are weighted by the distribution of students.

Table 3

Ethnic composition and ethnic segregation of primary schools in catchment areas as well as in micro-regions and larger municipalities (towns and cities) in 1980 and 2011

Average values	Larger catchment areas		Micro-regions		Towns and cities	
	1980	2011	1980	2011	1980	2011
Average number of students	5 153	3 324	6 668	4 235	4 723	3 139
Fraction of Roma students	0.05	0.11	0.06	0.13	0.03	0.08
Exposure of non-Roma students to Roma students	0.04	0.08	0.05	0.10	0.03	0.07
Exposure of Roma students to non-Roma students	0.86	0.69	0.85	0.68	0.90	0.75
Index of segregation	0.09	0.22	0.09	0.22	0.07	0.19
Index of dissimilarity	0.48	0.53	0.47	0.53	0.47	0.51
Number of observations	175	175	174	174	140	140

Note. Average values (using the benchmark imputations from 2006 onwards) weighted by the number of students (except for the average number of students, which is unweighted).

The first row of Table 3 shows the number of students. The most important information here is the uniform decline in the number of students by about 35 percent. The second row presents the fraction of Roma students. The figures show a strong increase: the fraction of Roma students in Hungarian primary schools more than doubled between 1980 and 2011. A small part of their growing share is due to the greater participation of Roma students in primary school education, but a large part is due to demographics.

The catchment areas shown in this table refer to areas that had two or more schools during the 1980 to 2011 period and thus do not cover the smallest catchment areas, which have only one school. In the part of Hungary that is covered by these two-or-more-school catchment areas, the share of Roma students was 5 percent in 1980 and increased to 11 percent by 2011. The micro-regions cover the entire country, and thus the figures in the corresponding columns refer to the overall fraction of Roma students in Hungary. From a 6 percent level in 1980, the share of Roma students in primary schools (grades 1 through 8) increased to 13 percent by 2011. The corresponding figures in the larger municipalities (towns and cities with two or more schools) are 3 percent in 1980 and 8 percent in 2011. The lower levels in the larger catchment areas and the even lower levels in the larger municipalities show that the Roma population is overrepresented in the smaller villages and that the degree of overrepresentation did not decrease over time.

The exposure of non-Roma students to Roma students increased, but at a slower pace than the growth in the share of Roma students, the theoretical maximum of the exposure index. Mirroring this trend, the exposure of Roma students to non-Roma students declined significantly, more than the decreasing share of non-Roma students would imply. Taken together, the trends in the indices indicate a growing trend in the segregation index.

Between-school segregation increased substantially in Hungary between 1980 and 2011. Taking the average of the catchment areas, the relevant geographic units in the system of free school choice in Hungary, the index of segregation raised from 9 percent to 22 percent. The intuitive content of these figures is that the chance of contact between Roma students with non-Roma schoolmates decreased from 91 percent of its theoretical maximum in 1980 to 78 percent of the maximum level by 2011.

Table 4

Ethnic composition and ethnic segregation of the primary schools in the catchment areas around the largest Hungarian cities in 1989 and 2011

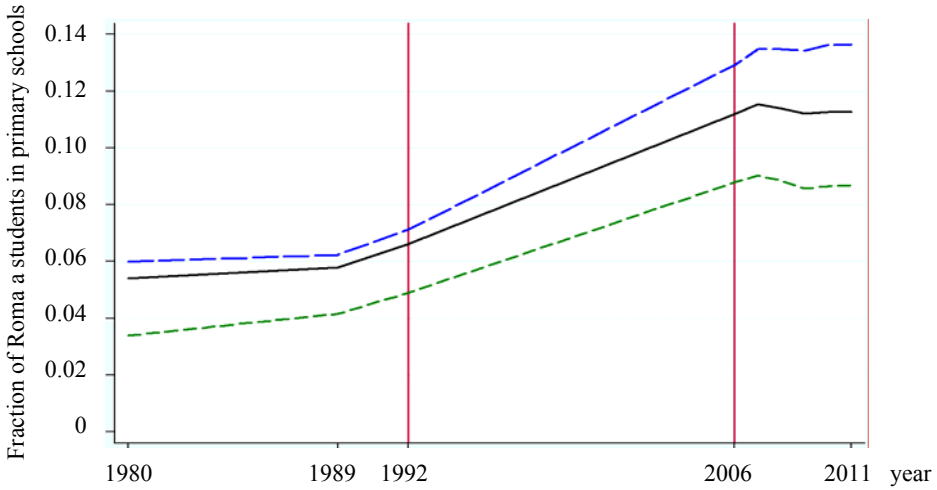
Year	Indicator			
	Number of students	Fraction of Roma students	Index of segregation	Number of municipalities
		Budapest		
1980	237 896	0.02	0.06	71
2011	165 931	0.07	0.26	
		Miskolc		
1980	35 255	0.09	0.13	33
2011	20 818	0.20	0.36	
		Debrecen		
1980	28 280	0.02	0.09	7
2011	20 361	0.04	0.26	
		Pécs		
1980	24 020	0.04	0.13	55
2011	15 489	0.08	0.16	
		Szeged		
1980	20 178	0.02	0.16	12
2011	14 311	0.03	0.05	
		Gyõr		
1980	19 736	0.02	0.06	37
2011	13 316	0.04	0.13	

Table 4 shows the number of students, the share of Roma students and the index of segregation for the catchment areas around the six largest Hungarian cities in 1980 and 2011. Similar to the national trends, these areas experienced a large drop of 30 to 40 percent in the number of students. Again, similar to the national trends, the share of Roma students got higher substantially in each area. The levels differ considerably, but the trends are rather similar except for the catchment area of Budapest where the increase was more than three-fold, from 2 percent to 7 percent. The highest share, both in 1980 and in 2011, was in the catchment area of the northern city Miskolc, while the lowest one was in the catchment area of the southern city Szeged.

Ethnic segregation strengthened considerably in most but not all of the catchment areas. The index of segregation grew almost threefold in the areas of Budapest, Miskolc and Debrecen. Segregation increased by a smaller amount in the Pécs and Győr areas, and it decreased substantially in the Szeged area.

The level of segregation in 1980 could be considered to be low; the level in 2011 is moderate. The US metropolitan areas that are characterized by the school segregation of African Americans and whites similar to the levels documented for large Hungarian areas include San Diego (0.28), Phoenix (0.31) or Los Angeles (0.33). These are not among the most segregated US cities: the segregation index is 0.45 in New York City, 0.57 in Chicago; while the most segregated metropolitan area is that of Detroit (0.71, see *Clotfelter* [1999] p. 494.).

Figure 1. Time series of the fraction of Roma students in primary schools in larger catchment areas, micro-regions and larger municipalities (towns and cities) from 1980 to 2011

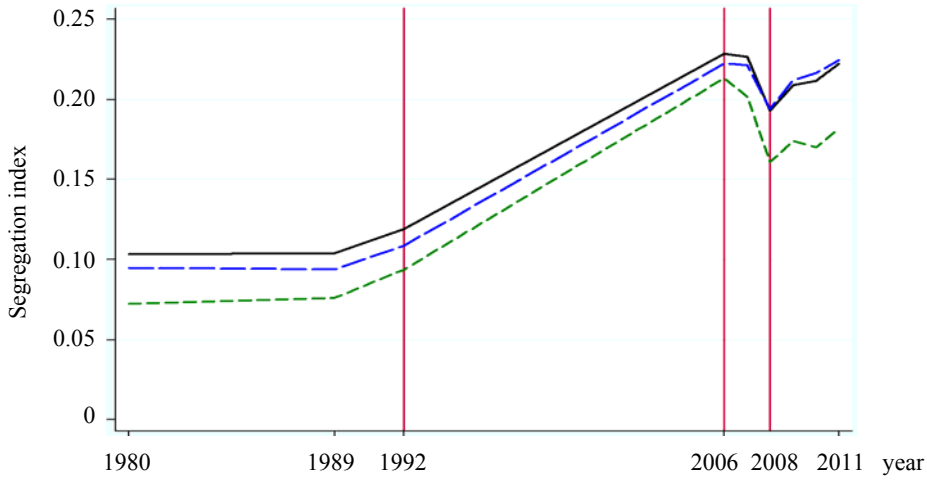


Note. The fraction of Roma students after 2006 is based using our benchmark imputations for missing data. The solid line indicates larger catchment areas, while the long dashed line is for micro-regions and the dashed line is for larger municipalities (towns and cities).

Figure 1 shows the times series of the fraction of Roma students as estimated using the benchmark imputation procedure. Figure A1 in Appendix shows the same time series together with the conservative lower and upper bounds. Recall that the bounds represent the most conservative imputations for the missing data. While we cannot rule out any figure within the bounds, our benchmark estimates use available information in a careful way and are thus likely to be close to the true figures. The post-2006 data are also noisier, although that noise is unlikely to have a significant effect on the aggregate figures. According to the benchmark results, the increase in the fraction of Roma students was concentrated in the 1989 to 2007 period, and it stopped afterwards. When one looks at the intervals between the lower and upper bounds in Figure A1, the apparent trend break is lost in the overall degree of uncertainty.

Figure 2 shows the time series for the index of segregation, and Figure A2 presents the uncertainty interval for our calculations using the lower and upper bound imputations for the missing data in 2006–2011. The figures show the time series of the index of segregation from 1980 to 2011 averaged over the geographic areas (catchment areas, micro-regions, and larger municipalities).

Figure 2. Time series of the average of the index of ethnic segregation between primary schools in larger catchment areas, micro-regions and larger municipalities (towns and cities) from 1980 to 2011



Note. The index after 2006 is based using our benchmark imputations for the missing data. The average of the index is weighted by the number of students. The solid line indicates larger catchment areas, while the long dashed line is for micro-regions and the dashed line is for larger municipalities (towns and cities).

According to Figure 2, between-school segregation by ethnic lines stayed constant between 1980 and 1989 but began to increase afterwards. By 2006, it reached a value that is more than double the 1989 level. This growth is large and is also robust

to the imputation method that we chose for the missing data. Our best estimate for the index shows a significant decline in between-school segregation in the 2006–2008 period that appears to be driven by the larger municipalities. The slope of the decreasing trend is comparable to that of the previous increase, resulting in a small drop because of the short time interval.

The trend breaks in the time series coincide with trends in the desegregation initiatives of the government of Hungary. A law introduced in 2004 banned segregation based on race, ethnicity and social background and divided the burden of proof between the plaintiffs and the defendants. In the following years, advocacies and offices of the central government pressured some of the towns and cities to close down segregated schools. By anecdotal evidence, these central government activities came to a halt after 2008. The link between desegregation in larger municipalities and the observed patterns of segregation is further supported by the fact that the trend breaks are largest for the largest municipalities, from 0.21 to 0.16. The drop was smaller, from 0.23 to 0.19 in the catchment areas that included not only the towns and cities but also some of the surrounding villages. This finding is consistent with the larger municipalities implementing desegregation within their administrative boundaries without the other parts of their catchment area following suit. Furthermore, some of the largest drops between 2006 and 2008 are observed in the cities that carried out changes in the composition of their schools as a result of desegregation plans (including, for example, Szeged, shown in Table 4). This evidence suggests that the observed trend breaks could be real.

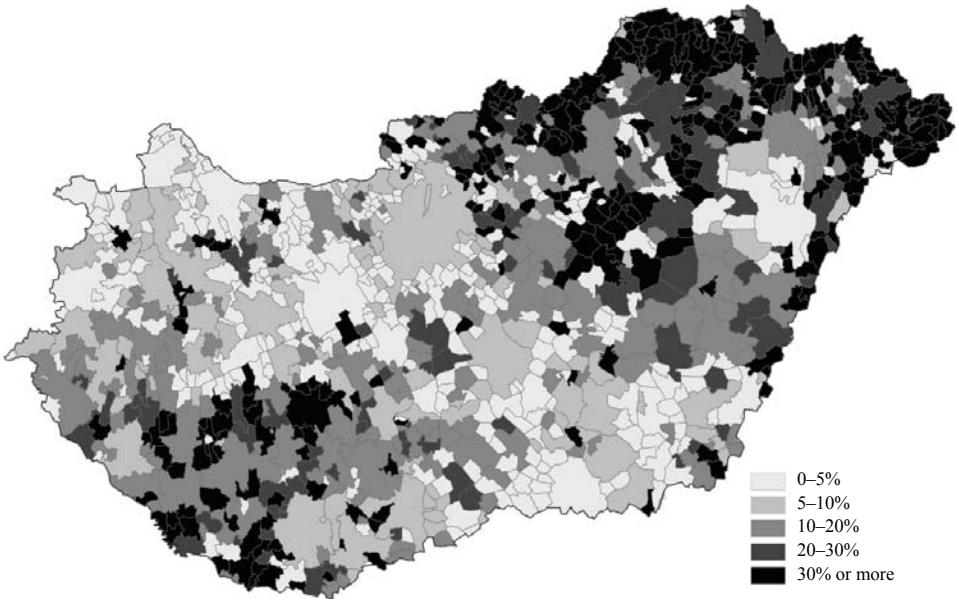
However, they also coincide with the apparent breaks in the time series of the share of Roma students, which is harder to understand. This trend implies that the estimated breaks in the segregation indices could be spurious. Indeed, while the large increase between 1992 and 2006 is robust to the imputation method used after 2006, the trend breaks after 2006 are not robust at all. Similar to the Roma share series, the benchmark estimates are surrounded by a very wide interval of possible values between the conservative lower and upper bounds, shown in Figure A2. As a result, the coincidence of the trend breaks with the desegregation activities could be completely spurious. Evidently, the missing information in the NABC data simply prevents us from identifying trends after 2006.

5. The geographic distribution of school segregation

The Roma population is distributed unevenly in Hungary. Using all data available up to 1993, *Kertesi–Kézdi* [1998] presented detailed maps on the geographic distri-

bution of the Roma population in Hungary. Using school-level information in a system characterized by school choice and the widespread commuting of students, we can present analogous maps at the level of the catchment areas for the 2000s. Figure 3 shows a map of Hungary divided into catchment areas (1 055 clusters of villages, towns and cities) with the fraction of Roma students for all areas in 2011.

Figure 3. The share of Roma students in primary schools in all catchment areas of Hungary in 2011



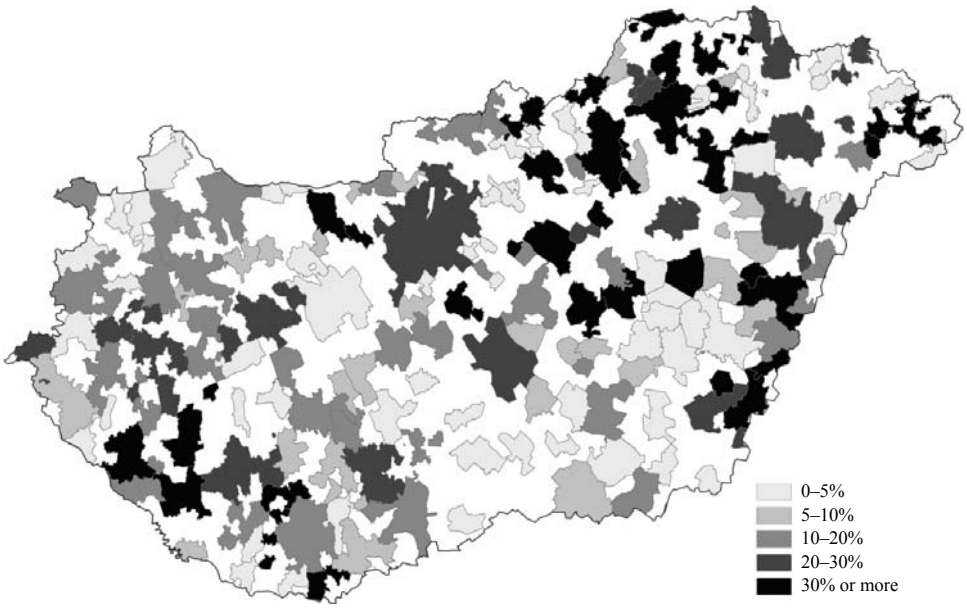
Note. Catchment areas are clusters of villages, towns and cities that are close in terms of student commuting. We defined these areas using the observed commuting patterns. The calculations are based on our benchmark imputations for the missing data.

As we presented formerly, between-school segregation is defined for the larger catchment areas. Figure 4 shows the map of the 175 largest catchment areas and presents the index of segregation in these areas.

Comparing the two maps suggests two patterns. The geographic distribution of school segregation is quite alike to the geographic distribution of the Roma students. This similarity indicates a positive and potentially quite strong relationship between the share of Roma students in the area and the level of ethnic segregation as regards primary schools. However, this correspondence is far from being perfect. The areas around Budapest, Pécs and Győr, for example, are characterized by relatively strong segregation but a low fraction of Roma students. This finding suggests that other mechanisms can also be important and that the size of the area

is likely to be related to these mechanisms. In the following section, we present regression results that show some more systematic evidence for these types of associations.

Figure 4. Ethnic segregation between primary schools in the larger catchment areas of Hungary in 2011



Note. Catchment areas are clusters of villages, towns and cities that are close in terms of student commuting. We defined these areas using the observed commuting patterns. The calculations are based on our benchmark imputations for the missing data.

6. School segregation and the size of the educational market, schools and the Roma population

In our final analysis, we show regression results with the index of segregation being the left hand-side variable and the size of the area (number of schools), the average size of the schools and the size of the Roma minority (fraction of Roma students) on the right hand-side. We first present the results from the cross-sectional regressions for 1980 and 2011. They show cross-sectional associations: whether, in a given point in time, the areas that are larger, have bigger schools or have a greater fraction

of Roma students in the schools are characterized by higher or lower levels of school segregation.[10, 11]

The results are shown in Table 5, and the summary statistics are in Table A1. The number of schools in the area was positively associated with school segregation in 2011, while the association was substantially weaker in 1980. The change is also statistically significant. In 2011, the standard deviation of the log number of schools was between 0.8 and 1.0 depending on the geographic area definition (see Table A1); the areas that are larger by one standard deviation were characterized by a one-tenth of a standard deviation higher index of segregation on average, holding ethnic composition and average school size constant. The average size of the schools is negatively, albeit weakly, correlated with the segregation between schools, with no clear pattern across years or definitions of the geographic area.

Table 5

School segregation and the sizes of the educational market, schools and the Roma population

Dependent variable – index of segregation	Larger catchment areas		Micro-regions		Larger municipalities	
	1980	2011	1980	2011	1980	2011
Log number of schools	0.022 [2.45]*	0.055 [4.98]**	0.020 [1.81]	0.066 [7.10]**	0.021 [2.06]*	0.062 [8.84]**
Log average school size	–0.024 [1.51]	–0.022 [0.58]	–0.032 [2.19]*	–0.067 [2.17]*	–0.056 [2.09]*	–0.036 [0.84]
Fraction of Roma students	0.439 [4.27]**	0.661 [6.86]**	0.247 [3.00]**	0.563 [8.40]**	0.624 [2.53]*	0.747 [6.06]**
Constant	0.142 [1.41]	0.076 [0.36]	0.200 [2.23]*	0.288 [1.80]	0.343 [2.02]*	0.121 [0.49]
Number of observations	175	175	174	174	140	140
R-squared	0.12	0.30	0.10	0.35	0.11	0.42

Note. Cross-sectional regressions for selected years. Robust *t*-statistics in brackets. * significant at the 5 percent level; ** significant at the 1 percent level. Observations are weighted by the square root of the number of students in the area.

[10] Apart from the missing information from some schools after 2006, our data represent the population of schools. We sew standard errors nevertheless, because we interpret our regressions as models that try to uncover more general tendencies in educational markets, characterized by the properties of the Hungarian educational markets in the observed years.

[11] Note that the Budapest agglomeration is an outlier in terms of size, and it experienced larger than average increase in both the share of Roma students and the index of segregation. Nevertheless, the estimated coefficients are very similar when we exclude Budapest.

The fraction of Roma students in the area is the strongest predictor of school segregation, with increasing magnitude over time and across geographic units (being the strongest predictor within towns and cities). Towns and cities that had a one percentage point greater fraction of Roma students in their schools were characterized by a 0.75 percentage point higher index of segregation. In terms of standardized coefficients, the towns and cities with a fraction of Roma students that is greater by one standard deviation (0.1) were characterized by a half of a standard deviation (0.14) higher index of segregation on average, holding the number of schools and the average school size constant.[12]

Table A2 shows the regression results for all years for the larger catchment areas. They suggest that the large increase in the coefficients took place between 1992 and 2006, and the years after 2006 are characterized by further increases, with ups and downs without any clear pattern.

Table 6

Changes in school segregation and in the sizes of the educational market, schools and the Roma population from 1980 to 2011

Dependent variable – change in index of segregation	Larger catchment areas	Micro-regions	Larger municipalities
Log change in number of schools	0.170 [3.23]**	0.116 [2.43]*	0.018 [0.42]
Log change in average school size	0.068 [1.30]	–0.01 [0.19]	–0.059 [1.08]
Change in fraction of Roma students	0.605 [4.31]**	0.792 [7.39]**	0.839 [3.84]**
Constant	0.098 [3.01]**	0.057 [2.05]*	–0.016 [0.56]
Number of observations	175	174	140
R-squared	0.17	0.23	0.14

Note. Regression results. Robust t-statistics in brackets. * significant at the 5 percent level; ** significant at the 1 percent level. Observations are weighted by the square root of the number of students in the area.

After the cross-sectional regressions, we turn to the regressions estimated in long differences: changes between 1980 and 2011. Table 6 shows the results, and Table

[12] These results are similar to the regression results of *Clotfelter* ([1999] p. 501.). In particular, the magnitudes of all three partial correlations are similar to our estimates. His regression has the log number of students as opposed to that of schools and the log average size of the school districts as opposed to that of the schools. Of course, his measure of segregation is between African American and white students. Our results are very similar if we include the log number of students instead of the log number of schools.

A3 has the appropriate summary statistics. Table A4 and A5 show the corresponding results separately for the communist period (1980 to 1989) and the post-communist period (1989 to 2011).

The results from these regressions show the extent to which the areas that experienced larger-than-average increases in the number of schools, school size or the fraction of Roma students tend to be characterized by larger-than-average growth in school segregation. When interpreting the results, one must keep in mind that, typically, school segregation strengthened, the number of schools decreased (except in the larger municipalities), the average school size became smaller (especially in the larger municipalities) and the fraction of Roma students grew during the observed period. These trends were the most pronounced during the post-communist period (1989 to 2011). On average, there were no significant shifts before 1989, but the variation in changes was substantial even then, so that interesting associations can be identified.

The results are qualitatively similar to the cross-sectional associations measured in 2011. Growth (drop) in the number of schools by 10 percent is associated with an increase (decline) in the index of school segregation by one to two percentage points in the larger catchment areas and the micro-regions. These magnitudes are actually stronger than the cross-sectional estimates in 2011: a one standard deviation (0.35 to 0.42) higher rise in the log number of schools is associated with an approximately one third of a standard deviation (0.14 to 0.16) increase in segregation. No association is present within the larger municipalities. The changes in the average school size are not associated with changes in segregation, holding the number of schools and the ethnic composition constant. Similar to the cross-sectional results, the change in the fraction of Roma students is the strongest predictor of changes in school segregation. The magnitudes are similar to the cross-sectional associations (a one standard deviation growth in the fraction of Roma students is associated with a half of a standard deviation increase in segregation).

7. Conclusions

In this paper, we documented the degree of between-school segregation of Roma versus non-Roma students between 1980 and 2011. We showed the long-run trends and geographic distributions as well as the regression estimates of some robust associations.

An important contribution of our paper was the definition of school catchment areas: clusters of villages, towns, and cities that are closed in terms of student commuting in the 2000s. This geographic aggregation allows school segregation to be ana-

lyzed at the level of the smallest and most relevant geographic area. The use of the catchment areas also allows school-level information to be used to estimate figures for the people living in those areas, such as the share of the Roma minority.

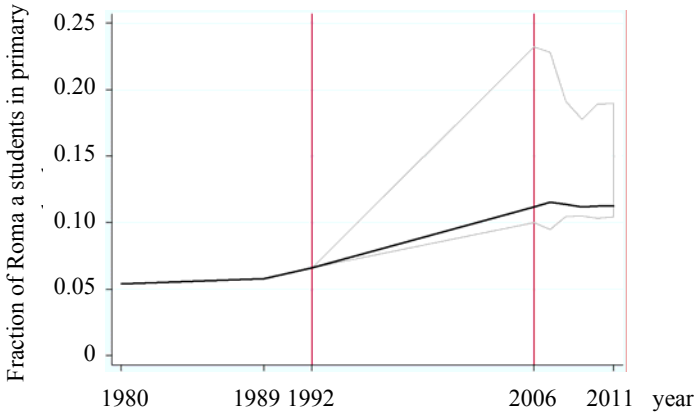
From a theoretical point of view, our most interesting results are the regression estimates. They show that the size of the educational markets (defined as the number of schools) is strongly and positively associated with between-school segregation. This association is consistent with the notion that school choice and selective commuting are among the most important mechanisms behind segregation, and the size of the market increases differentiation between schools, therefore providing a higher incentive to commute. This explanation is, however, not the only possible one. The fraction of Roma students in the area is an even stronger predictor of segregation. Explaining this association could be even harder. However, both associations are robust in the sense that they are identified from the cross-section as well as from the long differences, and analogous results for both are found in the US as well.

From a policy perspective, another interesting finding is the coincidence of an apparent trend break in segregation between 2006 and 2008, correspondent to the timing of the most intensive desegregation campaigns. Unfortunately, the quality of the data does not allow for a robust analysis here. Improving the data quality by implementing the full coverage of schools is necessary for fine analysis of the effects of desegregation policies and other aspects of school segregation in Hungary.

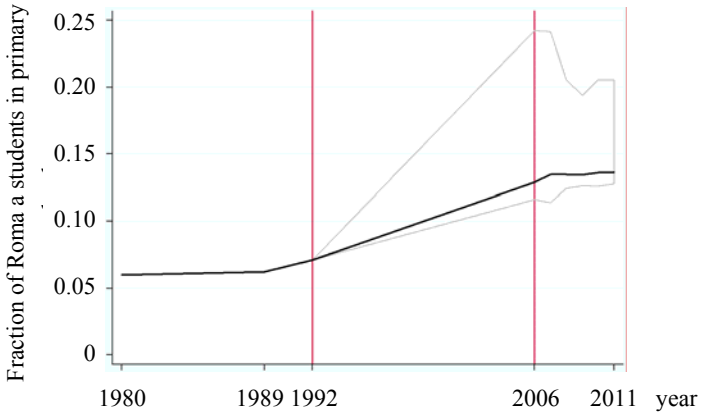
Appendix

Figure A1. Time series of the fraction of Roma students primary schools in larger catchment areas (panel A), micro-regions (panel B) and larger municipalities (towns and cities; panel C) between 1980 and 2011

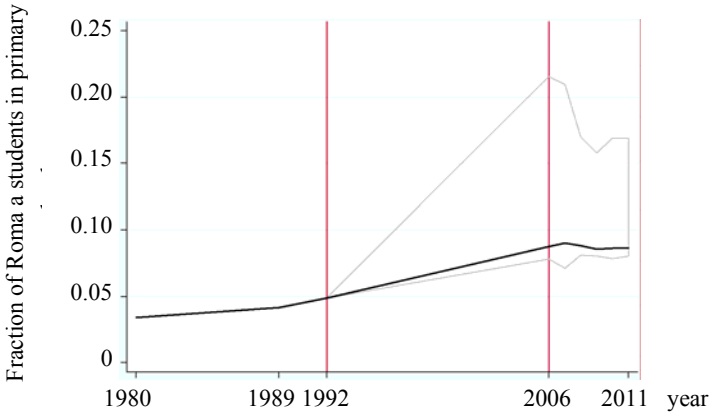
Panel A. Larger catchment areas



Panel B. Micro-regions



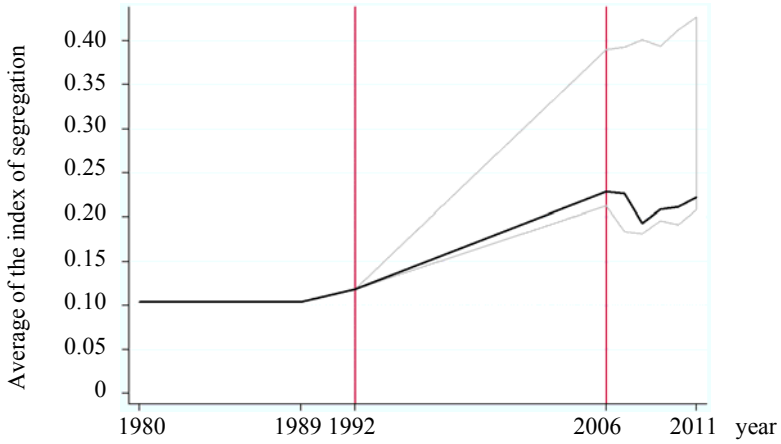
Panel C. Larger municipalities (towns and cities)



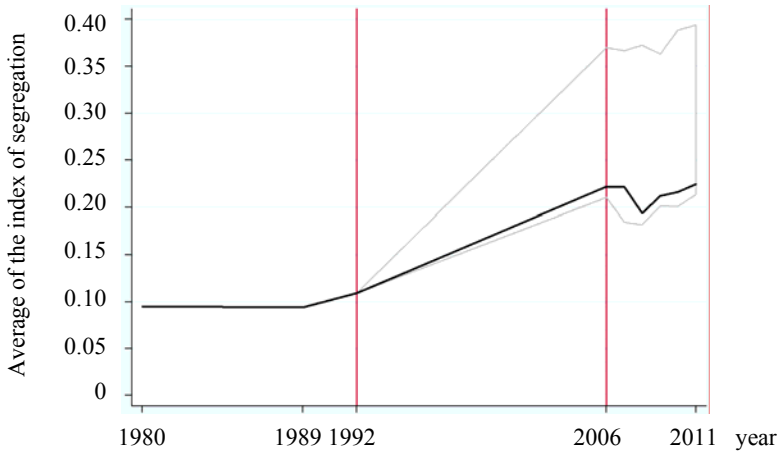
Note. The lines are based on our benchmark imputations for missing data after 2006. Grey area shows conservative lower and upper bounds using alternative imputations.

Figure A2. Time series of the average of the index of ethnic segregation between primary schools in micro-regions (panel A), catchment areas (panel B) and larger municipalities (towns and cities; panel C) between 1980 and 2011

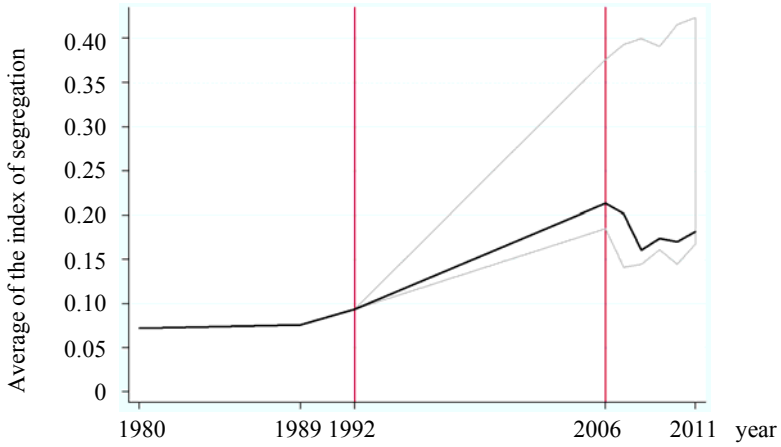
Panel A. Micro-regions



Panel B. Catchment areas



Panel C. Larger municipalities (towns and cities)



Note. The lines are based on our benchmark imputations for missing data after 2006. Grey area shows conservative lower and upper bounds using alternative imputations. Average of the index is weighted by number of students.

Table A1

Summary statistics for school segregation and for the sizes of the educational market, schools and the Roma population (corresponding to the regressions in Table 5 of the article)

Summary statistics	Larger catchment areas		Micro-regions		Larger municipalities	
	1980	2011	1980	2011	1980	2011
Mean (index of segregation)	0.08	0.17	0.10	0.20	0.06	0.12
Log number of schools	2.05	1.83	8.43	7.93	7.81	7.41
Log average school size	5.71	5.41	5.60	5.33	6.15	5.60
Fraction of Roma students	0.08	0.18	0.08	0.17	0.06	0.12
Standard deviation (index of segregation)	0.09	0.17	0.08	0.14	0.11	0.14
Log number of schools	0.91	0.93	0.72	0.80	0.90	0.89
Log average school size	0.44	0.31	0.41	0.27	0.39	0.28
Fraction of Roma students	0.06	0.14	0.06	0.14	0.05	0.10
Number of observations	175	175	174	174	140	140

Table A2

School segregation and the sizes of the educational market, schools and the Roma population

Dependent variable – index of segregation	Larger catchment areas								
	1980	1989	1992	2006	2007	2008	2009	2010	2011
Log number of schools	0.022	0.022	0.025	0.039	0.041	0.038	0.038	0.042	0.055
	[2.45]*	[3.00]**	[3.47]**	[3.38]**	[3.37]**	[4.05]**	[4.30]**	[3.89]**	[4.98]**
Log average school size	–0.024	–0.010	–0.006	0.073	0.065	0.051	0.050	0.027	–0.022
	[1.51]	[0.75]	[0.39]	[2.87]**	[2.46]*	[1.66]	[0.99]	[0.57]	[0.58]
Fraction of Roma students	0.439	0.464	0.511	0.555	0.620	0.635	0.595	0.617	0.661
	[4.27]**	[4.54]**	[5.36]**	[6.12]**	[6.45]**	[6.15]**	[5.52]**	[5.66]**	[6.86]**
Constant	0.142	0.055	0.019	–0.379	–0.354	–0.296	–0.274	–0.156	0.076
	[1.41]	[0.62]	[0.22]	[2.95]**	[2.53]*	[1.75]	[1.01]	[0.60]	[0.36]
Number of observations	175	175	175	175	175	175	175	175	175
R-squared	0.12	0.19	0.22	0.24	0.26	0.27	0.20	0.23	0.30

Note. Cross-sectional regressions for all years for the larger catchment areas. Robust t-statistics in brackets. * significant at 5 percent, ** significant at 1 percent. Observations are weighted by the square root of the number of students in the area.

Table A3

Summary statistics of the changes in school segregation and in the sizes of the educational market, schools and the Roma population from 1980 to 2011
(corresponding to the regressions in Table 6 of the article)

Summary statistics	Larger catchment areas	Micro-regions	Larger municipalities
		Mean	
Log change in index of segregation	0.09	0.10	0.06
Log change in number of schools	–0.22	–0.23	0.15
Log change in average school size	–0.30	–0.27	–0.55
Change in fraction of Roma students	0.10	0.09	0.06
		Standard deviation	
Log change in index of segregation	0.16	0.14	0.16
Log change in number of schools	0.42	0.35	0.35
Log change in average school size	0.41	0.30	0.39
Change in fraction of Roma students	0.10	0.09	0.07
Number of observations	175	174	140

Table A4

Changes in school segregation and in the sizes of the educational market, schools and the Roma population from 1980 to 1989 and from 1989 to 2011

Dependent variable – change in index of segregation	From 1980 to 1989			From 1989 to 2011		
	Larger catchment areas	Micro-regions	Larger municipalities	Larger catchment areas	Micro-regions	Larger municipalities
Log change in number of schools	0.085	0.002	0.112	0.171	0.095	–0.013
	[2.30]*	[0.05]	[1.24]	[2.84]**	[2.28]*	[0.26]
Log change in average school size	–0.001	–0.068	–0.022	0.061	–0.028	–0.072
	[0.02]	[1.86]	[0.26]	[1.04]	[0.56]	[1.30]
Change in fraction of Roma students	0.612	1.194	1.708	0.564	0.69	0.759
	[1.29]	[2.16]*	[1.73]	[4.24]**	[7.22]**	[4.27]**
Constant	–0.009	–0.007	–0.019	0.109	0.06	–0.005
	[1.42]	[1.05]	[1.33]	[3.32]**	[2.51]*	[0.17]
Number of observations	175	174	140	175	174	140
R-squared	0.06	0.08	0.13	0.19	0.25	0.19

Note. Regression results. Robust *t*-statistics in brackets. * significant at 5 percent; ** significant at 1 percent. Observations are weighted by the square root of the number of students in the area.

Table A5

Summary statistics of the changes in school segregation and in the sizes of the educational market, schools and the Roma population from 1980 to 1989 and from 1989 to 2011

Summary statistics	From 1980 to 1989			From 1989 to 2011		
	Larger catchment areas	Micro-regions	Larger municipalities	Larger catchment areas	Micro-regions	Larger municipalities
	Mean					
Log change in index of segregation	–0.01	–0.01	–0.01	0.10	0.11	0.07
Log change in number of schools	–0.03	–0.06	0.03	–0.19	–0.17	0.12
Log change in average school size	0.02	0.04	0.04	–0.32	–0.31	–0.59
Change in fraction of Roma students	0.00	0.00	0.00	0.10	0.09	0.06
	Standard deviation					
Log change in index of segregation	0.07	0.07	0.10	0.15	0.11	0.12
Log change in number of schools	0.19	0.15	0.22	0.38	0.30	0.32
Log change in average school size	0.18	0.14	0.23	0.41	0.28	0.34
Change in fraction of Roma students	0.02	0.01	0.03	0.09	0.09	0.07
Number of observations	175	174	140	175	174	140

References

CLOTFELTER, C. T. [1999]: Public School Segregation in Metropolitan Areas. *Land Economics*. Vol. 5. No. 4. pp. 487–504.

CLOTFELTER, C. T. [2004]: *After Brown. The Rise and Retreat of School Desegregation*. Princeton University Press. Princeton, Oxford.

HORN, D. [2012]: Early Selection in Hungary. A Possible Cause of High Educational Inequality. http://mta.academia.edu/DanielHorn/Papers/1646306/Early_Selection_in_Hungary_-_A_possible_cause_of_high_educational_inequality

KERTESI, G. – KÉZDI, G. [1998]: *A cigány népesség Magyarországon*. Dokumentáció és adattár. Socio-typo. Budapest.

KERTESI, G. – KÉZDI, G. [2011]: The Roma/non-Roma Test Score Gap in Hungary. *American Economic Review*. Vol. 101. No. 3. pp. 519–525.

Prisonization and/or Criminalization? Some Theoretical Considerations and Empirical Findings

Gábor Papp
Chief Councillor
Hungarian Central Statistical Office
E-mail: Papp.Gabor@ksh.hu

The study focuses on the correlations between criminalization and prisonization. On the grounds of earlier works, it shows that these two phenomena can be very hard to be separated from each other at theoretical level; they partly overlap. The empirical part of the article is based on a research conducted in a Hungarian medium- and maximum-security prison in the spring of 2010. It attempts to find the answers to the questions whether some attitudinal indicators of criminality and the nonconformity toward the staff expectations (as a frequently used indicator of the prisonization) are associated with one another or not; and if so, to what extent. The study also investigates whether criminality and prisonization can be related to the same factors, which may indicate the non-separable nature of the two concepts at the empirical level as well.

KEYWORDS:
Penalty.
Crime.
Criminal statistics.

One of the most basic questions of criminology concerns the role prison plays in criminality, as long as it has any remarkable impact at all. The widely used notion regarding this issue has been that prisons are the "schools of crime" in some respect since they intensify the criminal world-view of inmates. Hence, obviously these institutions are counterproductive for rehabilitation and re-socialization goals. This introduction may seem oversimplified; however, much of the research has dealt with the topic during the past decades. After all, the concept of the "schools of crime" has remained only an idea until nowadays. The methodological tools used by earlier researches have not been able to clarify the basic questions, and the theories were simply accepted with the recognition that prisons are harmful to society. Theorists do not appear to re-consider this statement. We do not know that if we were to indicate any associations between prisons and criminality, which factors would be found to influence this connection and in what context we could interpret it. The concepts of prisonization and criminalization, used in the title of this study, have surfaced in relation to this issue when thinking about prisons. The present study does not seek to support the "schools of crime" notion, but rather attempts to discuss the relationship between prisonization and criminalization based on data from a Hungarian empirical research.

1. Theoretical background

The authors of earlier studies were not consequent in considering these two concepts/ processes separated or closely interrelated. First of all, it is advisable to briefly review how prisonization and criminalization were treated at the level of theories, theoretical constructions, and measures.

At the level of theories it is worth going back to the beginnings. There is a semi-sentence of *Donald Clemmer* [1940 p. 299] on his definition of prisonization quoted in almost every work concerning the prisonization phenomena.[1] Moreover, the "universal factors" of prisonization have also been mentioned by most authors that quote Clemmer. At the same time, Clemmer's work is about a lot more than the issue of prisonization. He discusses it in a more nuanced way regarding both its implications and its association with criminalization. Although his narrow definition of prisoniza-

[1] "…the taking on, in greater or less degree, of the folkways, mores, customs, and general culture of the penitentiary."

tion does not contain dogmas, he mentions them as one of the objects of "taking on" and he asserts that their acceptance has a crucial significance in the process of prisonization. In this context, Clemmer mentions different opinions, attitudes regarding prisons, judges, and the police as parts of prison culture. He mentions the following examples of dogmas: negative attitudes towards parole board and government officials, distrust in and hate for prison guards, and finally believing that money is the universal solution. He states that prisoners' dogmas are harmful to themselves and to society, since they inhibit post-prison reintegration. Concerning the relationship between criminality and prisonization, there are two approaches in his writings. On the one hand, he asserts that one of the consequences of prisonization can increase criminality: "The phases of prisonization… are the influences which breed or deepen criminality and antisociality and make the inmate characteristic of the criminalistic ideology in the prison community." (*Clemmer* [1940] p. 300.) On the other hand, it seems that he regards this connection more as a potential than a necessary consequence: "No suggestion is intended that a high correlation exists between either extreme of prisonization and criminality. It is quite possible that the inmate who fails to integrate in the prison culture may be and may continue to be much more criminalistic than the inmate who becomes completely prisonized. The trends are probably otherwise, however, as our study of group life suggests" (*Clemmer* [1940] p. 302.).

One of Clemmer's contemporaries, *Reimer* stated the following about the "right guy" inmate role: "These men are so known because of the consistency of their behavior in accordance with the criminal or prison code…", as well as that: "…the 'right guy' to be definitely opposed to the law and its enforcement and the institution itself…" (*Reimer* [1937] pp. 152–153.) Another pair of authors posited that: "In the prison community, the chronic hostility between cons and screws – to some extent an extension of the progressive conflict between criminals and police on the outside…" (*Hayner–Ash* [1940] p. 579.)

If we have a look at the former trains of thought, we can get confused regarding the relationship between prisonization (and the inmate code on which it is based) and criminality. The basic rationale behind Reimer's comment is that he put an equal sign between criminal and inmate code. *Hayner* and *Ash* assumed a kind of continuity between the opposition to institutional officials (as a main component of the inmate code) and negative attitudes toward criminal justice system prior to the entry to prison, treating the former as originated in the latter. In his "narrow" definition Clemmer discussed changes of attitudes during prison sentence and he also mentioned external worldviews regarding the taking-on of dogmas. Some elements of dogmas are fairly close to favourable attitudes towards criminality. This is particularly important to note since it is assumed that the concepts of prisonization and criminalization are hard to distinguish from each other even at a theoretical level; their contents are overlapping in many respects. It seems that, according to earlier

theories, there is a sort of conceptual contamination regarding the relationship between prisonization (or rather the main tenets of the inmate code) and criminalization. Moreover, besides the "present tense" of the prison period, the argument points out the "past" and "future" as well. All this may explain why later researchers could translate this theory into the language of theoretical constructions and empirical research only imperfectly.

A later researcher, *Ohlin* emphasized the determination by past experience concerning inmate code: "This code represents an organization of criminal values in clear-cut opposition to the values of conventional society, and to prison officials as representatives of that society. … The code incorporates most of the values and orientations which inmates have shared in their criminal activities in the free community.… The prisoners' code reflects and adaptation of this criminal value system to the conditions of prison life." (*Ohlin* [1956] p. 28.) He was also the one who started to (mis)interpret Clemmer's prisonization theory and did not really manage to construe the attributes of prisonization theory in full compliance with the author's intention. According to Ohlin: "As Clemmer employs the term, prisonization reflects a continuous acculturation and *assimilation to the criminal values system* and the prisoner code of the inmate community." (*Ohlin* [1956] p. 39.) Likewise, *Wheeler* – who also interpreted Clemmer's thoughts about prisonization – asserted that "The net result of the process was the internalization of a *criminal outlook*, leaving the "prisonized" individual relatively immune to the influence of a conventional value system." (*Wheeler* [1961] p. 697.) In the same study, Wheeler's expression "*commitment to a criminal value system*" in connection with prisonization is also telling. At the same time it should be noted that this author, regarding the importation model (that he labeled "negative selection" model), unambiguously declared his standpoint about the relationship between criminalization and prisonization: "Their criminal acts indicate in varying degrees an opposition to conventional norms. It follows that the inmate culture should give expression to the values of those who are the most committed to a criminal value system – the long termers, those who have followed systematic criminal careers, etc. And if the culture is viewed as an outgrowth of the criminogenic character of inmates, it is reasonable to expect a reinforcement process operating throughout the duration of confinement." (*Wheeler* [1961] p. 708.)

However, another author, *Glaser* emphasized the interdependence of the two processes. He asserted that the adaptation to prison conditions (thus committing to the inmate code) is rather temporary, and it is far from certain that it has any impact after being released from prison (*Glaser* [1964]).

Later researchers aimed at translating these theories into theoretical constructions by interpreting the original ones. One of them carried on the "prisonization equals criminalization" approach. A good example for this is *Faine*'s study. It obviously reveals that he regarded the two concepts exchangeable both theoretically and practi-

cally (*Faine* [1973]). Although the title of his study includes the term "prisonization", his measures did not refer to this. Instead Faine took "inmate reference group orientation" as the indicator of prisonization, which he considered the "long range impact of institutionalization". Later, a similar practice was followed by *Walters* too, whose study contains the term "prisonization" but he measured criminal thinking and identity by his variables (*Walters* [2003]).

Another approach treated criminalization and prisonization (or its certain indicators) separately. Its one subtype considered both as juxtaposed components of a given group of views. *Bondeson* exemplifies this approach, constructing three scales to measure criminality: inmate solidarity, argot knowledge and criminality scales. Within criminality, Bondeson differentiated the sub-scales of criminal ideology, criminal association, and criminal identification (*Bondeson* [1989]). The common label "criminality" can be somewhat misleading in this instance, because only the last mentioned part of that is closely connected with criminality, while the two other scales rather relate to the taking-on of the institutional value system (for example prisonization). This is fairly similar to the approach of another author, *Schwartz*, who labelled "criminal value-orientation", "conformity to the inmate code" and "peer identification" scales with the common name of "inmate perspectives" (*Schwartz* [1971], [1973]). Thus, in the case of Schwartz, the principles of juxtaposition and parity predominate. Although he distinguished between criminalization and prisonization, he considered them as parts of the same group.

A different perspective predominates in the publications of *Thomas* and his associates. They mixed up and investigated "cause and effect" relationships in their studies, following the logic of temporal arrangement. Although the importation model was mentioned by Thomas in several works of him, stating that the components of the criminal value system are rooted in pre-prison socialization, it was the "consequence" approach that dominated in his theoretical models (*Thomas* [1977a], [1977b]; *Thomas–Hyman–Winfree* [1981]; *Thomas–Petersen* [1977]; *Thomas–Petersen–Cage* [1981]; *Thomas–Poole* [1975]; *Zingraff* [1975]).[2] Thomas and his colleagues discuss "short-term" and "long-term" consequences of prisonization (or rather imprisonment), their variables were defined accordingly in their models. They classified the opposition to prison and the priority of interpersonal relationships with other inmates as short-term, criminalization (or rather criminal identification) and attitudes toward the law and justice system as long-term consequences. Here we are interested in the last two attitudes. The approach of "long-term consequences" was jus-

[2] It must be noted that his publications are characterized by ambiguity concerning the consequences of what he is writing about. Thomas uses the expressions of "consequences of confinement" (or of imprisonment) and "consequences of prisonization" simultaneously. Logically, this can be accepted only if we consider imprisonment and prisonization equal, thus we accept that imprisonment cause prisonization as a logical necessity. If we adopted this perspective, it would be unnecessary to investigate the prisonization phenomenon.

tified by a logical slip. For example: "…the adoption of attitudes and values that increase the likelihood of reinvolvement in criminality upon release from the institution" (*Thomas* [1977a] p. 58.), or elsewhere: "…the greater the degree of criminal identification, the greater the probability of criminal involvement after release" (*Thomas–Foster* [1972] p. 230.). It proves that Thomas hypothesized an essential connection between intra-prison thinking and post-prison law-breaking behaviour.

Compared to the above mentioned works there is an inverse logic in the study of *Rhodes* [1979]. In his theoretical chapter, he apparently took on the "consequence" viewpoint, but at the same time he emphasized the "antecedent" role regarding both opposition to the law and justice system and criminal identity variables (which he treated separately). However, it is noteworthy that only the importation aspect is used in the analysis. The tested model of another researcher, *Alpert* [1978] was based on the "attitudinal and ideological import" approach, whereby he investigated the role of attitudes toward justice system (besides different kinds of alienation) as imported views.

Finally, we need to mention another trend developed almost simultaneously with the start of prisonization research. It focused on inmates and, instead of investigating the conformity-nonconformity dichotomy, it studied their attitudes towards criminal justice and attitudinal changes (*Watt–Maher* [1958], *Hulin–Maher* [1959], *Mylonas–Reckless* [1963], *Cleaver–Mylonas–Reckless* [1968], *Maher–Stein* [1968], *Mylonas–Cleaver–Reckless* [1968]). The aim of these authors was the same as that of prisonization researchers: they considered the prison as a "black box" which generates changes in the thinking of inmates and inhibits their reintegration into society.

2. About the reduced model and the research

One common feature of the aforementioned studies is that they failed to answer at least two questions: what changes occur in prison and what are their consequences for the post-prison period. One of the main reasons for this is that prisonization studies were mainly based on cross-sectional design. Besides, there were some panel studies which attempted to reveal changes in thinking only during the course of several months (*Glaser* [1964], *Alpert* [1978], *Bondeson* [1989]). The cross-sectional studies explored the consequences of imprisonment purely logically. Frequently, it was declared in the theoretical part of these works that certain variables would be viewed as consequences (see for example the writings of Thomas). Another technique was creating certain synthetic groups by taking the time spent and time remaining into account (for example "early" and "late career phases", see *Wheeler* [1961]). In theory, panel studies would be suited for revealing changes during and af-

ter the prison period; however, it is rather illusory to think that these processes occur only in a few months. Cross-sectional studies are even less appropriate for answering such questions, but they can be suitable for other purposes.

In my own cross sectional study, I treated the different attitudes as what they really are, in other words, juxtaposed with each other, having not causal but correlational relationship. Hence I investigated the correspondence between certain indicators of prisonization and criminalization. That is to say, I did not regard them either as imported thinking patterns or ones that prevailed following the imprisonment. I exclusively considered them as patterns that were present right at the time of imprisonment. In some respects this approach means a step back, compared to earlier ones because the very components are missing from the model which would replace theory and which, in fact, typically justify the scientific investigation of prisonization, that is, the ones related to the effects of prison. Since – as I have already mentioned – there was not accordance between the theory and empirically measured phenomena in earlier studies, it is necessary to move back. The cornerstone (and maybe the advantage) of my approach is not showing more than the conclusion that we can draw from the data. However, it would be also important to explore the causes and potential consequences, but it should be emphasized again that we cannot venture on this due to the limitations of the current research.

The present study looks for the answer to three questions. *Firstly,* to what extent the different components of criminal world-view are associated with one another and whether they are integrated into a coherent system. *Secondly*, to what extent they are correlated with the possible indicator of prisonization, in other words, with the rejection of staff expectations. The *third* question awaiting answer is that which factors influence the criminal views and prisonization, and whether these results show a uniform pattern.

To answer these questions, I used data from a survey which was conducted in Vác Maximum and Medium Security Prison in Hungary, in March 2010. The aim of the study was to explore attitudes of the total population of adult male inmates with a definitive sentence in this institution. The number of potential respondents was 618 at the reference date (1 March). The questionnaires were properly filled out by interviewers in 365 cases, thus the response rate was 59.1 percent. It should be noted that there was no preliminary sampling in the prison. Posteriorly, however, it was possible to liken to the total inmate population with respondents of the study using data from institutional records (of course, observing the rules of anonymity and personal rights). On the grounds of this comparison, it seems that the respondent population is representative to the total inmate population at an acceptable level. There were no significant differences between the total and the sub-populations with regard to such factors as security and custody levels, age, marital status, nature of committed crimes, doing a job in prison or not, length of sentence. The Goodman–Kruskal's

gamma (γ) coefficients were used in this analysis, in accordance with the ordinal level of measurement of the examined variables (*Goodman–Kruskal* [1954]).

3. Analysis and findings

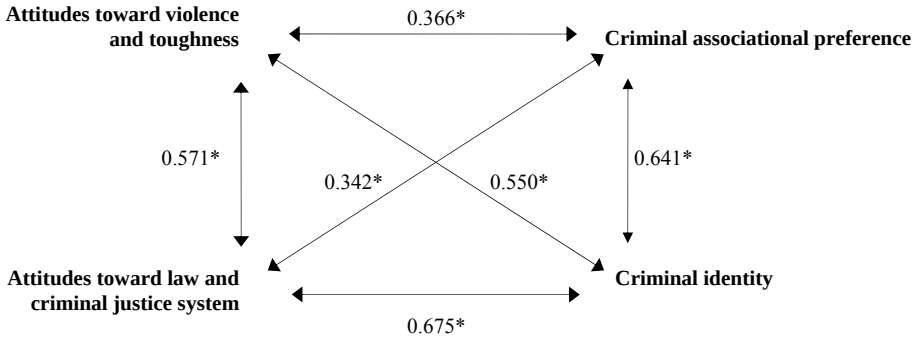
Before we discuss the analysis in detail, it is important to note that the expression of "criminalization" in the title of this study is actually incorrect (as I have mentioned already), since criminalization refers to a process, while a cross-sectional survey can show only a "snapshot". Therefore, it is more proper to talk about criminal world-view and its indicators, which were constructed in conformity with the previous literature. Accordingly, regarding the components of the criminal world-view, the following variables were distinguished by principal component analysis: attitudes toward relationship with criminals (criminal associational preference), acceptance or rejection of self-definition as a criminal (criminal identity), opinions on the law and criminal justice system, and attitudes towards violence and toughness.[3] The total scores of these variables were trichotomized. Value 3 indicates that the inmate accepts the criminal value orientations, 1 is the opposite, and 2 is the medium category.

Associations between the different criminal views and their relations to prisonization

The first hypothesis – stating that criminality indices there are interconnected – points to the fact that these variables are interchangeable, they may reflect the same phenomenon. The data presented in Figure 1 shows that all the relationships between the indicators of procriminal views are positive and significant. On the grounds of moderately strong associations, it is obvious that the delinquent self-identification also means that one has got a negative orientation towards law and the justice system, as well as this type of identification implies supporting close relationships with criminals. Similarly, the criminal identity is closely associated with attitudes towards violence and toughness, while the latter one is connected with the opposition to the law. Being less tightly related to the aforementioned, yet the views on law and criminal justice do correlate with the level of accepting or rejecting criminal friendships. This is accompanied by opinions supporting violence. These findings suggest that there is a relatively coherent criminal value system amongst inmates.

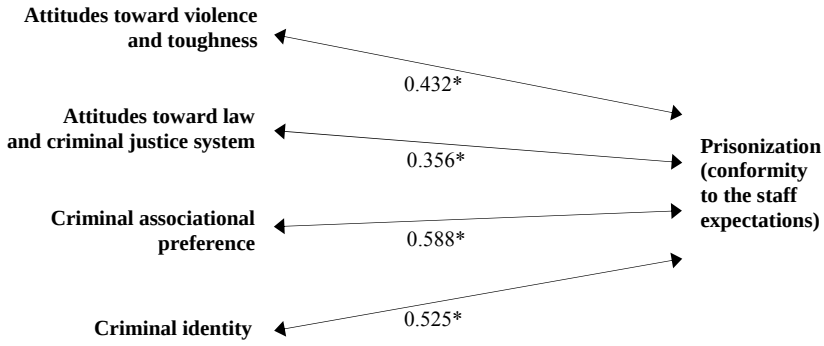
[3] The items used to operationalize these variables are included in the Appendix B.

Figure 1. Associations between criminal views
(gamma coefficients)



* $p < 0.001$.

Figure 2. Associations between criminal views and prisonization
(gamma coefficients)



* $p < 0.001$.

The second hypothesis pertains to the relationship between institutional nonconformity (prisonization) and each component of criminal world-view. Prisonization was measured with the use of a tool developed by *Wheeler* [1961]. It was based on some hypothetical conflict situations in which the respondents have to decide between two alternatives. One option represents the expectations of the prison staff, the other those of inmate society.[4] I also applied Wheeler's guidelines when constructing the variable. That is, if the respondent agreed with the answers that represented the

[4] The situations can be found in Appendix A. They were almost identical with the ones used by Wheeler, with the exception that actors had been given Hungarian last names, as well as in the case of the "hiding" story the respondents did not have to decide about money but Rivotril pills (which means higher risk-taking).

supposed staff expectations at least in four out of five situations, he was categorized as "conform" inmate. Those who answered this way in zero or only one situation became the "non-conform" (or "prisonized") type. The latter was coded with 3, the former with 1, and the medium category with 2. The data shows that the non-conform answers are more prevalent among those who consider themselves as criminals, who did not seclude themselves apart from close associations with criminals, who has got violent attitudes, and who is up against the law. (See Figure 2.)

The third hypothesis is linked to the factors associated with antisocial perspectives: implications of the criminal past and the prison period. Since the parts of criminal world-view are undoubtedly associated with each other as well as with prisonization, it is worth exploring that to which factors these thinking components can be traced back or whether they can be described with a uniform pattern. It is important because if these components are related to certain background variables in similar vein, this would indicate that criminal thinking and nonconformity towards institutional staff expectations cannot be distinguished from each other, thus these two can be considered one single phenomenon.

Regarding nonconformity and the components of criminal world-view, I investigated the role of those factors that had been used in earlier studies. These variables can be divided into two "rough" groups according to their place in time (past or present). The reason of using the word "rough" is justified by the following reason. It is hard to decide which time level is represented by certain variables, similarly to establishing what phenomenon is shown by the given indicator. The variables that represent the past are casted into two groups. One of them refers to criminal past (juvenile arrest, having a clean record or not) and the other to the conventional aspects of the past (educational attainment, school behaviour and scholastic records, truancy, longest time spent in a given job). It is loosely related to the aforementioned group of variables that indicate the type of offense: violent crimes (or not) and sexually oriented crimes (or not). Since I investigated a homogenous population regarding gender (there were only adult male inmates in the prison), the demographic viewpoint was limited to age. The other group of variables refers to the "present tense" of imprisonment. The security and custody level, whether he works in prison or not, together with prison rewards and punishments fell into this group because this research was conducted in one institution only. The table contains a summary of the associations between these factors and attitude variables.

a) Factors related to the past. The assumption concerning criminal past is that the earlier and more often one has got into contact with the agents of social control or the representatives of the criminal justice system, the more probable that it has strengthened his pro-criminal views. However, the opposite is not likely to happen on a theoretical basis. Our data seem to support this hypothesis, showing that people who were arrested on suspicion of a crime as juveniles (aged 14–17) tend to adopt crimi-

nal interpersonal relations and stronger criminal identity, they are up against the law and its representatives, they accept violence and support nonconformity in prison. However, it must be noted that juvenile arrest is a "soft" indicator in some respect, since in itself it does not necessarily mean that the person committed a crime. The results connected to the other criminal past variable (namely whether he has ever been in prison) are consonant with that of juvenile arrest. Accordingly, in contrast with first-termers, those who have ever been imprisoned before the current sentence accept pro-criminal views much more and stand clearly in opposition to staff expectations. The unambiguous findings suggest that a criminal maturation process (in other words criminalization) do exist. But it must be kept in mind that these are official indicators. That is, whether a person gets arrested and later convicted are determined partly by how the criminal justice system operates, partly by the extent of how "skilful" or maybe "lucky" the criminal is, or rather by the interplay between these factors. Consequently, being arrested and sentenced can be an indicator of the extendedness of one's criminal history. Moreover, if the individual had not been caught by investigator authorities, it can be the best indicator of his criminal potential. Thus in this case we are groping in the dark, since the presence or absence of prior arrest or conviction would equally indicate strong and weak connection to crime.

Perhaps it is not an exaggeration to state that there is only one conventional mobility channel in our age: school. Schooling was measured by official educational attainment as well as subjective indicators such as scholastic records, school behaviour, and truancy. The explanatory mechanism behind is that early broken or fragmented school career leads the later perpetrators to alternative ways of living, to crime, which very likely generate some changes in their way of thinking. As can be seen from data, this hypothesis was not unfounded either, since favourable attitudes toward crime and opposition to the demands of the institutional staff are more prevalent among the poorly educated inmates and the ones who defined themselves as frequent truants. The coefficients that show the relationship between school behaviour and certain antisocial views point in the same direction. However, the past scholastic records are surprisingly only associated with negative attitudes toward law and prisonization, while in the cases of the three other types of views the coefficients are weak and not significant. We can only assume what may stand in the background. Questions about past school behaviour and scholastic records seem to relate to the same "unit", but this is only an artifact. The former refers to a more or less narrower phenomenon: following or breaking regulations concerning teachers and fellow students. However, regarding the latter, the answers are based on more complex experiences, which would lead to some kind of distortion. Another reason for the lack of associations can be that school failures have different significance in various social groups; in certain ones, they can be highly respected instead of stigmatization. Due to subjective retrospection and the special conditions, the reliability of the answers about school records is questionable.

Not independently from past school experiences and educational attainment, it is worth looking at the relationship between criminal views and labour market performance. Although there were questions about the respondents' typical and last occupation in the questionnaire, the variables that were constructed from them have proved unusable. In fact, they should be treated as constants rather than variables, since almost all inmates fall into the category of unqualified blue-collar workers. That is why we utilized the continuity or discontinuity of the labour market position as an indicator, based on the longest time spent at a workplace. We confirmed the hypothesis that the shorter time an inmate spent at a given workplace, the more willing he is to accept some criminal views as well as the more hostile towards staff expectations. It is important to emphasize that "workplace" also refers to non-registered types of employment in this study. If it would be restricted to legal or registered jobs, presumably the findings would be even stronger.

The table also shows that the age of inmates is inversely related to the acceptance of criminal views and prisonization. In other words, the younger prisoners tend to accept such opinions that support close association with criminals, the use of violence and the rejection of the law and criminal justice system more than the older ones do. It seems there is a "fading" process with aging. The question about other factors that may influence this association arises again. It is fairly conceivable that the data do not indicate a kind of real prosocial maturation process but the strengthening of concealing, outwitting and manipulating the outside word. Thus there can be a latent "manipulation" factor which would influence the observed relationship. It is not impossible either that the younger and older inmates have committed different types of offences or they have dissimilar social backgrounds, which would explain why members of different age groups think differently.

The fact that a given person committed a certain kind of crime (for which he was sentenced to imprisonment) can indicate a kind of orientation to law, hereby to crime itself. The criminal cases are considerably diversified and grouping them – even if the penal code contains one classification – can be problematic. Within the framework of the present study, the crimes, or rather the perpetrators were categorized according to two criteria. One is based on the distinction between violent and non-violent crimes, while the other differentiates between sexually and non-sexually oriented crimes. The logic behind these distinctions is that violent crimes are at the top of the crime hierarchy, while sexually oriented crimes are at the bottom of that, and the position in this stratification is supposedly associated with the individual's own values and views. However, the findings suggest that the types of committed crimes in the past are not or only slightly linked with antisocial thinking. There is no significant relationship between the violent nature of crimes and the acceptance/rejection of the inmate perspectives; the values of the coefficients are almost zero. The lack of the latter association is interesting because the rate of the violent perpetrators is approximately 80 percent of

the total convicted population in the studied prison. There were questions also about attitudes toward violence and toughness, but the hypothesis was rejected again. A possible reason is that a part of violent crimes were motivated by passion or were non-intentional, thus they were not linked with an ideology that supports violence.

Committing sexually motivated crimes is associated with only some views in the predicted direction. So the opposition to staff expectations and accepting associations with criminals are less prevalent among those who committed sexual crimes than among other inmates. One's attitude towards relationships with law-breakers is probably not independent from his position in the prison. An inmate's opinions about his fellows can be shaped by others' general orientation toward him. The perpetrators of sexually motivated crimes (or even its suspicion) can be stigmatized by the public within the prison what can lead to loneliness. The items/statements of the associational preference scale explicitly refer to attitudes towards criminals and, inseparable from that, towards fellow inmates. This special socially excluded position within the prison may explain the similar strength and direction of the association between non-conformity and committing sexual crimes. The utilized situations are based on two important aspects of prisonization: the conflict between solidarity among inmates and, as its complementary, the opposition to institutional staff. Interestingly, however, the otherwise closely associated three other components (attitudes toward law, criminal self-identification, violent views) do not show any relationships with the sexual-based grouping. A tentative explanation for this can be that inmates who committed sexually motivated crimes have more contradictory attitudes towards the label of being a "criminal" than other prisoners. Although violence is a typical motif and often used also in the case of sexual crimes, the answers do not reflect that it was used as an intentional strategy. Similarly, the attitudes towards law and criminal justice were not influenced by the type of crimes committed in the past.

b) *Immediate context of the prison, factors of the present*. The study was conducted in an institution in which inmates serve their sentence at either medium- or maximum-security level. Security levels reflect the conditions of imprisonment. Actually, they are partly related to the present and partly to the past, because court sentences decide on them on the basis of criminal law that defines the conditions and seriousness of a crime. From at least two respects, it would be logical to assume that the inmates in more rigorous maximum-security sites tend to accept the investigated antisocial opinions. Firstly, the maximum-security punishment is characteristically imposed upon perpetrators of more serious crimes by the court, who are likely to have deeper and more criminal motivations. Secondly, the conditions of confinement are harsher than at medium-security level and this can make inmates responsive to accepting antisocial orientation (as stated by the deprivation model of *Sykes* [1958]). The data suggest that these assumptions do not hold. Security level has no influence on inmates' views. It should be emphasized that inmates in either medium- or maxi-

mum-security units are housed separately. These units are theoretically different but practically not or only imperceptibly. Nevertheless, it is also possible that if we had got data from inmates housed in minimum-security prison (this kind of security level was not present in the investigated institution), we would have succeeded in showing significant relationships.

Another possible comparison would be based on inmate custody level which is not determined by the past but the present behaviour in the institution. Custody levels range from 2 to 4, where the numbers refer to the harshness of circumstances. At admission, every inmate is classified into medium level (level 3), and in the case of good behaviour or low-level security risk and other requirements, the inmate gets into level 2, while in the instance of high-level risk, into level 4. The majority of inmates in the studied prison generally stay at medium level, the rates of downwardly and upwardly mobiles are 7-8 percent, respectively. Since this classification regarding custody levels takes inmate thinking and general orientation into consideration, we can assume that stricter custody level goes together with higher level of acceptance of criminal views and of nonconformity. This hypothesis is basically supported by the data. Although there is a significant relationship in only one case, associations with the other four inmate perspectives point in the same direction.

Whether an inmate has a job in the institution can also indicate his attitudes towards getting on in conventional life, that is, towards the opposite of crime. The findings regarding this assumption are considerably contradictory. Violent thinking and prisonization are not at all affected by participating in prison work activities. The acceptance of criminal self-conception and opposition to the law are more typical amongst non-workers, although the relationships are not too strong. On the contrary, workers give priority to contacts with criminals more often. These results raise the question that how wide the "range" of these criminal associations is, with whom they want to maintain closer relationships. Since the working inmates are housed in separate living units and they work with their cellmates or with other inmates from the same living unit, it would be logical that these individuals would be their potential post-release associations. It may explain the negative sign of the coefficient.

Institutional behaviour, more specifically rule infractions and rewards can be important indicators of the position held by prisoners within the institution. The higher number of the former and the lower number of the latter can be indicative of the fact that inmates have such a way of thinking that facilitate rather than hinder criminality. In the case of few punishments and many rewards, we can hypothesize the opposite. Previously, rule infractions (institutional punishments) were frequently utilized as the indicators of the behavioural aspect of prisonization. However, the role of institutional rewards was not investigated. Within the framework of the present research, it was possible to match questionnaire data with institutional records. The problem was mainly with the construction of these behavioural variables. The time factor has a

significant impact on both of them, as it does matter how long it takes for someone to attain a certain amount of rewards or punishments. To this end, the number of punishments and rewards were divided by the total time spent from the present sentence (in months). The derived values were grouped in two different ways. (See the table.)

Factors associated with the measures of criminal views and prisonization
(gamma coefficients)

Variables investigated	Criminal associational preference	Criminal identity	Attitudes toward the law	Attitudes toward the violence	Prisonization
Juvenile arrest (no, yes)	0.400**	0.549**	0.335**	0.327**	0.397**
Number of times he has been in prison (once, more)	0.134	0.434**	0.274**	0.287**	0.246**
Educational attainment (completed 0–7 grades, 8 grades, high school)	–0.120	–0.295**	–0.265**	–0.106	–0.132*
Scholastic records (good, medium, bad)	0.122	0.045	0.224**	0.044	0.166*
School behaviour (good, medium, bad)	0.386**	0.176*	0.211**	0.175*	0 .385**
Truancy (no, yes)	0.428**	0.497**	0.340**	0.200*	0 .455**
Longest time spent at a given workplace (0–1 months, 1–6 months, 6 months–1 year, more than 1 year)	–0.399**	–0.497**	–0.316**	–0.321**	–0.445**
Age group (18–29, 30–39, 40–49, 50–X years)	–0.356**	–0.283**	–0.166*	–0.261**	–0.417**
Violent crimes (yes, no)	–0.038	–0.033	–0.019	–0.095	–0.156
Sexually oriented crimes (yes, no)	0.307*	0.133	0.004	–0.038	0.300*
Security level (medium, maximum)	–0.022	0.025	0.100	0.054	–0.021
Custody level (2, 3, 4)	0.195	0.345**	0.202	0.163	0.155
Whether he works in prison (yes, no)	–0.159	0.201*	0.133	0.068	0.028
Rewards (yes, no) – absolute	0.067	–0.134	–0.102	–0.074	–0.018
Punishments (yes, no) – absolute	0.244*	0.223*	0.159	0.131	0.135
Rewards (few, medium, many) – relative	0.045	–0.282	–0.206	0.075	0.019
Punishments (few, medium, many) – relative	0.270	0.294*	0.401**	0.359*	0.287*

* $0.01 \le p < 0.05$.
** $p < 0.01$.

On the one hand, absolute measures were developed in which the inmates were classified into two categories: those who got at least one reward and those who re-

ceived none at all. We followed the same procedure for punishments. On the other hand, since these are fairly "rough" indicators, relative measures were also constructed. In the case of both rule infractions and rewards, the values were trichotomized, hereby the constructed variables refer to relatively few, medium and many rewards (and punishments). The assumption is that punishments are more frequent among those who prefer criminal views and nonconformity, while institutional rewards are more frequent in the group of those who respect law and staff expectations and those who reject the criminal value system. These hypotheses – irrespective of using either the absolute or the relative measures – were only supported in the instance of rule violations. The associations were moderately strong in the case of the three-way relative categorization and weaker when using the absolute measures. It can be concluded that the relative number of rule infractions are higher among those who consider themselves as criminals, clearly oppose the law and the criminal justice system, support violence and reject staff expectations. The acceptance of close relationships with criminals is also associated with the higher number of rule violations, although this result is statistically not significant.

There are no such relationships regarding institutional rewards, although in some cases (criminal self-identification, negative attitudes towards the law) the values of coefficients are between 0.2–0.3 and negative in the three-way categorization. It suggests that inmates who accept pro-criminal views received relatively more rewards. However, it must be noted again that these associations are not significant. In other cases, the values of coefficients are very close to zero. It would be reasonable to suppose that reward and punishment are opposites of each other, but it seems they are not. In order to clarify this relationship, we would need to understand the mechanisms that are at work when imposing rewards and punishments. For example, it is important that (in principle) rule violations are not limited in a given period, while imposing rewards are. It is very likely that negative and positive sanctions are initiated by different actors and through different "channels": the former falls within the cognizance of custodians, the latter within that of the treatment staff. Another important question is whether the same "threshold" level applies to rewards and punishments. It is fairly conceivable that in a stricter prison it is easier to commit rule infractions than to get rewards.

4. Summary and discussion

The data presented in this study shows that there is some kind of combination of procriminal views among inmates, which are closely associated with nonconformity

to institutional staff. Even though the theoretical model applied in this study makes a distinction between the rejection of staff expectations and the components of the criminal value system, our findings suggest that for the most part the same factors influence both of them and in the same direction. So inmates with more extensive criminal background, lower education, only temporary labour market experience, stricter custody level, and more instances of previous incarcerations are more often characterized by criminal and non-conform views. These findings – even if partially – would take us closer to the topic of "the schools of crime". On the one hand, it is obvious that criminals are kept in prison, so the widespread acceptance of the criminal value system is not unexpected. On the other hand, results concerning pre-prison indicators suggest that the roots of criminal views go back to the period prior to the entry into the institution. Unfortunately, the current investigation leaves an important question unanswered: to what extent prison forms and deepens the already existing value system. However, it is very unlikely that general nonconformity of both the pre-prison and prison period would essentially change after inmates will be released.

Regarding prisons as crime-intensifying institution, it would be necessary to conduct a research (following the logic of impact analysis) which would aim at exploring control and experimental groups by a follow-up study. It would cover – at least in theory – such criminals who have never been in prison. It is the only way to distinguish general temporal changes in criminal views from the real effects of imprisonment. The investigation should hypothetically start before the occurrence of the crime itself, which, certainly, is absolutely impossible. Hence it appears that the concept of the schools of crime remains an open question. Maybe even the question itself is wrong, since the real motives of crime should not be sought inside the prison but outside its walls.

Appendix A

Hypothetical conflict situations used to measure the prisonization are enumerated below.

Conformity to staff role expectation

1. An inmate, *Gulyás* is working as much as he can in the institution. Some other inmates threaten him because he does more work than anybody else in the crew. He works as hard as he can just like earlier.

2. Inmate *Szabó* goes before a committee that makes job assignments. He is given a choice between two jobs. One job would call for a hard work, but it would give Szabó training that might be useful to him "outside". The other would allow him to do easier time in the institution. But it provides no training for a job outside. Szabó decides to take the easier job.

3. An inmate, without thinking, commits a minor rule infraction. He is given a "write-up" by a correctional officer who saw the violation. Later three other inmates are talking to each other about it. Two of them criticize the officer. The third inmate, *Fodor* defends the officer, saying the officer was only doing his duty.

4. Inmates *Deák* and *Budai* are very good friends. Deák has 50 pieces of Rivotril pills that were smuggled into the institution by a visitor. Deák tells Budai he thinks the officers are suspicious, and asks Budai to hide the pills for him for a few days. Budai takes the pills and carefully hides it.

5. Inmates *Kocsis* and *Pintér* are planning an escape. They threaten inmate *Szűcs* with a beating unless he steals a rope for them where he works. While he is trying to smuggle the rope into the cell house, he is caught by an officer, and Szűcs is charged with planning to escape. If he doesn't describe the whole situation, may get into serious trouble. He can avoid it by blaming Kocsis and Pintér.

Appendix B

Items used to measure the major attitude variables are listed below.

Attitudes toward law and justice system

It's all right for a person to break the law if he doesn't get caught.

The only bad thing about breaking the law is the chance of getting caught.

It's hard to have much respect for the law after I think about how I've been treated by the police and court.

It's all right to bend the law as long as you don't actually break it.

It's hard to have much respect for the law after I think about how I've been treated by people who are supposed to support the law.

There's nothing wrong with breaking the law as long as nobody gets hurt.

Stealing is just another way to make a living.

Laws are so often made for the benefit of small selfish groups that a man cannot respect.

The law is more rotten than core.

A hungry man has the right to steal.

Laws are the enemy of freedom.

Laws are for the poor to obey and for the rich to ignore.

A person should obey only those laws which seem reasonable.

Criminal identity

I would rather lead a life of adventure and dishonesty than that of a law-abiding type with a regular job.

I'm more like people who are after easy money than I'm like people who grind away at a job.

It is better to do a few illegal things to make money than work at a job with regular, fixed hours in the same place every day.

I'm more like the people who can make a living outside the law than I'm like those who only break the law occasionally.

I would define myself as a criminal.

I would like to be able to take things easy and not have to work hard.

The life of most people who follow the rules and have a steady job is dull.

People who have been in trouble with law are more like me than people who don't have trouble with the law.

A man is fool to work for a living if he can get by some easier way, even if it means violating the law.

I would define myself as a law-abiding person.[5]*

I think more like other inmates than people outside.

People who have been in trouble with the law have the same sort of ideas about life that I have.

You've really got to respect a guy who's smart enough to break the law and get away with it.

Criminal associational preference

I would rather associate with people who obey the law than those who don't.*

When I get out I don't want to associate with the kind of people that are always getting into trouble.*

I want to keep in touch with inmates I have met in here after I get out.

Upon my release, I will avoid all friends I have here.*

I don't care to associate with the kind of people that are in prison.*

Most of the friends I have made in prison are not like the friends I would make in the streets.*

The only kind of persons I take as a friend is one who respects the law.*

The kinds of guys I hang around with here are really a lot like most of the people I knew on the street.

I am friendly with a group of guys who work hard and feel that an inmate should try "to better himself" while in prison.*

Most of the inmates in here are like the people I ran with in the free world.

Attitudes toward violence and toughness

In order to survive in prison, an inmate has to establish a "tough guy" reputation.

The best way to get respect around here is to act tough.

You have to be hard to make it here.

If you ever do have to fight, you're wise to do a good enough job on the other guy that he'll never come back for more.

Knowing that you are tough is sufficient. You don't have to show it by force.

I believe in the use of force to overthrow the law.

There is never a good reason to use psychical violence no matter what the situation might be.*

It's not smart to look for trouble, but once it comes you can't back away from it and still be a man.

[5] Here and hereinafter * means reversed items.

References

ALPERT, G. P. [1978]: A Comparative Study of the Effects of Ideology on Prisonization: A Research Note. *LAE Journal of the American Criminal Justice Association.* Vol. 41. No.1. pp. 77–86.

BONDESON, U. V. [1989]: *Prisoners in Prison Societies*. Transaction Publishers. New Brunswick.

CLEAVER, P. T. – MYLONAS, A. D. – RECKLESS, W. C. [1968]: Gradients in Attitudes Toward Law, Courts, and Police. *Sociological Focus.* Vol. 1. No. 2. pp. 29–40.

CLEMMER, D. [1940]: *The Prison Community*. Holt, Rinehart and Winston. New York.

FAINE, J. R. [1973]: A Self-Consistency Approach to Prisonization. *Sociological Quarterly*. Vol. 14. No. 4. pp. 576–588.

GLASER, D. [1964]: *The Effectiveness of a Prison and Prison System*. Bobbs-Merrill. Indianapolis.

GOODMAN, L. A. – KRUSKAL, W. H. [1954]: Measures of Association for Cross Classifications. *Journal of the American Statistical Association.* Vol. 49. No. 268. pp. 732–764.

HAYNER, N. S. – ASH, E. [1940]: The Prison as a Community. *American Sociological Review*. Vol. 5. No. 4. pp. 577–583.

HULIN, C. L. – MAHER, B. A. [1959]: Changes in Attitudes Toward Law Concomitant with Imprisonment. *Journal of Criminal Law, Criminology, and Police Science*. Vol. 50. No. 3. pp. 245–248.

MAHER, B. – STEIN, E. [1968]: The Delinquent's Perception of the Law and the Community. In: *WHEELER, S.* (ed.): *Controlling Delinquents*. John Wiley and Sons. New York. pp. 187–221.

MYLONAS, A. D. – CLEAVER, P. T. – RECKLESS, W. C. [1968]: A Comparative Study of Attitudes Toward Law and Law-Enforcement Agencies in English-Speaking and French-Speaking-Canada. *Criminology*. Vol. 6. No. 3. pp. 30–40.

MYLONAS, A. D. – RECKLESS, W. C. [1963]: Prisoners' Attitudes Toward Law and Legal Institutions. *Journal of Criminal Law, Criminology, and Police Science*. Vol. 54. No. 4. pp. 479–484.

OHLIN, L. E. [1956]: *Sociology and the Field of Corrections*. Russell Sage Foundation. New York.

REIMER, H. [1937]: Socialization in the Prison Community. In: *Proceedings of the American Prison Association*. American Prison Association. New York. pp. 151–155.

RHODES, M. L. [1979]: *The Impact of Social Anchorage on Prisonization*. Unpublished PhD dissertation. A&M University. College Station.

SCHWARTZ, B. [1971]: Pre-institutional vs. Situational Influence in a Correctional Community. *Journal of Criminal Law, Criminology, and Police Science*. Vol. 62. No. 4. pp. 532–542.

SCHWARTZ, B. [1973]: Peer versus Authority Effects in a Correctional Community. *Criminology*. Vol. 11. No. 2. pp. 233–257.

SYKES, G. [1958]: *The Society of Captives*: *A Study of a Maximum Security Prison*. Princeton University Press. Princeton.

THOMAS, C. W. [1977a]: Prisonization and Its Consequences: An Examination of Socialization in a Coercive Setting. *Sociological Focus*. Vol. 10. No. 1. pp. 53–68.

THOMAS, C. W. [1977b]: Theoretical Perspectives on Prisonization: A Comparison of the Importation and Deprivation Models. *Journal of Criminal Law and Criminology*. Vol. 68. No. 1. pp. 135–145.

THOMAS, C. W. – FOSTER, S. C. [1972]: Prisonization in the Inmate Contraculture. *Social Problems*. Vol. 20. No. 2. pp. 229–239.

THOMAS, C. W. – HYMAN, J. – WINFREE, T. L. [1983]: The Impact of Confinement of Juveniles. *Youth Society*. Vol. 14. No. 3. pp. 301–319.

THOMAS, C. W. – PETERSEN, D. M. [1977]: *Prison Organization and Inmate Subcultures*. Bobbs-Merrill. Indianapolis.

THOMAS, C. W. – PETERSEN, D. M. – CAGE, R. J. [1981]: A Comparative Organizational Analysis of Prisonization. *Criminal Justice Review*. Vol. 6. No. 1. pp. 36–43.

THOMAS, C. W. – POOLE, E. D. [1975]: The Consequences of Incompatible Goal Structures in Correctional Settings. *International Journal of Criminology and Penology*. No. 3. pp. 27–42.

TORO-CALDER, J. – CEDEÑO, C. – RECKLESS, W. C. [1968]: A Comparative Study of Puerto Rican Attitudes Toward the Legal System Dealing with Crime. *Journal of Criminal Law, Criminology, and Police Science*. Vol. 59. No. 4. pp. 536–541.

WALTERS, G. D. [2003]: Changes in Criminal Thinking and Identity in Novice and Experienced Inmates: Prisonization Revisited. *Criminal Justice and Behavior*. Vol. 30. No. 4. pp. 399–421.

WATT, N. – MAHER, B. A. [1958]: Prisoners' Attitudes Toward Home and the Judicial System. *Journal of Criminal Law, Criminology, and Police Science*. Vol. 49. No. 4. pp. 327–330.

WHEELER, S. [1961]: Socialization in Correctional Communities. *American Sociological Review*. Vol. 26. No. 5. pp. 697–712.

ZINGRAFF, M. T. [1975]: Prisonization as an Inhibitor of Effective Resocialization. *Criminology*. Vol. 13. No. 3. pp. 366–388.

Practical Examples of Key Index Numbers Measuring Market Domination Abuse in the Electricity Sector

András Sugár
Associate Professor
Corvinus University
of Budapest
E-mail: andras.sugar@uni-corvinus.hu

The study focuses on two index number groups: the index numbers of concentration in competition law practice and the Lerner index. By means of the example of electrical energy, the study illustrates the use of index numbers, and evaluates the advantages and disadvantages of those. Primarily, provided the illustration of the power station market, the study argues that a raw analysis of data might come as misleading in drawing consequences, as a stand against the market power inherently alters the structure of the market analyzed. The study involves the effects of import and the arrangements motivating regulative competition policies and, thus, provides a more realistic picture on market power in the sector examined.

KEYWORDS:
Market power.
Concentration and measurement of concentration.
Hirschmann-Herfindahl index (HHI) and Lerner index.

In competition policies, the notion of market power is a fundamental category. Practically the majority of market competition law regulations focus on this; therefore, its measurement is of paramount importance (*Bishop–Walker* [2010]).

Several ways exist in which market power can be defined and understood. Here we use the definition of market power as the capability of companies to sustain their prices for prolonged periods of time above the competition (*Bishop–Walker* [2010]). Hence this phenomenon is interpreted in relation to effective competition, whose measurement is again not trivial.

Market power may demonstrate itself in a number of ways, wherever the need for its statistical quantification emerges. This article considers a few of the key indicators, focusing on their measurement and problems of interpretation from an economic perspective.

The phenomena examined and their measurements are: *1*. market shares and the extent of concentration; *2*. the general market power measuring figure, the so-called Lerner index, the capability of raising price above the marginal cost.

Prior to studying these two areas in depth, it is worth mentioning and linking two (statistically measurable) elements. Market power partially emerged from monopolistic or oligopolistic market competition, that is, a monopoly has the ability to sustain high prices over a prolonged period of time. Measurement and discussion of market domination abuse is therefore linked to the number of companies functioning on the market and their absolute and relative concentration. At the same time, a fairly small number of competitors are able to generate tense competition, making raising prices improbable. (A classical example is the case of Coca Cola's and Pepsi Cola's competition, but the tight race that evolved on the mobile telephone market at a relatively small number of competitors is far less likely to occur on the landline market.) The latter is clearly linked with the naturally monopolistic setup of providing such service, but this article does not cover this area (*Sugár* [2011b]). Thus, one element of market domination is monopoly, the degree of concentration that *limits competition from the supply side* (*Stigler* [1964]).

Market domination abuse has its *prerequisites or provoking factors on the demand side as well*. A (nonexhaustive) list of these includes: *1*. lack of information and general behavioural inflexibility of consumers; *2*. rendering a supplier or a service provider switch more difficult using administrative tools and withholding information; and *3*. inelasticity of demand on the short and longer term.

Lack of information of consumers or their failure to follow information closely is not the area of classical economics, rather that of behavioural economics. Espe-

cially in the service sectors, many consumers are unaware that more favourable conditions are available and therefore fail to make use of the advantages of a switch. (An employee of a mobile operator company recently leaked verbally that several clients have not changed their packages over the past 10-14 years, paying approximately four times than a modern package's price that contains the same services.) Price inelasticity of demand (often linked with the aforementioned factor) may also cause consumer inflexibility to change. The more inelastic the demand, the less likely the consumers are to switch service providers. A good example of this is that however there is tight competition on the food market and a large number of gas or electricity utility is made available by the service providers, practically no switches took place in the residential sector in the past years (especially on the electricity market). In contrast to compulsory third party insurance, where price elasticity is far more significant and the consumers are also better informed and transaction costs are lower.

Perhaps the oldest field with the most information available is the measurement of market shares and concentration, even though it may not be the most successful method. Several textbooks, studies and case studies have been committed to this topic and it forms part of standard microeconomics and statistics curricula as well. The notions of absolute and relative concentration also belong here, for example the uneven distribution of the number of competing vendors and market shares. (See *Hunyadi–Vita* [2008] or in more detail on the specific indicators *Kotz–Johnson* [1982–1988].)

The simplest indicators of the measurement of absolute and relative concentration are the number of market actors and the concentration quotient (the per cent of revenue shared by the largest player on the market) and the Hirschmann–Herfindahl index (hereinafter referred to as HHI). These are in general use when studying competition policies too. (Another favoured tool of social sciences is the Lorenz curve, to which this article does not extend, as it is not widespread in competition analyses.)

The HHI index is the weighted average of the relative frequencies multiplied with themselves:

$$HHI = \sum z_i^2 \ ,$$

where z_i is the relative frequency (assuming individual data are available).

It is well-known that the lower limit of the HHI index is 1/n, which is the actual value of the index if and only if the population is distributed evenly and its upper limit is 1 with a single market player whose amount is not 0. In practice, when instead of the relative values being squared, they are first multiplied by 100 and then

the upper limit of the HHI will be 10 000. Many competition regulation authorities use this order of magnitude (which is naturally equivalent to the relative value). Based on the practice of the recent decades (for example in the case of the European or American authorities), some indicative values have also developed. The range of 1 500–2 500 is not dangerous, but above 2 500 the concentration is said to be significant (DoJ – FTC [2010]).

Two remarks are worth making in consideration with the usage of the HHI index for practical competition policy applications.

Markets with a different number of actors may be compared or the number of actors may change due to the integration or the separation of companies. In this case, the lower limit of the HHI index depends on the number of elements (market actors). Although in practice generally a few market actors are considered at a time, this may still cause a comparison problem. One option to resolve this is to transform the indicator to be between 0 and 1, by calculating:

$$HHI^* = \frac{HHI - 1/n}{1 - 1/n},$$

where HHI* is the transformed indicator whose value will fall between 0 and 1.

This may have practical consequences even in the case of a small number of market actors. Residential consumers in Hungary may purchase electricity from a choice of six service providers in the 2010s (the so-called universal service providers).

In 2010, the electricity provided by the universal service providers to the consumers was as shown in the table below:

Table 1

The distribution of electricity per service provider

Electricity providers	Distribution (percent)
ELMŰ	28.8
DÉMÁSZ	13.7
E.ON Dél-dunántúli Áramszolgáltató Zrt.	13.3
E.ON Észak-dunántúli Áramszolgáltató Zrt. (ÉDÁSZ)	17.7
E.ON Tiszántúli Áramszolgáltató Zrt.	14.1
ÉMÁSZ	12.4
Total	*100.0*

Source: Hungarian Energy Office.

HHI index in this case is 0.171 or 1 710. The largest provider is ELMŰ (Budapest), followed by ÉDÁSZ, but apart from ELMŰ as market leader, the rest of the distribution can be said to be fairly even. Considering, however, that there are only three proprietor actors on the market (ELMŰ and ÉMÁSZ are property of RWE), E.ON is the market leader with a share of 45.2 percent in the case of these three service provider groups. The value of the HHI index is 0.384 or 3840, which at first sight shows a considerably greater level of concentration, but it may be misleading, as the theoretical lower limits are 0.167 in the first case and 0.333 in the second. Transforming the HHI index as shown formerly, for six service providers HHI* = 0.005, and for the three owners HHI* = 0.076. While it is difficult to judge the relative concentration for the first indicator due to the uncertainty of the lower limit, the adjusted index clearly shows that the relative concentration due to the proprietorship background is considerably higher.

It is worth mentioning that the aforementioned transformation of the HHI index is often disputed in practice; in this case, the HHI only measures relative concentration, while in its original form the smaller number of actors (increasing the value of the indicator) considers the extent of absolute concentration as well. Because of this, the untransformed version of the index is in more general use. Based on the former data, the concentration is acceptable on the market of the six providers, but alarming for the three owners that may play a key role in abuse of market domination.

The values of all the parameters describing the economic phenomena for all the market actors are often not known (for example total revenues) only those of the more significant ones. This case is not covered by the textbooks, but often emerges in competition law practice.

Let us consider, for instance the market for large power stations in Hungary (of capacities greater than 50 MW). Seventeen such power stations operated in Hungary in 2011. The following table shows some of their key data.

Table 2

Data of large-scale power stations in 2011

Power station	Net capacity (MW)	Net released electrical energy (GWh)	Utilisation rate (percent)
1. Paks Nuclear Power Station	1 892	14 741.3	88.9
2. Dunamenti Power Station	1 869	1 501.7	9.2
3. Mátra Power Station	849	5 762.2	77.5
4. Tisza II Power Station	864	1 158.8	15.3
5. Gönyű Power Station	425	995.0	26.7
6. Csepel Power Station	403	1 831.4	51.9

(Contiued on the next page.)

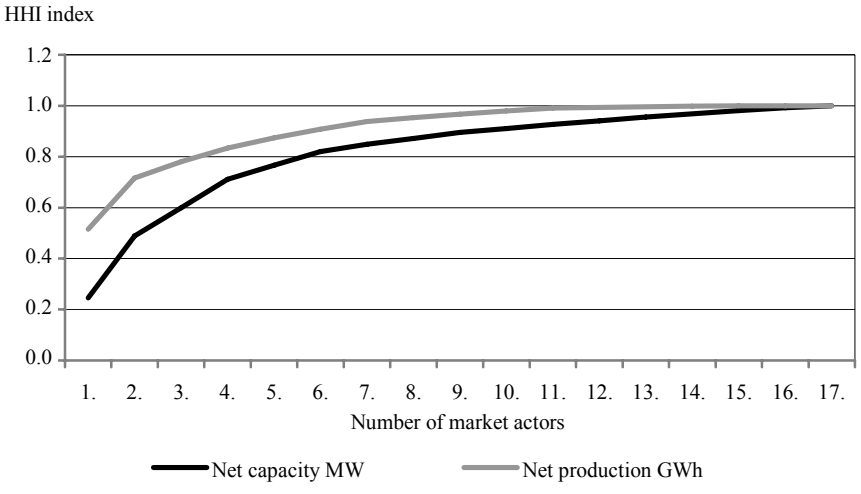
(Contunuation.)

Power station	Net capacity (MW)	Net released electrical energy (GWh)	Utilisation rate (percent)
7. Oroszlány Power Station	224	858.6	43.8
8. Tiszapalkonya Power Station	172	9.9	0.7
9. Kelenföld Power Station	190	422.0	25.4
10. Borsod Power Station	116	78.7	7.7
11. Pannon Thermal Power Station	122	47.0	4.4
12. Bakonyi GT Power Station	113	10.4	1.0
13. Kispest Power Station	108	377.0	40.0
14. Újpest Power Station	106	389.0	42.1
15. Ajka Power Station	88	35.3	4.6
16. DKCE Power Station	93	304.2	37.3
17. ISD Power Station	58	88.5	17.4
Total	*7 692*	*28 611.0*	42.5

Source: Hungarian Energy Office.

For these large power stations, the extent of concentration is shown by Figure 1 prepared based on the concentration ratios and the HHI index.

Figure 1. Concentration ratios of net capacity and production



The HHI index is for capacity 0.15 and for sales 0.32, in other words, based on both the figure and the HHI, the concentration according to sales is far more signifi-

cant. The main reason for this is that the utilisations of the respective small power stations are different from those of the two large ones standing out by far (Paks and Mátra, as these are the cheapest in Hungary at the moment).

Going back to the question of not having sufficient data, it is known that besides the seventeen large power stations, several hundreds of small ones are in operation (generally in linked generation, that is, focusing on selling thermal energy). In 2011, these accounted for 6 percent of the total electricity generated, not influencing significantly the large power stations' market, but their values could not be taken into consideration anyway as their individual generation volumes are not available. (Data provision liability to the Hungarian Energy Office is only in force for power stations with a generation capacity exceeding 50MW.) These small power stations lag behind the large ones by far in terms of capacity as well as actual generation.

If we wish to take these power stations into consideration upon estimating the HHI index, then the market shares and the HHI indices must be recalculated as a first step. In the case of the electricity released, the HHI index decreases to 0.28 but the calculated index only concerns the seventeen large power stations covering 94 percent of the market.

It is known that the market share of the largest power station that follows is below 0.033 percent. (This is also the market share of the Tiszapalkonya Power Station among the seventeen large stations.) Based on this, the number of small power stations is at least 6/0.033, that is, a minimum of 182 small power stations (there is more in reality, but the estimation can be used even if their real number is not known). The value of the HHI index is a maximum at the value of 0.28, being $182 \cdot 0.00033^2 = 0.00002$, in other words, there is only an insignificant increase.

A similarly well-known property is that the HHI index of concentration can be directly correlated to distribution as the more the individual values are distributed, the less the extent of concentration. Formally, the following relation holds:

$$HHI = \frac{\sum V^2 + 1}{n},$$

where V is relative distribution. This expression is only included here for completeness' sake; its only relevance in competition analysis being that individual data does not seem to be available at first sight, only the figures for average and distribution. In these cases, however, HHI can be calculated indirectly, further request and analysis of data therefore needs consideration.

Competition analyses widely use another indicator, measuring market power with an entirely different logic: the so-called Lerner index. (See the original definition in *Lerner* [1934] and more on its application in *Bishop–Walker* [2010].)

The Lerner index measures the market power of a particular firm by determining its relative margin, that is, the capability of the company to raise its selling price above the marginal cost. The index is defined as:

$$L = \frac{P - MC}{P},$$

where P is the price defined by the company and MC is the marginal cost. If the company is aiming at maximising its profit, then marginal cost equals marginal revenue (MR). In this case, it can be shown that the Lerner index is inversely proportional to the own price elasticity.

$$MR = \frac{d(PQ)}{dQ} = P + Q\frac{dP}{dQ} = P\left(1 + \frac{QdQ}{PdQ}\right) = P\left(1 + \frac{1}{\varepsilon}\right)$$

using the $MR = MC$ equality:

$$MC = P\left(1 + \frac{1}{\varepsilon}\right) \rightarrow \; L = \frac{P - MC}{P} = \frac{-1}{\varepsilon}$$

in other words, the Lerner index is obtained as the inverse of the own price elasticity multiplied by –1.

Theoretically, this is a well-established indicator to measure the abuse of market domination. As it was mentioned earlier, its value shows the extent to which a company is able to maintain its selling price above the marginal cost, which is increasingly a potential possibility as the price elasticity of the company's product decreases, that is, the less responsive the consumers are to changes in the selling price.

At this point, it is worth clarifying that although the phenomenon of concentration and the capability to raise prices are both sources of market domination abuse, they do not necessarily occur simultaneously. As the two measurement logics introduced formerly show, concentration may result in market advantage in a monopolistic market: a small number of actors may exploit the market due to the possible lack of competition. The price raising capability measured by the Lerner index may only occur if the price elasticity of the product is low, and if this is the case, then a number of companies may abuse the market through domination. The two phenomena often occur together, naturally, but this is not a prerequisite. To illustrate this, it is worth considering the following four cases:

a) High concentration of the market with a low price elasticity of the product: The market can be abused according to both perspectives.

This is typically the case with the naturally monopolistic public utility sector (electricity, gas, water, canalisation, district heating, etc.). It is no coincidence that there is local government intervention, or that the price is limited by the authorities, or that a price control is in force in this sector.

b) Low concentration of the market with a high price elasticity of the product: This is the case when no market domination abuse is expected. Examples are foods sold in the retail sector and general household items.

c) High concentration of the market with a high price elasticity of the product: Due to the small number of service providers, market domination abuse is theoretically possible but the high price elasticity of the product causes tough competition still. Examples generally quoted are mobile operators or compulsory third party car insurance. Naturally, the condition of the formation of market competition is the possibility of simple and easy service provider switch. (The case of banking services would be similar, but this latter condition is not available as changing a bank is difficult and costly).

d) Low concentration of the market with a low price elasticity of the product: In this case, the consumers have a proper range of suppliers to choose from, but market abuse is still possible due to the low price elasticity of the product. This is the least probable case as the ease of switching a supplier or vendor limits raising the selling price. Although the price of bread is inelastic, it is still difficult to keep prices high due to intense competition. Examples also exist in the energy sector; the non-residential consumers in the electricity and gas sectors are characterised by this. As of 2008, they are only able to purchase electricity from the free market. The price elasticity of their demand is low (we have conducted several surveys over the past years, based on which the demand price elasticity of electricity from small businesses on the short run is about –0.3 – –0.4). Because of this, although several suppliers offer electricity to them, supplier switches are marginal in volume. Market domination abuse has probably evolved as a result. After liberating the market, price control by the authorities remained in control in the residential sector, while large companies were in a better negotiating position than large consumers. The suppliers attempted to compensate the profit thus foregone from the small businesses sector, resulting even in price hikes of 30-40 percent in 2008.

Returning to the basic topic of this article, to the indicators, the Lerner index raises problems in practice as well as theory. The following theoretical problems emerge upon its application:

a) The marginal cost is usually an ineffective method of pricing. In a number of industrial fields, the price contains the return of the investment cost (fixed cost) and the capital cost.

b) The Lerner index can only be used effectively for single product companies, where the particular costs cannot be cross-loaded to other products. (See the distorting effect of multiple product cases in Appendix 3.1 of *Bishop–Walker* [2010]).

Computing the Lerner index

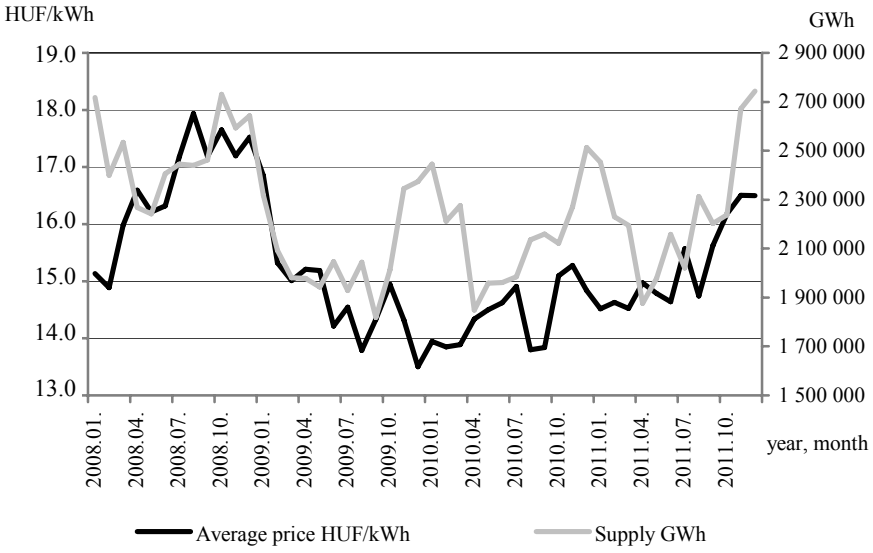
The theoretical considerations are not particularly relevant concerning the electricity generation described in the article. The power stations are indeed homogeneous producers of a single product (electricity) and theoretically the selling price may be close to the marginal cost as many of them are obsolete, write-off power stations not capable of generating capital cost significantly in the tough competitive (primarily import influenced) market situation. (As a result of this, a number of power stations have ceased operations, such as Tisza II and several blocks of Dunamenti Power Station.)

If we wish to quantify the Lerner index, it can be done from the cost-price side. The marginal cost per power station and the actual prices need to be known, but presently this information is unavailable. In practice, this latter (cost-price side) principle is generally followed, resulting in the price inelasticity of demand. As in this case it is not possible to take this option, the index is approached from the other side.

To estimate the price elasticity, time series methods were used at first. The non-residential electricity sector has been fully liberalised since 2008. Figure 2 shows the evolution of the average price at the power stations and their supply (assuming that this is the sector that purchases the most on free market basis) between 2008 and 2011, based on monthly data.

The data have been seasonally adjusted applying the TRAMO_SEATS method. Using both the original and the seasonally adjusted data, calculating in the case of time series with the Cochrane-Orcutt algorithm, the price elasticity of demand is positive, which contradicts economic rationale.

Figure 2. Average price and supply of electricity between 2008 and 2011

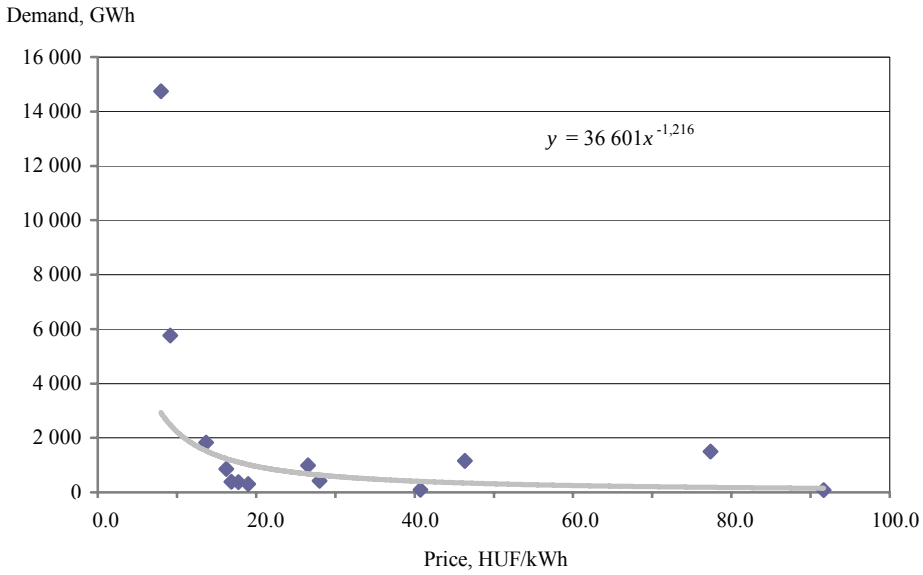


Source: Hungarian Energy Office.

The fundamental reason for this is that there is a reversed correlation between price and demand, it is not the demand that determines the selling price, but it is the price that depends on the size of demand (and in the case of a special product such as electricity, it depends on the size of supply). This has also been examined with the Granger causality test, which proved our hypothesis. The smaller the demand to satisfy, the lower the average price, higher prices may be realised in the peak consumption periods. (This is the main reason why the price is often around the marginal cost as the occasionally occurring high demands at peak consumption periods allow for high profits. This correlation demonstrates itself most clearly in stock market prices).

We also attempted to estimate the price elasticity of demand using power station data from 2011. Although the individual power station prices were not known, we knew that they sold electricity at an average price of 16.9 HUF/kWh in 2011. Assuming that price is proportional to the utilisation rate of the power station, and in this case Paks sells electricity at an average price of 8 HUF/kWh, with Mátra Power Station at 9 HUF/kWh which more or less correspond to the actual prices. The prices thus obtained and the corresponding quantities were used to establish a simple exponential regression function. The XY figure (scatterplot) and the estimated function are shown in Figure 3.

Figure 3. Scatterplot of prices and demand and the regression curve



It can be seen that the price elasticity of demand is –1.2; in other words, a price change of 1 percent will generate an average change of 1.2 percent of demand in the opposite direction. The modulus of price elasticity is above 1, that is, the demand can generally be said to be price elastic albeit not significantly.

The Lerner index based on the above is 0.8: the market price may be as high as fivefold the marginal cost.

Alternative calculations of HHI and Lerner index

Both the HHI index and the Lerner index demonstrate a possibility of considerable market domination abuse in the electricity generation sector in Hungary.

This result is in coherence with the economic situation of large power stations having a strong absolute concentration in the area of generation capacities and the relative concentration is also significant, while in the area of generation, the degree of concentration is even higher due to the greatly varying capacity utilisations. (This very difference of capacity utilisation that shows that power stations sell energy at a wide range of prices, being forced to match it to their costs. It provides room for deducing the prices from the capacity utilisation, which match other information as

well.) The demand price elasticity of power station generation was estimated and this gave a clue to estimate the magnitude of the Lerner index, which again showed that power stations are indeed able to abuse their market domination, establishing their prices well above the marginal cost.

Still, the scenario described is somewhat contradictory to the general pricing practice followed by power stations in Hungary. The prices – according to this scenario – follow the marginal costs; there are explicitly "cheap" power stations where the marginal cost is lower: these are typically the large power stations with the capacity to abuse their market positions and there are smaller and technologically less advanced large power stations who sell more expensively. The generally tough market competition is often quoted as an explanation to this, principally fuelled by energy imports. This indicates that the role of imports cannot be neglected when evaluating the possibilities of market domination abuse.

The interpretation and the incorporation of imports into the measurement as a factor strengthening or diluting concentration raise a number of methodological and economic questions.

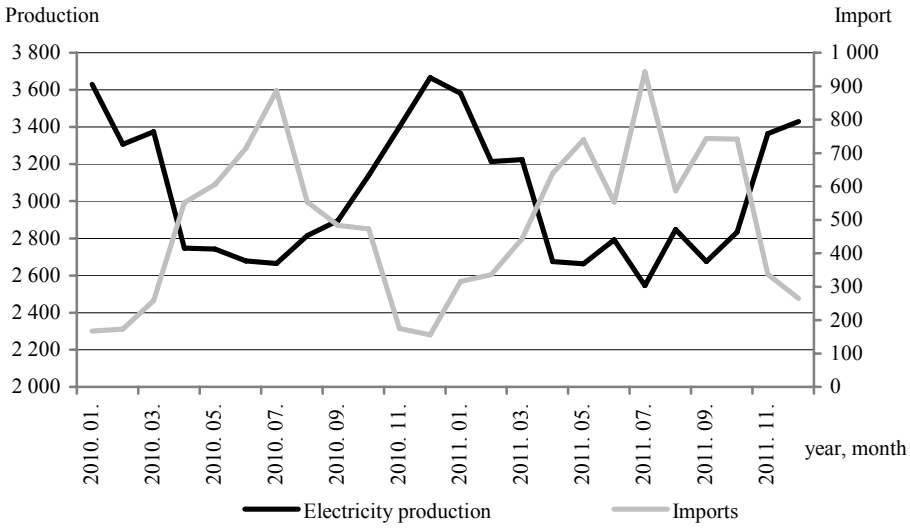
a) There are branches where imports emerge as a true monopoly. Such is the natural gas sector, where 86 percent of consumption is import based with the vast majority entering the country within the framework of E.ON Trade's long-term contract established in 1996 (back then between MOL and Gazprom). Although other sources of procurement are available (MOL, Tigáz, and previously Emfesz), because of this, the Hungarian Energy Office rated E.ON Trade on the wholesale market as a supplier with great market power, due to its definitive 55 percent market share (as a thumb rule, this rating is automatically achieved at 40 percent of the market share). As the importing company dictates the prices (greatly influenced by the long-term contract's prices, naturally), import here is typically a factor increasing market domination. This is probably the cause of Hungarian gas prices being higher than the European average and that E.ON maintains a particularly strong negotiation position against the state of Hungary as a regulatory authority.

b) In certain branches there is no import competition at all, mostly where transportation costs render this impossible. Such are basic building materials, diesel, or petrol. In the case of the latter, some petroleum derivatives from abroad do enter certain regions of the country, but their share does not exceed 20 percent. Therefore, on the wholesale market of fuels, MOL could have absolute market domination with its 80 percent share and its capacity to raise prices could also be signifi-

cant. The increasing price elasticity of the demand for fuels supports this scenario. However, none of the competition authority checks managed to put MOL in the wrong (to date) for its pricing practices, the main reason being that MOL's pricing is not cost based – meaning that Lerner index supported examinations are meaningless in this case – is linked instead to Mediterranean and Rotterdam index prices, that is, to global trends and changing exchange rates, whose correct nature has not been called into question with success so far. (See more detail in *Sugár* [2011a].)

c) There are examples placed between the two extremes mentioned before, such is the market for electricity. The share of imports fluctuates between 15 and 20 percent, neither definitive nor significant.

Figure 4. Production and import of electricity between 2010 and 2011
(GWh)



Source: Hungarian Energy Office.

Imports may be managed upon modelling and measurements by (and this often is the practice) that import volume is considered a separate power station. Imported electricity comes to Hungary mainly from Slovakia and Romania and in smaller quantities from Austria; the main reason being its low price. (Besides Slovakian and Ukrainian electricity, the price of Polish and Czech electricity is also lower.) The import quantity is limited by the boundary capacity (which has to be bidden for sepa-

rately). Another option would be the consideration of the imports from each neighbouring country as a power station separately, but data are not available in such breakdown. One solution (which we have opted for too) is to separate the import volume into two parts, the cheaper but greater volume season (summer) and the more expensive but lower volume (winter) season: the periods from April to September and from October to March. (Examples of regional modelling are also available (*Kiss–Barquín–Vázquez* [2006].)

The other factor significantly influencing monopolistic situations is that Hungarian regulations (in conformity with the European ones) consider companies in monopolistic positions with those in significant market power (as mentioned already in the case of natural gas) and brings about monopoly resolving decisions in consequence.

The Hungarian Energy Office has brought about four resolutions establishing significant market power (SMP), of which two – concerning the wholesale and the system level service markets – were issued in the summer of 2008 and two further ones concerning the retail market in the spring of 2009. The last SMP resolution was issued in November 2011. The wholesale SMP resolution has obliged Hungarian Electricity Company (MVM) as a company of significant market power to carry out the following:

> *a*) Based on the corresponding legal regulations, the company has to auction a quantity of electricity from the quantity available to reduce its market share calculated without electricity below 40 percent.
>
> *b*) Besides the establishment of the auctioning obligation, the SMP resolution also established maximum prices concerning electricity sales by MVM. The Office established the company's maximum selling price in 2009 as 19.05 HUF/kWh. The resolution of 2011 deleted this point considering MVM's decreasing market power and the level of prices that formed.

In order to interpret the resolution properly, it must be understood that MVM fixed a significant part of the capacity of the Hungarian power stations earlier with long-term contracts. Although these contracts had to be broken as a result of EU examinations, the subsequent replacement contracts were also made between MVM and the large powers stations mostly. MVM definitively ties down the production capacity of Paks Nuclear Power Station and Mátra Power Station. As the resolution of the Hungarian Energy Office shows, MVM is obliged to sell part of this capacity in auctions, this quantity also having an effect on the extent of market concentration. The consideration of this is ambiguous. In our case, we chose the solution whereby

the larger quantities auctioned off in 2011 and their prices were considered a competitive power plant package. The capacities of Paks and Mátra power stations were decreased proportionally by these quantities.

By considering all these, the new (fictitiously prepared) power station portfolio is as shown by Table 3.

Table 3

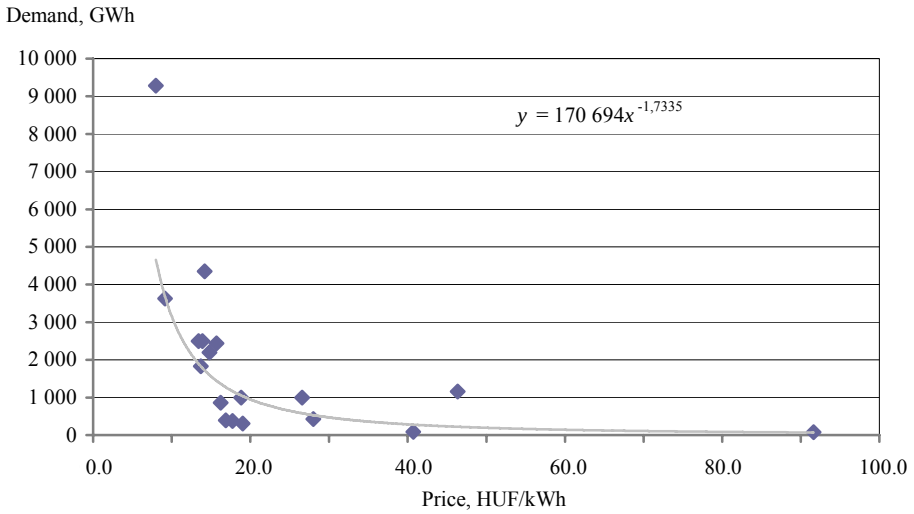
Corrected data of large power stations, auctions and imported volumes in 2011

Power station	Electricity sold (GWh)	Electricity sold (percent)	Estimated average price (HUF/kWh)
Paks Nuclear Power Station	9 287	25.86	8.0
Summer imports	4 356	12.13	14.2
Mátra Power Station	3 630	10.11	9.2
1st auction	2 500	6.96	13.9
2nd auction	2 500	6.96	13.4
Winter imports	2 439	6.79	15.7
3rd auction	2 200	6.13	14.8
Csepel Power Station	1 831	5.10	13.7
Dunamenti Power Station	1 502	4.18	77.4
Tisza II. Power Station	1 159	3.23	46.4
4th auction	1 000	2.78	18.8
Gönyű Power Station	995	2.77	26.6
Oroszlány Power Station	857	2.39	16.2
Kelenföld Power Station	422	1.17	28.0
Újpest Power Station	389	1.08	16.9
Kispest Power Station	377	1.05	17.7
DKCE Power Station	304	0.84	19.0
ISD Power Station	89	0.25	40.7
Borsod Power Station	79	0.22	91.6
Total of large power stations	*35 916*	*100.00*	16.8

The data thus modified alter the values of both indicators significantly. The HHI index decreases greatly, from the initially calculated 0.32 down to 0.07, indicating that the extent of concentration is not significant if import volumes and auctions are taken into consideration.

Estimating the value of the elasticity coefficient we obtain a value of –1.7, that is, the price elasticity is greater, a 1 percent change in price will result in a change of 1.7 percent of demand in the opposite direction as indicated in Figure 5.

Figure 5. Scatterplot of prices and demand and the regression curve



The –1.7 percent price elasticity corresponds to a Lerner index of 0.6, showing a far smaller potential to abuse the market; the price may exceed the marginal cost by 67 percent. Naturally this is not an insignificant value, but far smaller than the potential to raise the prices fivefold to that of the marginal cost before the introduction of the corrective measures.

The correction of the import volumes and the auctions therefore yielded the results whereby market concentration was not significant on the source side of the electricity generated (imported or auctioned) and the Lerner index indicated a far weaker power to raise prices. This means that the obstruction of the development of abusive market power can be efficiently achieved by both the intensification of competition (imports) and administrative measures (auctions) at least based on the statistical figures available.

Finally, it must be remarked that the uncorrected data that indicated the possibility of market domination abuse may be misleading, as they do not present the supply of power stations in pure market conditions, but beside the presence of a company, MVM that has dominated the market for decades. MVM determined fully the quantity and the price of electricity taken over in the past and still influences it today. A perfectly clear picture would be available if the power stations were selling closer to pure market conditions without the presence of a dominating wholesaler. There are historic examples available for this: in the beginning of the 2000s, all long-term electricity purchase contracts in California were terminated and power stations were forced to take their capacities to the stock market; this drastic intervention step did result in prices close to the marginal cost. A similar construction cannot be envi-

sioned given the Hungarian conditions, although many researchers of economic statistics dream about such a scenario, which would be a close approach of the very rare case of controlled experiments in economics.

Our article primarily focused on the practical calculation of the indicators used to measure market power and as intended, it linked mechanical measurement to economic analysis to obtain as clear information as possible on the market situation and the intensity of market power.

References

BISHOP, S. – WALKER, M. [2010]: *The Economics of EC Competition Law: Concepts, Application and Measurement*. Sweet and Maxwell Ltd. London.

DOJ – FTC (U.S. DEPARTMENT OF JUSTICE AND THE FEDERAL TRADE COMMISSION) [2010]: *Horizontal Merger Guidelines*. Washington, D.C.

HUNYADI, L. – VITA, L. [2008]: *Statisztika I–II*. Aula. Budapest.

KISS, A. – BARQUÍN, J. – VÁZQUEZ, M. [2006]: Can Closer Integration Mitigate Market Power? – A Numerical Modeling Exercise. In: *LaBelle*, *M*. – *Bán*, *T*. – *Kaderják*, *P*. (eds.): *Towards More Integration of Central and Eastern European Energy Markets.* Regional Centre for Energy Policy Research. Budapest. http://rekk.uni-corvinus.hu/c3em/pdf/modeling_study.pdf

KOTZ, S. – JOHNSON, N. (eds-in-chief) [1982–1988]: *Encyclopedia of Statistical Scienses*. Vol. 1–9. John Wiley and Sons Inc. New York.

LERNER, A. [1934]: The Concept of Monopoly and the Measurement of Monopoly Power. *Review of Economic Studies*. Vol. 1. No. 3. pp. 157–175.

STIGLER, G. J. [1964]: *Competition and Concentration*. Challenge. January.

SUGÁR, A. [2011a]: A hazai benzin és gázolaj árszintjének és árazásának empirikus elemzése. *Statisztikai Szemle*. Vol. 89. No. 6. pp. 624–643.

SUGÁR, A. [2011b]: The Political Economy of Price Control. *Society and Economy*. Vol. 33. No. 2. pp. 321–345.

Was the Financial Crisis of 2008 Forecastable?

Barnabás Ács
Senior Account Manager
Thomson Reuters, London
E-mail: acsbarnabas@gmail.com

The fall of Lehman Brothers in September 2008, surprised the financial markets around the globe. The financial collapse immediately afterwards hit everybody unexpectedly. This article discusses whether deeper analysis of the US macroeconomic data would have been able to give hints to traders about the approach of the crisis of 2008.

KEYWORDS:
Financial crisis.
Forecast.

The National Bureau of Economic Research (NBER) published on 1 December 2008 that the United States had been in the state of recession since December 2007. Hereby recession became official, 3 months after the collapse of Lehman Brothers and 5 months after the historic high of the West Texas Intermediate (WTI) benchmark. Then almost 8 months had elapsed since Bear Sterns was saved by JP Morgan. 2008 passed by with the S&P 500 Index falling almost by 41 percent.

Several authors in literature (*Faber* [2009], *Ritzholtz* [2009], *Morris* [2009]) state that the current financial crisis was built in the system, and several renowned economists have blamed the regulatory bodies, mainly the Federal Reserve System (FED), being far too inactive preventing it (*Fleckenstein–Sheenan* [2008]). Based on their opinion, the crisis was encrypted in the system not only since 2001, but since the 1970s, the start of the real estate market boom. However, it seems that its amplitude, depth and arrival have surprised the general public. The aim of this study is to examine whether the 2008 economic crisis was forecastable, by means of statistical tools and time series models, and if there were any visible signs in the economic databases noticeable to everyone, to clarify whether its suddenness was the result of global blindness or the "storm clouds" were only visible to a few "insiders" only.

The macroeconomic indicators published by the national statistical agencies and supranational organisations (IMF, OECD, World Bank) are available basically for free to all investors and decision makers. This is especially true for the macroeconomic indicators of the United States.

The analysis spectrum of this paper is between the first quarters of 1985 and 2010. In this given 25 years, the NBER, the ultimate reference for economic cycles, identified three macroeconomic downturns.

The Japanese Banking Crisis of 1990, followed by the Dotcom Crisis of 2000 that had started with the burst of the Internet equities bubble and bottomed after the terror attacks of September 2001, and finally the financial meltdown starting at the end of 2007, referred to as the Crisis of 2008.

The length of the data series and the number of historical events make the analysis of the typical economic characteristics possible. The question of this article is: Would analysing the time series and their interactions with naive data mining tools have made possible to sense the approach of the crisis, or do even the most accurate statistical tools fail to unveil the signs?

However, the non-profit character of the study must be emphasized. It does not aim to find the most secret, always profitable investment strategy and indicator constellation. Neither does it try to criticise the main theses of various macro economic theories.

The article, however, sets the objective of clarifying the role of the different macro indicators eagerly followed by investors that influence their trading decisions, the real interactions thereof, as well as their short and long term effects on one another. It aims to uncover non-realised anomalies. In a holistic view, the goal is to check whether the general attitude towards statistics and econometrics is true: only afterwards they are clever. Hence, I am going to test numerous hypotheses.

It is assumed that the output indicators, in consequence of self-fulfilling prophecies, are not behaving as results in statistical models but as root causes. Moreover, it is presumed that not the indicators followed most intensively by investors carry the maximum amount of information about the state of the economy. Nota bene, the role of these indicators can change during time.

It can be hypothesized that if not their own trend-change but the change of interactions between indicators, can indicate the start of a recession. It can be experienced in several cases that most of the regressing equations lose their stability in the times of crisis. Hence the paper strives to find indicator pairs that are either stable in crucial situations or behave the same way before each crisis.

It is probable that a recession is indicated later due to error correction mechanism among variables in dynamic relationships. Moreover, it is also supposable that the explosion or birth of the dynamic relationships among variables carries useful information for the trend changes of the variables indicating the recession.

The study is based on the discussion of the definitions, nature and characteristics of the crises (for example *Kindleberger* [1989], *Fisher* [1933], *Minsky* [1977], *Dewald* [1972]. The main reasons of the current crisis (for example *Faber* [2009], *Fleckenstein–Sheenan* [2008], *Harris* [2008], Morris [2008], *Ritzholtz* [2009], *Király–Nagy–Szabó* [2008]) are also incorporated in the hypotheses and the findings. It demonstrates the purposes of the applied econometric methodology of analysis, goes through the results (model estimations, hypothesis tests) and points to further examination and research possibilities. Based on my goals, the following hypotheses were verified.

– *First hypothesis*: The output indicators, in contrast with prior expectations, and probably due to the self-fulfilling prophecies, are not endogenous but exogenous variables.

– *Second hypothesis*: When examining particular variables in different variable categories, it is important to analyze others than the ones followed by market participants as it is not sure that the emphasized variable carries the most information. Besides, the highlighted role can change from one period to the other.

– *Third hypothesis*: It is not the trend of the macroeconomic indicators that might be able to indicate the start of a recession but the change in the interaction between them.

– *Fourth hypothesis*: It is presumable that due to the necessary error correction mechanism among variables in dynamic relationships, the start of a recession can be hidden or indicated later, as similar events tend to characterize upswing times.

– *Fifth hypothesis*: The explosion or birth of the dynamic relationships among variables carries useful information for the trend changes of the variables indicating the recession.

In order to prove my hypotheses, data mining and knowledge discovery were applied.

1. Variables included

The article is analysing 140 macroeconomic indicators of the US, issued on a quarterly basis covering a time span between 1985 and 2010. As the recession of 2008, in contrast to that of 1990, started undoubtedly in the USA, it wasn't necessary to consider incorporating the macroeconomic indicators of other countries.

These 140 pieces of indicators – 99 chosen and 41 derived – enable to cover the whole economic sphere of the US, starting from GDP through price and interest rate levels and production data to the number of filings for bankruptcy protection. Those variables were chosen that fulfilled at least one of the two following criteria:

– They are in spotlight, i.e. the financial markets pay attention to them. This principle is fulfilled if they are highlighted in the Thomson Reuters data bases.

– They help to cover all aspects of the US economy, in order to enable the identification of the possible latent indicators. This goal is based on the "Guide to Economic Indicators" edited by *The Economist* [2006].

The chosen indicators fall into nine categories (*The Economist* [2006]). These are: *1*. indicators of value added, *2*. employment indicators, *3*. fiscal indicators, *4*. consumption indicators, *5*. investment and savings indicators, *6*. indicators of industry and commerce, *7*. indicators of balance of payments, *8*. money and financial markets indicators, and *9*. indicators of prices and wages.

It was not intended to have the same number of elements in each category. To eliminate absolute values, units and the effect of inflation, 41 indicators are expressed as a ratio of either current or constant priced GDP (*Hajdu–Virág* [1993]).

2. The structure of the analysis

For the purpose of meeting my research goals, I analysed the formerly mentioned data series. Firstly, I examined whether the chosen variables are stationary, so I applied the Dickey-Fuller regression for each y_t variable:

$$\Delta y_t = \mu + \beta y_{t-1} + \alpha_1 \Delta y_{t-1} + \ldots + \alpha_r \Delta y_{t-r} + \varepsilon_t \ ,$$

where $\beta = \delta - 1$. The existence of unit root in the augmented Dickey-Fuller test is proven by accepting the null hypothesis of:

$$H_0 : \delta = 1 \qquad H_1 : \delta < 1 \ ,$$

that is

$$H_0 : \beta = 0 \qquad H_1 : \beta < 0 \, .$$

Hence, we are considering a time series to be stationary, if the null hypothesis is rejected. The τ-test was used as test statistic:

$$\tau_\beta = \hat{\beta} \Big/ \left(se\left(\hat{\beta}\right) \right) \ ,$$

where $\hat{\beta}$ is the estimation of β and $se\left(\hat{\beta}\right)$ is the standard error of the estimated coefficient. We are considering a time series being stationary if $\tau_\beta > \tau_{critical}$.

Differencing was applied for all the nonstationary variables until the given variable was deemed to be stationary. Hence the order of integration of each variable could be specified.

In the second step, Granger-causality tests were used to identify the relationships between each y_t time series.

Applied to stationary time series – $x_t \sim I(0)$ and $y_t \sim I(0)$ – according to the null hypothesis x does not Granger-cause y, if no better estimate can be given to y compared to the case when only the past values of y are analysed. Id est:

$$H_0 : MSE\left(\hat{y}_t \,|\, y_{t-1}, y_{t-2}, \ldots\right) = MSE\left(\hat{y}_t \,|\, y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots\right)$$
$$H_1 : MSE\left(\hat{y}_t \,|\, y_{t-1}, y_{t-2}, \ldots\right) > MSE\left(\hat{y}_t \,|\, y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots\right)$$

where MSE stands for the mean squared error and $\hat{y}_t$ marks the estimated values of y. Based on this assumption, the following regression equation was applied on all the possible combination of the same order integrated variable pairs (x_t, y_t):

$$\hat{y}_t = \alpha_0 + \alpha_1 y_{t-1} + \ldots + \alpha_l y_{t-l} + \beta_1 x_{t-1} + \ldots + \beta_l x_{-l} + \varepsilon_t .$$

The hypothesis system can be altered in the following:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$
$$H_1 : \exists j, \beta_j \neq 0 .$$

The test of this hypothesis system is quite straightforward with the Wald-test described also by *Jones* [1986]. So:

$$F_{emp} = \frac{MSE\left(\hat{y}_t \left| y_{t-1}, y_{t-2}, \ldots\right.\right)}{MSE\left(\hat{y}_t \left| y_{t-1}, y_{t-2}, \ldots, x_{t-1}, x_{t-2}, \ldots\right.\right)}$$

with $(2l, T-2l-1)$ degree of freedom.

The null hypothesis is rejected, that is, it is presumed that x does Granger-cause y, if $F_{emp} > F_{(2l,T-2l-1)}$.

In the case of first and second order integrated time series, the analyses follow the same idea. Only the hypothesis system, the regression equations applied for the variable pairs and the Wald-test are modified with the first or second order differentiation.

To be able to measure the interdependency of the chosen variables, a so-called *cause-reason matrix* is compiled. This is a table with rows containing exogenous variables (being 'cause' in the Granger-causality tests) and columns including endogenous variables (reasons).

In the *cause-reason matrix* the exogenous and endogenous codes refer to the identifier of the given variable in the database (1 to 140). The cells of the table marked with "X" indicate the relationships where the *F*-values of the Granger-causality Wald-tests were significant. The matrix reveals that personal consumption (variable 1, referred to as: 1) does Granger-cause private investment (2), however private investment (2) does not Granger-cause personal consumption (1).

Using this matrix, a so-called "causality carpet" can be made, hence the relationships between variables can be easily qualified and quantified. It reveals, for example, that the free cash flow of institutions (7) – in column – is Granger-caused by the added value of households and institutions (5) and new orders of industries (8). At

the same time, however, the free cash flow of institutions (7) – in row – does Granger-cause amongst both private investments (2) and new industrial orders (8). Hence, it can be stated that a feedback relationship exists between the free cash flow of institutions (7) and new industrial orders (8).

Cause-reason matrix of the variables

Cause of #?	Exogenous Code	Reason of #?										
		10	68	16	43	32	0	32	40	33	4	...
		Endogenous Code										
		1	2	3	4	5	6	7	8	9	10	...
64	*1*	–	X		X	X			X	X		
29	*2*		–			X						
7	*3*			–								
48	*4*		X		–	X						
15	*5*					–		X				
0	*6*						–					
38	*7*		X					–	X			
58	*8*		X		X	X		X	–	X		
41	*9*		X			X			X	–		
5	*10*										–	
...	...											–

Note. The "Reason of #?" row states how many variables do Granger-cause the given variable. The "Cause of #?" column identifies the number of variables the given variable is a Granger-cause of.

The cause-reason matrix shows that personal consumption (1) is the Granger-cause of 64 variables, and there are only 10 variables that do Granger-cause personal consumption (1).

The main goal is to be able to classify whether a particular variable behaves as an exogenous or endogenous variable. To this end, normally the Hausman-test is used (*Hausman* [1978]); however, its programming in EViews is not possible. That is why a simple rule of thumb is applied. Based on the numerosity of the Granger-causalities, the variables are classified into four categories.

Classifying variables based on Granger-causalities:[1]

– Exogenous (ex): The given variable does Granger-cause twice as many variables than it is the cause of, i.e. *"Cause of #"* $\geq 2^*$ *"Reason of #?"*

[1] In the list, # stands for "how many".

– Rather exogenous (r-ex): The former criterion is not met, however, the given variable does Granger-cause more variables than it is the cause of, for example: *"Reason of #?" < "Cause of #?"< 2* "Reason of #?"*

– Not applicable (n.a.) "Reason of #?" = "Cause of #?"

– Rather endogenous (*r*-en): The given variable is Granger-caused by more variables than it does Granger-cause, for example: *"Cause of #?" < "Reason of #?"< 2* "Cause of #?"*

– Endogenous (en): The given variable is Granger-caused by twice as many variables than it does Granger-cause, for example: *"Reason of #?" ≥ 2* "Cause of #?"*

Regarding the former classification, each variable can be identified as being exogenous or endogenous.

In the third step, the variable pairs showing significant Granger-causality were described by the equation:

$$y_t = \alpha + \beta x_t + u \ .$$

It is examined whether the different recession times cause structural breaks in the particular regression relationships. That is why each regression equation was tested through the Chow breakpoint test (*Chow* [1960]). Its main point is to examine whether the parameters of the regression equations of the sub-periods of a given period, identified by one or more break-points, differ from each other, that is:

$$\begin{aligned} y_t &= \alpha_1 + \beta_1 x_t + \varepsilon_1 \\ y_t &= \alpha_2 + \beta_2 x_t + \varepsilon_2 \end{aligned} .$$

The hypothesis system according to this:

$$\begin{aligned} H_0 &: \alpha_1 = \alpha_2, \beta_1 = \beta_2 \\ H_1 &: \exists\left(\alpha_1 \neq \alpha_2, \beta_1 \neq \beta_2\right) \end{aligned} .$$

A structural break is identified, if equality of either regression parameters is rejected. The following Wald-test is applied for the test. The *F*-statistic is based on the comparison of the restricted and unrestricted sums of squared residuals and in the simplest case involving a single breakpoint, is computed as:

$$F_{emp} = \frac{\left(\tilde{e}'\tilde{e} - \left(e_1'e_1 + e_2'e_2\right)\right)/k}{\left(e_1'e_1 + e_2'e_2\right)/\left(T - 2k\right)} \ ,$$

where $\tilde{\varepsilon}'\tilde{\varepsilon}$ is the restricted sum of squared residuals, $e_2'e_1$ is the sum of squared residuals from subsample before the break, $e_2'e_2$ is the sum of squared residuals from subsample after the break. T is the total number of observations and k is the number of parameters in the equation, in this case $k=2$. This formula can be generalized naturally to more than one breakpoint.

We consider the breakpoint to be significant, if $F_{emp} > F_{(k,T-2k)}$. There is a constraint regarding placing the breakpoint, the Chow-test can only be carried out, if the size of the subsamples generated by the breakpoint is greater than the number of the estimated parameters, that is $T_i > k$.

In order to be able to examine the effect of recessions on a continuous basis, by indexing the breakpoints, the break-point test is conducted for each of the quarters between Q1 1988 and Q4 2008 (conforming the $T_i > k$ constraint). This way 84 pieces of Chow-tests are conducted and 84 pieces of *F*-values are calculated, that are used to build a time series of *F*-s for all the regressions:

$$z_t = \left(F_{1988q1}, F_{1988q2}, \ldots, F_{2008q3}, F_{2008q4}\right).$$

The creation of *F* time series enables the followings:

– As the critical F-value is the same in all of the tests (in this case $F_{(k,T-2k)} = 3.09$), if all the breakpoints F-values stay under this critical level, all the breakpoints should be considered not significant, which means that the regression is considered to be stable for the whole period, free from any structural breaks.

– By charting the empirical *F*-values, the quarters, where structural breaks occur, can be identified easily. This happens when either of the *F*-values exceeds the critical *F*-value. This way the breaks in the regression relationships preceding a recession can be identified. As previously mentioned, the time series analysed contain three recession periods identified by NBER: Q3 1990 to Q1 1991, Q1 2001 to Q4 2001, and Q4 2007 onwards.

Variable pairs with a structural break before the recessions at least two times out of the three can be identified easily.

In the fourth step the former regression pairs were put under scrutiny with the help of the Johansen-test checking for cointegration. Its purpose is to determine whether a group of non-stationary series is cointegrated or not.

EViews uses an identification method so that the error correction term has a sample mean of zero. We identify the part inside the error correction term by regressing the cointegrating relations $\beta' y_t$ on a constant (and linear trend).

When testing cointegration, the analyzed periods were:

– Q1 1986 to Q1 2010 to see which variable pairs were cointegrated in the whole period.

– Q1 1986 to Q4 2007 to make the comparison with Q1 1986 to Q1 2010, and identify those variable pairs that are not affected by the recession at all.

– The so-called inter-recession bands to identify the variable pairs with similar cointegration parameters in all three periods (Q1 1986 to Q3 1990, Q2 1991 to Q1 2001, Q1 2002 to Q4 2007).

– Q1 2002 to Q1 2010 to identify the variable pairs with changed parameters after the burst of the dotcom bubble.

The aim is to identify those variable pairs, that were cointegrated with the same β' parameters for more periods.

3. Results

The analysis works with a data set of 140 macroeconomic indicators looking back 25 years, published quarterly. When analysing Granger-causality, the significance of 11 084 variable pairs was checked and 843 Granger pairs were categorized. When running the Chow-test, 303 576 *F*-values were calculated and charted in 3 614 graphs. Doing the individual visual analysis of these, 828 characteristic regressions were chosen.

When testing cointegration, 4 968 Johansen-tests were run and analysed in six different time periods.

As the analysis was based on a wide literature and conducted with the utmost care, it can be stated that the main tendencies could be identified, and all material factors were discovered.

In congruence with the presumptions, it was proved that the majority of the economic time series are non-stationary. Out of 140 variables, 102 turned out to be first order integrated, meaning that the result of the regressions run on the non-transformed form of these variables will be biased. That is, the traders running the classic regression and correlation tests on these variables without differencing them, will base their decisions on wrong results.

Among others, the Granger-causality analysis revealed that in contrast with expectations, economic indicators measuring added value are primarily exogenous variables. The GDP is in feedback relationship with both the financial sector and the monetary policy.

The employment indicators can be solidly considered endogenous variables, however, no common factor can be identified that does Granger-cause the employment variables.

In contrast to expectations, the fiscal indicators are strongly endogenous and their primary Granger-causes are the profit and output indicators of the non-financial sectors.

In the case of consumption variables, it turned out that energy consumption and crude oil consumption handled similarly previously have completely different characters, the former shows endogenous, while the latter exogenous characteristics.

As for the investment indicators, it was proven that the expectations are manifested in profit indicators since the profit indicators are exogenous and in contrast, the stock indicators are endogenous.

Among indicators of commerce, the housing ones are rather endogenous, whereas those of car sales have no clear characteristics. Common factors can be identified, as short term consumption variables Granger-cause housing indicators, while it is the long term consumption indicators that determine the car market.

Neither do the indicators of the current accounts carry a common character, nor do common factors play a part in determining the budget deficit and current account balance.

A surprising finding is that the FED fund rate and the monetary base are exogenous variables.

Price indicators behave, as expected, clearly as exogenous variables.

The Chow-tests discovered inter alia that the added value of households and institutions have the most stable regression relationships among the value added indicators. It was shown that only the aggregate GDP and the GDP created by the government are in stable relationship with inflation.

The regressions including the employment indicators are not stable, this is one of the most important findings of the Chow-tests. Out of the five employment indicators only the participation rate has classifiable regressional relationship with other variables. Another interesting fact to note is: the Granger-causality tests showed that employment indicators are clearly endogenous variables. However, solid relationships are only built when they bear exogenous roles.

Out of ten fiscal indicators, merely four show solid endogenous relationships, hence it is also proven that this indicator group is also lacking steady relationships. It has to be emphasized, however, that government consumption and investment form stable relationships with all of its exogenous variables. The solid relationship between the fiscal and monetary indicators stresses the presumption that economic policymakers of the US are relying on both tools.

It is shown that the disposable income to GDP is in steady relationships with all the indicator categories. The crude oil consumption, in contrast to the energy consumption, forms solid relationship only with commodities price index, a sign that oil consumption is price-flexible.

Based on the Chow-tests, it can be stated that the profit of the financial institutions builds up solid relationships with the participation rate and the factors affecting their economic environment. However the profit of non-financial institutions is influenced by the indicators of the broader economy, such as consumption expenditures, consumer price index, and industrial production.

In case of the housing indicators, the focus has to be put on the housing inventory. This variable is influenced by the profit of the financial institutions, the trade balance, and the EUR/USD exchange rate in a solid way. The EUR/USD exchange rate tends to strengthen if the US economy is slowing down, which is in congruence with increasing housing inventories, the housing price decrease, and the declining profits of the financial institutions. Based on this, it could be expected that trade balance and the EUR/USD exchange rate are also in steady relationships; however, these tests show the opposite in both relations.

The retail sales of new passenger cars proved to be a stable variable, as it has stable relationships with all of its influencing indicators.

It is stated that the relationship between the consumer price index and the consumption expenditures on durables always breaks at times of recessions.

The indicators of money supply (M0, M1, M2) are in lagging relationships with their exogenous variables, the relationships always become weak after the start of the recessions.

There is a stable feedback relationship between the S&P500 index and the profit of the financial institutions to the GDP. This relationship states that long term price appreciation is not possible without the health of the financial institutions.

Based on the cointegration tests, it can be stated that the added value of the non-financial institutions is cointegrated in a constant parameter manner with three other variables; the consumer expenditures on services, the consumer expenditures, and GDP.

The ceasing character of the cointegration is proven in the case of the following variable pairs: *1*. Manufacturing production and the gross value added by businesses; *2*. Industrial production and constant price GDP; *3*. Private business sector production and manufacturing production.

Manufacturers durables new orders to the GDP and the private business sector production demonstrate the strength of the current recession, hence the cointegration parameters of the cointegrations equations are the same between Q1 1986 and Q3 1990 as well as Q1 1999 and Q1 2001.

The cointegration throughout the whole Q1 1986 to Q1 2010 period (referred to as holistic cointegration) between the government investment to GDP and the personal consumption stresses the automatism of the fiscal policy.

In the followings, the results are compared with the hypotheses set up.

First hypothesis: *The output indicators, in contrast with prior expectations, probably due to the self fulfilling prophecies, are not endogenous but exogenous variables*.

The result of Granger-causality tests showed that surprisingly the output indicators are strongly exogenous and they have feedback relationships with both the financial sector and monetary policy. Based on these facts, the first hypothesis can be accepted.

The self-fulfilling prophecies seem to be true in the case of the monetary policy as well; hence both the monetary base and the FED fund rate show exogenous characteristics. Special attention should be devoted to the monetary base, as its relationships with other indicators are stable. It has to be stated, however, that fiscal policy is an endogenous variable, which could be worrisome for those who are expecting an immediate result from fiscal policy.

Second hypothesis: *When examining particular variables, not only the ones followed by market participants but also others shall be analyzed, as it is not sure that the emphasized variable is carrying the most information. Besides, the emphasized role can change from period to period*.

This hypothesis can be accepted as well, mainly based on the behaviour of the employment indicators. These variables carry endogenous characteristics; however, they do not have any common influencing variables as Granger-causes. Apart from that, their regression relationships are not stable. Out of five employment indicators, only the regression relationships of the participation rate can be characterized. Still, the market does not follow this indicator.

Similar revelations can be made regarding the consumption of crude oil and that of energy. Firstly, the Granger-character of the two variables is totally different – the former is rather exogenous, the latter is endogenous –, secondly, energy consumption is basically in solid relationship with most of its variable pairs, while crude oil consumption has only one stable relationship. Thirdly, energy consumption is cointegrated with numerous variables, whereas crude oil consumption shows no cointegration.

These findings are important, because it shows that the focus of the financial markets is on the wrong indicator – huge volatily is occurring after the weekly crude stocks report (proxy indicator for consumption), but attention is paid to the energy consumption reports. All the traders want to know the non-farm payroll figures as they are released, however, as it seems from the foregoing, it is not a correct indicator to follow.

Third hypothesis: *Not the trend of macroeconomic indicators but the change in their interaction is able to indicate the start of a recession.*

This hypothesis is partially confirmed. The characteristic breaks of the regression relationships are only confirmed in the case of a few indicators. Regarding the output indicators, merely the constant price GDP and the output of the non-financial institutions to the GDP regressions break in a characteristic way.

As far as the employment indicators are concerned, only the regression relationships of the participation rate break in several cases, nor the lagging style break is proven. Within the fiscal, consumption, investment and commerce variable categories, the regression relationships of only the following variables break in a characteristic way: governmental consumption and investment to GDP, consumption expenditures for transportation, private investment, and aggregated industrial production. Neither the current account balance, nor the money market indicator has characteristic regression breakpoints, and the same can be stated for the price indicators.

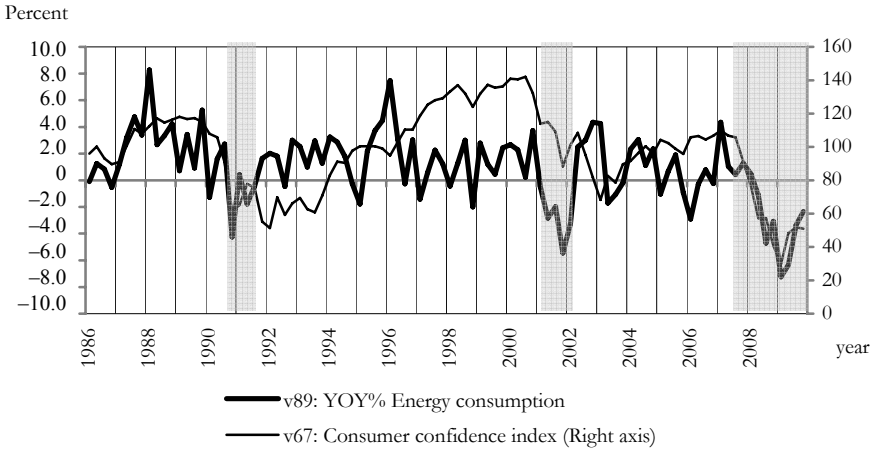
Fourth hypothesis: *It is presumable that, due to the necessary error correction mechanism among variables in dynamic relationships, the start of recession can be hidden or indicated later, and the same happens also in upswing times*.

This hypothesis is also partially proven, as only the holistic cointegration characteristics could cause these anomalies. However, holistic cointegration makes up only one third of the specific cointegration relationships, and it is mainly typical for consumption, investment, and commerce indicator groups.

Fifth hypothesis: *The explosion or birth of the dynamic relationships among variables carries useable information for the trend changes of the variables indicating recession*.

This hypothesis is clearly confirmed due to not only the numerosity of these cointegration relations but also the fact that starting and ceasing cointegrations are typical for specific indicators. Since the burst of the 'dotcom balloon', cointegrations ceased to exist in the case of the aggregated industrial production. However, the numerous cointegrations came to existence in the case of energy consumption and consumer satisfaction.

Energy Consumption and consumer confidence



Source: Thomson Reuters datastream.

In times of economic expansion, cointegrations that characterise the current price GDP, the output of non-financial institutes, energy consumption, the free cash flow of companies prognosticate new single family home sales and non-farm payroll costs.

Energy consumption proved to be a very important indicator of the state of the economy, as it is clearly endogenous in character, it has stable regression relationships, and carries numerous characteristic cointegrating relationships. Consumer satisfaction is to mention among exogenous variables. The importance of these two variables is enhanced by the fact that consumer satisfaction does Granger-cause energy consumption, the regression is stable, and since 2002 onwards these two variables has been cointegrated.

4. Conclusions

By using statistical methodology, the study identified variables and constellations that could have helped discovering the approach of an economic crisis. Based on the 'depth and width' of the database, it is presumable that only a fraction of the context was discovered.

The results support a further test on whether these variables can be 'wrapped into' latent variables (factors, main components) and the results can be generalised by means of these main components. However, it would need numerous changes in methodology, for example, in the dynamic factor models (*Tusnády–Zierman* [1987]).

Although my analysis was carried out at the macro level, it would be also worth analysing mezo- and micro-level data based on the usage of the same variables applicable to both industries and companies.

Since the credit crisis definitely started off from the US, this country was focused on. But it would be also important to check whether the same could be foreseen for other G7 economics.

References

CHOW, G. [1960]: Tests of Equality Between Sets of Coefficients in Two Linear Regressions. *Econometrica*. Vol. 28. No. 3. pp. 591–609.

DEWALD, W. G. [1972]: The National Monetary Commission: A Look Back. *Journal of Money, Credit and Banking*. Vol. 4. Issue 4. pp. 930–956.

DICKEY, D. A. – FULLER, W. A. [1979]: Distribution of the Estimators for Autoregressive Time Series with a Unit Root. *Journal of the American Statistical Association*. Vol. 74. Issue 366. pp. 427–431.

FABER, D. [2009]: *And Then the Roof Caved in*: *How Wall Street's Greed and Stupidity Brought Capitalism to Its Knees*. John Wiley & Sons. Hoboken.

FISHER, I. [1933]: The Debt Deflation Theory of the Great Depression. *Econometrica*. Vol. 1. No. 4. pp. 537–557.

FLECKENSTEIN, W. A. – SHEEHAN, F. [2008]: *Greenspan's Bubbles*: *The Age of Ignorance at the Federal Reserve*. McGraw-Hill. New York.

HAJDU, O. – HERMAN, S. – PINTÉR, J. – RAPPAI, G. – RÉDEY, K. [1994]: *Statisztika I-II*. JPTE Kiadó. Pécs.

HAJDU, O. – VIRÁG, M. [1993]: Pénzügyi viszonyszámokon alapuló vállalatminősítés többváltozós statisztikai módszerek felhasználásával. *Ipargazdaság*. Vol. 44. No. 7. pp. 23–32.

HARRIS, E. S. [2008]: *Ben Bernanke's Fed: The Federal Reserve After Greenspan*. Harvard Business Press. Boston.

HAUSMAN, J. A. [1978]: Specification Tests in Econometrics. *Econometrica*. Vol. 46. No. 6. pp. 1251–1271.

HUNYADI, L. [2004]: Wald-próba a regresszióban. *Statisztikai Szemle*. No. 9. Vol. 82. pp. 866–869.

JONES, J. D. [1986]: Consumer Prices, Wholesale Prices, and Causality. *Empirical Economics*. Vol. 11. Issue 1. pp. 41–55.

KINDLEBERGER, C. P. [1989]: *Manias, Panics and Crashes: History of Financial Crises*. Basic Books. New York.

KIRÁLY, J. – NAGY, M. – SZABÓ, E. V. [2008]: Egy különleges eseménysorozat elemzése – a másodrendű jelzáloghitel-piaci válság és (hazai) következményei. *Közgazdasági Szemle*. Vol. LV. No. July–August. pp. 573–621.

MINSKY, H. [1977]: A Theory of Systematic Fragility. In: *Altman*, *E. I. – Sametz*, *A. W.* (eds.): *Financial Crises*: *Institutions and Markets in a Fragile Environment*. Wiley International. New York.

MORRIS, C. R. [2008]: *The Trillion Dollar Meltdown*: *Easy Money, High Rollers, and the Great Credit Crash*. Public Affairs. New York.

ODED, M. – ROKACH, L. [2005]: *Data Mining and Knowledge Discovery Handbook*. Springer. New York.

RAPPAI, G. [2010]: A statisztikai modellezés filozófiája. *Statisztikai Szemle*. Vol. 88. No. 2. pp. 121–141.

RITZHOLZ, B. [2009]: *Bailout Nation*: *How Greed and Easy Money Corrupted Wall Street and Shook the World Economy*. John Wiley & Sons. Hoboken.

SIPOS, B. [1986]: A Kondratyev-ciklus empirikus vizsgálata és prognosztizálása. *Statisztikai Szemle*. Vol. 64. No. 12. pp. 1209–1237.

SOREN, J. [1995]: *Likelihood-Based Inference in Cointegrated Vector Autoregressive Models*. Oxford University Press. Oxford.

THE ECONOMIST [2006]: *Guide to Economic Indicators*: *Making Sense of Economic*. Sixth Edition. Profile Books. London.

TUSNÁDY, G – ZIERMANN, M. [1987]: *Idősorok analízise*. Műszaki Kiadó. Budapest.

A Method to Maximize the Information of a Continuous Variable in Relation to a Dichotomous Grouping Variable: Cutpoint Analysis

András Vargha
Professor
Károli Gáspár University
of the Reformed Church
in Budapest
E-mail: vargha.andras@kre.hu

Lars R. Bergman
Professor
Stockholm University
E-mail: lrb@psychology.su.se

In statistical analyses the researcher should normally use all the relevant information in the data. This argument has been used to advise against the habit of dichotomizing (approximately) continuous variables. However, if, for instance, a continuous variable is not normally distributed, it is possible that an optimal dichotomization can reveal relationships between variables otherwise obscured. Two analytical situations when this might apply were treated: *1*. The study of the relationship between an independent dichotomous grouping variable and a dependent continuous variable and *2*. the discrimination between two groups by identifying an optimal cutpoint in one or more continuous variables, treated as the predictor(s). For these purposes, cutpoint analysis (CPA) is introduced as a method for finding an optimal categorization of a continuous variable together with a computer package (ROPstat) to carry out the analysis. Three empirical examples are given that show the usefulness of CPA as compared to conventional analyses.

KEYWORDS:
Group comparison.
Best discriminating point.
Detailed comparison of distributions.

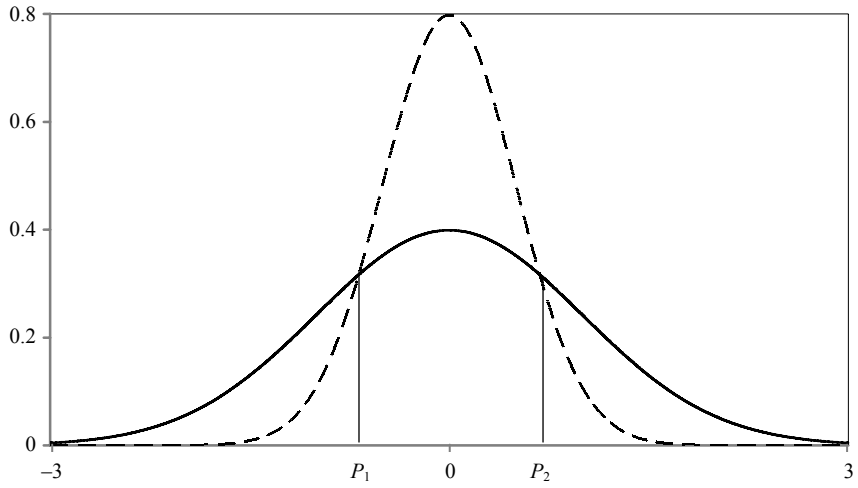
This paper takes its starting point from either of two analytical questions:

> *Case 1*: A dichotomous grouping variable is regarded as the independent variable and it is asked how two groups differ in a continuous dependent variable.
>
> *Case 2*: A grouping variable is regarded as the dependent variable and it is asked how well you can discriminate between the groups using information about the values of a continuous variable, which is regarded as the independent variable, that is, the predictor of group membership. Of course, in practice the continuous variable is usually not really continuous, by "continuous" we mean here a variable that is at an approximate interval scale level and taking many values, or is at least ordinal, taking many values.

Case 1 is often addressed by a two-sample t-test of the mean difference and Case 2 is frequently covered by discriminant analysis (DA) or logistic regression analysis (LRA). However, it must be assumed that the population means carry the important information about group differences in the first case and directly using the continuous variable is the best way of representing the information contained in it in the second.

We claim that sometimes in Case 1 or 2, the information value of the continuous variable is not maximized by treating it as an interval scale variable and working only with means but instead by an optimal categorization of its value range. By "information value" we mean all possible information by which one can draw conclusions from one variable to the other. Such a situation can occur in Case 1 when group membership relates to the value of the continuous dependent variable differently in various ranges of it. It might be well that there are no mean differences between groups but still the percentages belonging to one category of the categorized continuous variable vary between groups (see intervals $(-3; P_1)$, $(P_1; P_2)$, and $(P_2; 3)$ in Figure). This line of reasoning receives some support from the not infrequent finding that dichotomous variables, as compared to the corresponding continuous variables, can be surprisingly good at detecting relationships between variables (*Farrington–Loeber* [2000]). In Case 2 an appropriately categorized version of the continuous variable may be more useful to discriminate between the groups than the original variable. This can occur when there are threshold effects of the independent variable.

Two distributions with identical means but different dispersions



The purpose of the present paper is to present a method and a computer program to identify optimal cutpoints for categorizing a continuous variable and evaluating the usefulness of this categorization. This is done for each of the two cases described formerly and the method is called cutpoint analysis (CPA). We will only treat the case when the grouping variable is dichotomous but the findings are easily extended to the multi-group case.

First we present a brief discussion of the distributional conditions of the continuous variable in the different groups that must hold for the categorization approach to possibly detect group differences otherwise not detected or achieve a discrimination otherwise not achieved.

We note that in CPA the restrictive assumption of interval scaled continuity can be relaxed. The minimal constraint imposed is ordinality.

1. Necessary conditions for preferring an optimal categorization method

Case 1. Suppose one is interested in comparing Group A and Group B with regard to the values of a quantitative variable *X*. Let *X* be denoted in Group A by X_A, and in Group B by X_B. If X_A and X_B are normally distributed and $\sigma_A = \sigma_B$, the

only difference that can occur between the distributions of X_A and X_B is a mean difference, denoted by *d*:

$$F_A(x) = F_B(x + d), \qquad /1/$$

for the distribution functions F_A, and F_B for all possible values *x*, and for some value of *d*.

If the assumption of normality holds but variance homogeneity is strongly violated, large differences between the two distributions can occur even under $\mu_A = \mu_B$. (See Figure.)

If the assumption of normality is seriously violated, which occurs quite frequently (*Micceri* [1989]), then several situations can occur: Group A and Group B can differ in several special ways in terms of the distribution of *X* also when $\mu_A = \mu_B$. One such type of difference is when Group A members exist with a substantially lower or higher proportion under (or above) a certain value *c* than Group B members, that is when

$$F_A(c) < F_B(c) \text{ or } F_A(c) > F_B(c).$$

If such a cutpoint *c* exists on the scale of *X*, it suggests that different ranges of the scale of *X* represent different qualities. The exploration of one or more cutpoints over the range of *X* in relation to the distribution of X_A and X_B may then provide important information about the relationship between the independent and dependent variables.

It is also important to note that if X_A or X_B is non-normal, the inequality

$$F_A(c) \neq F_B(c) \qquad /2/$$

for some value of *c* can hold even if the population means of the two groups are equal, that is, $\mu_A = \mu_B$. Accordingly, the violation of normality is always an indication that differences between F_A and F_B other than a mean difference may occur.

Case 2. Whenever /2/ holds for some value *c*, this can serve as information for discriminating between Group A and Group B. If X_A and X_B are normally distributed, and variance homogeneity holds, then $\mu_A - \mu_B$ contains all information regarding the differences between Group A and Group B. However, if these assumptions are violated, identification of values *c* for which /2/ holds may suggest that there is a better rule for discriminating Group A and Group B than the one obtained in standard DA (*Farrington–Loeber* [2000]).

2. Description of the CPA method

Case 1. It is helpful if prior knowledge exists about defining the range in which the two groups differ. In a substantial portion of cases, however, the researcher has no idea about the locations of the optimal cutpoints over the range of the dependent variable *X* but still their existence might be surmised and merit investigation. A simple solution would be to compare the two distributions in all of the observed values of *X*, but this approach would necessarily yield high alpha error inflation, which is statistically unacceptable. By the CPA method presented here one can search for cutpoints discriminating sharply the two distributions without the danger of alpha inflation.

The main idea of CPA is as follows.

1. Chose a limited number of cutpoints (c_1 , …, c_k) from the value range of variable *X*.

2. For each c_i ($i = 1, \ldots, k$) dichotomize *X* at cutpoint c_i, defining variable X_i as follows: $X_i = 0$ if $X < c_i$ and $X_i = 1$ if $X \geq c_i$.

3. Compare Group A and Group B in terms of each X_i by performing a 2×2 chi-square test or a Fisher-exact test, determining the *p*-value of the significance.

4. In order to avoid alpha inflation, multiply *p* by *k*, the number of all cutpoints, that is, the number of performed tests: $p_{adj} = k \cdot p$. This is the well-known Bonferroni method for adjusting *p*-values in multiple comparisons (*Maxwell–Delaney* [2004] pp. 202–208).

The only questions left open are the choice of the value *k* and the selection of the c_i ($i = 1, \ldots, k$) cutpoints. Whenever variable *X* has k_X different values, and k_X is less than k_{max}, the maximal allowed value for *k* is set to $k = k_X$. In all other cases, set $k = k_{max}$. As a rule of thumb – based on our empirical experience – we suggest that the maximal allowed value for *k* be 10 in most comparisons, that is, $k_{max} = 10$. Increasing the value of *k* would decrease the power of CPA, and a decrease of its value may increase the chance that relevant cutpoint(s) will be unidentified. In a later section we will also provide some empirical support to this choice.

In the identification process of the *k* cutpoints, we propose the following two criteria if the number of different values of variable *X* in the sample exceeds k_{max} :

1. Let $c_1 \geq x_\varepsilon$ and $c_k \leq x_{1-\varepsilon}$, where x_ε and $x_{1-\varepsilon}$ are percentile values in the empirical pooled distribution of variable *X* with certain small ε

values ($\varepsilon < 0.20$). Hence, we do not compare the groups at the lower and upper ε part of the pooled distribution. The value of ε can be fixed freely by the user – its recommended value may be between 0.01 and 0.05, depending on the sample size (in larger samples ε can be smaller, enabling CPA to detect differences between the distributions over a larger range of *X* values). In ROPstat the default value for ε is 0.025.

2. The estimated $P(c_i \le X \le c_{i+1})$ probabilities, where $1 \le i \le k-1$, should be as similar as possible in the total sample containing both groups.

The description of the computer program can be found in the Appendix. For more details about the program output, see the empirical examples provided in the next section.

Case 2. If the grouping variable is regarded as the dependent variable, an important aim of the analysis can be to predict group membership based on the value of the continuous variable. This model is well-known, for instance, in research evaluation and comparison of the performance of diagnostic tests (*DeLong–DeLong–Clarke-Pearson* [1988]).

The key concepts are as follows. One of the two groups is regarded as the criterion group, the other as the control group. Based on the continuous variable *X*, it is decided if a subject belongs to the criterion group or not. The decision is made by means of a threshold value x_c, on the scale of *X*. Subjects having *X* values greater than or equal to x_c will be regarded as belonging to the criterion group. A threshold value x_c works well if most subjects from the criterion group will be judged as belonging to it, in other words, if

$$\text{Sensitivity}\,(x_c) = Pr\left(X \ge x_c \middle| \text{criterion group}\right) \qquad /3/$$

is close to 1, and most subjects from the control group will be judged as not belonging to the criterion group, that is, if

$$\text{Sensitivity}\,(x_c) = 1 - Pr\left(X \ge x_c \middle| \text{control group}\right) \qquad /4/$$

is close to 1. The choice of an appropriate threshold value x_c can be made by means of a receiver operating characteristic (ROC) curve (*DeLong–DeLong–Clarke-Pearson* [1988]).

For an optimal x_c value, the $Pr(X \ge x_c)$ proportions for the criterion and control group might differ substantially. Accordingly, cutpoints identified in a CPA will

carry diagnostic information for discriminating the two groups. Starting from several continuous variables to be used for the group discrimination and then creating one or more new dichotomous variables by means of the identified cutpoints, the set of these derived variables may serve as predictor variables in a DA or LRA for arriving at an efficient discrimination.

3. Examples

Example 1. *The femininity of applicants to psychology major*. In an examination of admittance to the Psychology major at Eötvös Loránd University, Budapest in 1981, the number of males and females were $m = 16$, and $n = 78$, respectively. Among these 94 applicants, 12 males and 70 females filled out a short Hungarian version (including 300 items) of the California Personality Inventory (SCPI) (*Oláh* [1985]). One of the scales of SCPI is "femininity" (Fem), which informs about the feminine character and focus of the interest of the subject. In order to test the validity of this scale, we compared the male and female samples. For testing the equality of theoretical means (the sample means were 12.08 for males and 14.00 for females), the two-sample *t*-test ($t(80) = 2.954$, $p = 0.0041$) and the Welch test ($W(13) = 2.372$, $p = 0.0337$) were applied. For examining the stochastic equality of males and females, the Mann–Whitney test ($z = 2.339$, $p = 0.019$) and the Brunner–Munzel test ($BM(12) = 2.108$, $p = 0.0566$) were performed (about stochastic equality, see *Vargha– Delaney* [1998], [2000]). The estimated *A* measure of stochastic superiority, which assesses the stochastic dominance of males versus females in terms of the Fem scale was 0.29. This shows that if we compare two randomly selected male and female persons among the applicants, the chance of having a larger male Fem score is about 0.29, and the chance of having a larger female score is about 0.71.

For a detailed comparison of the two distributions, a cutpoint analysis was performed by means of the group comparison module of ROPstat. The program divides the scale of the dependent variable, *X* into many narrow intervals so that the cutpoints define intervals with as equal proportion of the total sample as is possible, and if the number of the different values of *X* does not exceed 100, each value will be placed in the inner part of a separate interval. If the number of different values of *X* exceeds 100, some values of *X* may fall to the edges of these intervals. The program computes the value of the empirical distribution function for the upper limits *c* of these intervals separately for the compared groups (males, $F1(c)$ and females, F2(*c*)), and tests their difference at $k \leq 10$ different points. In the present example,

the program found 10 different values (between 8 and 18), but for the two lowest c values (8.05 and 10.05) the pooled cumulative percentage value was less than 0.05, and for the largest (18.05) was greater than 0.95. Hence in this case $k = 7$ and $\varepsilon = 0.05$. Based on the F1(c) and F2(c) values corresponding to the selected 7 cutpoints, the program computes for each c the φ contingency-coefficient measuring the strength of association between the grouping variable (in the present case "gender") and the dichotomized dependent variable (in the present case femininity), using either the 2x2 chi-square test or the Fisher-exact test with the corresponding unadjusted and adjusted two-tailed probability values. Due to the relatively small sample sizes in the present case, the Fisher-exact test was always performed. The results are summarized in Table 1.

Table 1

Results from a cutpoint analysis comparing males and females based on their California Personality Inventory/femininity level
($n = 82$)

Detailed point-wise comparison of the two distribution functions							
c	$F1(c)$	$F2(c)$	$F1–F2$	Phi	Chi|Fish	p-value	Adjusted p
8.05	0.083	0	0.083	0.27			
10.05	**0.417**	**0.029**	**0.388**	**0.49**	**Fisher**	**0.0005**	**0.0036***
11.05	0.500	0.100	0.400	0.39	Fisher	0.0027	0.0191**
12.05	0.583	0.229	0.355	0.28	Fisher	0.0314	0.2196
13.05	0.667	0.414	0.252	0.18	Fisher	0.1259	0.8814
14.05	0.750	0.629	0.121	0.09	Fisher	0.5250	1.0000
15.05	0.833	0.757	0.076	0.06	Fisher	0.7232	1.0000
16.05	1.000	0.900	0.100	0.13	Fisher	0.5861	1.0000
17.05	1.000	0.943	0.057	0.09			
18.05	1.000	1.000					

Note. The significance of phi is tested via chi-square or Fisher-exact test. Tail probability (p) is adjusted by means of Bonferroni method. * for $p < 0.05$ ** stands for $p < 0.01$;

To analyze the identity of the two distributions, Kolmogorov–Smirnov's two-sample test is applied: $J^* = 1.280$ ($p = 0.0754$)

The format of the table follows that of the corresponding computer output of ROPstat with some modifications.

Based on the results summarized in Table 1, the most significant difference between the two genders was obtained for the cutpoint $c = 10.05$ (the corresponding row in Table 1 is indicated by a bold type face). Below this value (that is in the range

0–10) we find 41.7 percent of males and 2.9 percent of females. The difference is 39 percent, which is highly significant (adjusted $p = 7 \cdot 0.0005 \approx 0.0036$). It is interesting that the Kolmogorov–Smirnov test shows only a tendency to significance ($p = 0.0754$). This weakness of the test is characteristic and is due to the fact that it performs a global comparison of two distributions, taking into account every possible type of difference, whereas CPA focuses on a small number of potentially informative cutpoints.

The psychological explanation of the obtained results may be as follows. The femininity of males and females differs from each other mostly in the fact that there exists a certain level of minimal femininity ($\mathrm{Fem} = 10$), below which we find almost exclusively males. Among the females almost everybody (in the present sample 68 out of 70) shows this minimal level of femininity. Such a strong differentiation, however, does not occur at the higher region of the Fem scale, which means that there is not a high level of femininity that would mainly be characteristic of females. This information is not revealed by standard analyses.

Example 2. *The relationship of birth rank to personality*. This study concerns the relationship of birth rank to adult personality (*Mózes–Vargha* [2007]). Studying women, we compared first born subjects ($m = 35$) with the rest of the sample ($n = 49$) in terms of six scales of Parker's parental bonding instrument (*Parker* [1989], [1990]). Among these six scales we present results concerning father's care. The measure is "retrospective", meaning that the women reported how they remember their father cared for them during their first 16 years.

Table 2

Comparison of first born and other women's responses to the father's care scale in the parental bonding instrument
($n = 84$)

Group	Size	Mean	SD	Minimum	Maximum	Skewness	Kurtosis
First born women	35	23.17	11.44	0	36	–0.829*	–0.428
Other women	49	21.96	8.075	3	34	–0.651+	–0.508

Note. Dependent variable is father's care. The significant skewness indicated non-normality. + stands for $p < 0.10$; * for $p < 0.05$.

Testing the equality of population variances: *1*. O'Brien test (Welch type): $F(1.0;\ 45.6) = 4.851$ ($p = 0.0327$); *2*. Levene test (Welch type): $F(1;\ 55.4) = 3.645$ ($p = 0.0614$).

Testing the equality of population means: *1*. Two-sample *t* test: $t(82) = 0.570$ ($p = 0.5704$); *2*. Welch's modified t test: $W(57.4) = 0.538$ ($p = 0.5924$).

The format of the table follows that of the corresponding computer output of ROPstat with some modifications.

For testing the equality of theoretical means, two-sample *t*-tests were applied but none of them indicated any significant difference between the two groups. (See Table 2.) Nonparametric rank tests comparing the two groups were also far from being significant ($p > 0.20$). However, in the present case, both the normality assumption and the variance homogeneity assumption are violated. On the one hand, this may invalidate the two-sample *t*-tests, on the other hand, it raises the possibility that some other types of differences may appear using CPA. The results of this analysis are summarized in Table 3.

Table 3

Results from CPA comparing first born and other women in terms of the father's care scale in the parental bonding instrument
(n = 84)

Detailed point-wise comparison of the two distribution functions							
c	$F1(c)$	$F2(c)$	$F1–F2$	Phi	Chi|Fish	p-value	Adjusted p
0.18	0.057	0.000	0.057	0.18			
1.26	0.086	0.000	0.086	0.23			
3.06	0.114	0.020	0.094	0.20			
5.22	0.143	0.041	0.102	0.18	Fisher	0.1223	1.0000
6.30	0.171	0.041	0.131	0.22			
7.02	0.171	0.061	0.110	0.18			
8.10	0.171	0.082	0.090	0.14			
9.18	0.171	0.102	0.069	0.10			
10.26	0.171	0.122	0.049	0.07			
11.34	0.171	0.143	0.029	0.04			
12.06	0.171	0.163	0.008	0.01			
13.14	0.200	0.184	0.016	0.02	Fisher	1.0000	1.0000
14.22	0.229	0.224	0.004	0.00			
16.02	0.229	0.245	–0.016	–0.02			
18.18	0.257	0.306	–0.049	–0.05	Fisher	0.8069	1.0000
19.26	0.286	0.347	–0.061	–0.06			
20.34	0.314	0.367	–0.053	–0.06			
21.06	0.343	0.388	–0.045	–0.05	Fisher	0.8191	1.0000
22.14	0.371	0.449	–0.078	–0.08			
23.22	0.429	0.510	–0.082	–0.08	Fisher	0.5112	1.0000
24.30	0.486	0.510	–0.024	–0.02			
25.02	0.514	0.612	–0.098	–0.10	Fisher	0.3826	1.0000
26.10	0.514	0.633	–0.118	–0.12			

(Continued on the next page.)

(Continuation.)

Detailed point-wise comparison of the two distribution functions

c	$F1(c)$	$F2(c)$	$F1-F2$	Phi	Chi|Fish	p-value	Adjusted p
27.18	0.600	0,673	–0.073	–0.08	Fisher	0.4993	1.0000
28.26	0.629	0.776	–0.147	–0.16	Fisher	0.1522	1.0000
29.34	0.629	0.816	–0.188	–0.21			
30.06	0.629	0.857	–0.229	–0.26	Fisher	0.0202	0.2017
31.14	0.657	0.939	–0.282	–0.36			
32.22	0.714	0.959	–0.245	–0.35			
33.30	**0.714**	**0.980**	**–0.265**	**–0.39**	**Fisher**	**0.0005**	**0.0051****
34.02	0.914	1.000	–0.086	–0.23			
36.18	1.000	1.000					

Note. To test the identity of the two distributions, Kolmogorov-Smirnov's two-sample test was applied: $(J^* = 1.273)$ ($p = 0.0784$).

The format of the table follows that of the corresponding computer output of ROPstat with some modifications.

$F1$ refers to the distribution function for first born women and $F2$ to the corresponding function for other women. + stands for $p < 0.10$; * for $p < 0.05$, and ** for $p < 0.01$.

In our case the best discriminating point is $c = 33.30$ (the corresponding row in Table 3 is indicated by a bold type face). A lower value, that is, $X \le 33$ occurred for 71.4 percent (25 out of 35) of first born women, and 98 percent (48 out of 49) of other women. These two proportions differ from each other significantly (the two-tailed probability of the Fisher-exact test is $p = 0.0005$). This is highly significant even after performing the Bonferroni adjustment ($p_{adj} = 0.005$). Accordingly, we can claim that first born women are significantly more likely (in the present sample the chance is 28.6 percent) to report an extreme high level ($X > 33$) of father's care as compared to other women (in this latter sample the chance is 2 percent). The conclusion is that a very high level of experienced father's care is almost only found among first born women.

Example 3. *Discrimination of psychotic and normal women by means of psychiatric rating scales*. In the framework of a longitudinal study launched in 1967, 230 psychotic and 41 mentally normal women were investigated by means of *Overall's* [1968] factor construct rating scale (FCRS) and *Rockland* and *Pollin's* [1965] questionnaire (RPQ) for psychiatric rating (*Pethő* [2001]). In the current analysis we used 17 elementary scales of FCRS ($F1$, ..., $F17$) and 34 elementary scales of RPQ ($R1$, ..., $R33$, $R35$). A value of zero on these scales reflects the lack of some psychiatric symptom, and values close to the maximum show the strong presence of a symptom. Preliminary analyses indicated that several of the scales were non-normally distrib-

uted. For these variables we addressed the following question concerning the discrimination of psychotic and normal subjects: If we dichotomize the continuous variables, using cutpoints identified by a CPA, and then perform DA and LRA, will we arrive at a better group discrimination as compared to conventional analyses based on the continuous variables? The following statistical analyses were undertaken:

1. First a CPA was carried out for all continuous variables. In the subsequent analyses we retained only those for which the CPA revealed at least one significant cutpoint. For these scales we performed a dichotomization at the cutpoint that had the lowest *p*-value. These scales were as follows: *F*1–*F*14, *F*16, *F*17, *R*1–*R*5, *R*9, *R*11–*R*16, *R*18, *R*20–*R*23, *R*25–*R*30, and *R*33, altogether 40 variables.

2. Subsequently, we performed stepwise DA and LRA with first the original 51 continuous variables, then with the 40 dichotomized variables to predict group membership. The results are summarized in Table 4.

Table 4

Percentage of correct identifications in stepwise discriminant analyses and binary logistic regression analyses for the factor construct rating scale and Rockland and Pollin's questionnaire scales in original and dichotomized form

Group	Discriminant analysis with original variables	Discriminant analysis with dichotomized variables	Logistic regression analyses with original variables	Logistic regression analyses with dichotomized variables
Psychotic (n = 230) (percent)	78.7	87.0	95.7	95.7
Normal (n = 41) (percent)	92.7	100.0	82.9	85.4
Total (percent)	80.8	88.9	93.7	94.1
Number of selected variables	11	7	10	8

Based on Table 4 we can draw the following conclusions.

1. Using non-normal independent variables in DA may lead to substantially weaker discrimination than LRA.

2. Using derived dichotomized variables may lead to surprisingly good results parallel to those found using the original variables, and CPA can be an efficient tool for identifying appropriate cutpoints for the dichotomization. As an example, we obtained 88.9 percent correct identification percentage in DA with 7 selected dichotomous variables, compared to 80.8 percent with 11 original variables.

3. Also in LRA the dichotomized variables performed well (with 8 variables resulting in 94.1 percent correct classifications, as likened to using 10 original variables resulting in 93.7 percent correct classifications).

4. Summary and conclusion

It is an almost trivial observation that in statistical analyses the researcher should normally use all the relevant information in the data. In the literature there are many arguments against the habit of dichotomizing continuous variables, which is usually performed for the purpose of simplifying the analyses and presentation or for handling interactions. This attitude is seen in its most extreme form in an editorial in the *Journal of Consumer Research*, entitled "Death to Dichotomizing" (*Fitzsimons* [2008]).

The warnings against dichotomization are often good advice but the arguments build on assumptions of normality and linearity (for example *Cohen* [1983], *Maxwell–Delaney* [1993]). If these assumptions are valid, the argument against dichotomization seems solid, however, frequently psychological variables do not follow the normal distribution (*Micceri* [1989]) and the relationships might not be linear. In such situations it is possible that the arguments against dichotomization of a continuous variable break down. For instance, take the case of studying the relationship between one continuous independent variable, regarded as the risk factor, and one continuous dependent variable, regarded as the outcome. Theoretically, it is possible that there is a threshold effect in the independent variable so that there is no risk increase for a bad outcome below a certain level of the value in the risk factor but then, suddenly, a strong increase in the risk occurs. Or it is possible that the outcome is really generated by a normal mixture model with a relationship between the risk factor and the distribution membership.

Within the context discussed formerly, we examined two analytical situations where dichotomization may be appropriate. The first one concerned the study of the relationship between a dichotomous grouping variable, regarded as the independent variable, and a continuous variable, considered as the dependent variable. We devised a dichotomization method, cutpoint analysis, in which a limited number of cutpoints (usually not exceeding 10) in the dependent variable distribution are used for different dichotomizations, selecting the one that maximizes the association between the independent variable and the dependent variable. In two empirical examples, one concerning gender and femininity and the other regarding birth rank and personality, the results indicated significant relationships not revealed by standard analyses.

These findings were partly explained by clear departure from normality in the dependent variable and by the fact that the relationship had a different form in different regions of the variables.

The second analytical situation that we discussed and where dichotomization may be appropriate concerned the discrimination between two groups by identifying an optimal cutpoint in one or more continuous variables, treated as the predictor(s). CPA can then be used to find an optimal dichotomization of the continuous variable(s) in the sense that the prediction of group membership is maximized. In a third empirical example, CPA was used for dichotomizing a number of psychiatric rating scales that were used in DA or LRA. This resulted in a higher or at least as high discrimination power between psychotic and normal women as was achieved using the original continuous variables. It appears that, for discrimination purposes, the essential information in the scales was largely binary, of qualitative nature.

DeCoster, Iselin, and *Gallucci* [2009] revealed also several situations in which the use of dichotomization is appropriate. Specifically, they argue that it is acceptable for researchers to use dichotomized indicators in the following circumstances:

> *1*. The study uses extreme group analysis.
>
> *2*. The purpose of the research is to investigate how a dichotomized measure will perform in the field.
>
> *3*. The underlying variable is naturally categorical, the observed measure has high reliability, and the relative group sizes of the dichotomized indicator match those of the underlying variable.

CPA is similar to the search for an "optimal cutpoint" in biostatistics. A common aim in biomedical research is to investigate whether a certain continuous variable (regarded as covariate) has potentially prognostic relevance for a time outcome dependent variable like survival. The optimal cutpoint is that value of the covariate, which corresponds to the most significant relationship between the covariate dichotomized at this cutpoint and the dichotomous outcome variable (*Heinzl–Tempfer* [2001]). Since the term "optimal" may give the false impression that this method is superior to other ones, *Altman*, *Lausen*, *Sauerbrei*, and *Schumacher* [1994] suggested that the method be called the "minimum *P*-value approach". When using this method, some researchers are ready to ignore the multiple testing and alpha inflation problems since their decisions are based on the unadjusted *p*-value of the "optimal" cutpoint. This practice may lead to inconsistent results in medical prognostic research (*Heinzl–Tempfer* [2001]). In contrast, CPA is protected against alpha inflation by performing only a limited number of two-group comparison tests (≤ 10) of scattered cutpoints and by applying a Bonferroni adjustment to the *p*-values of the selected cutpoints.

For comparing two survival curves, the most common statistical test is the logrank test. This is a type of chi-square test, asymptotically equivalent to the likelihood ratio test, which is based on observed and expected frequencies of a certain time event belonging to different time points of the two survival curves (*Bland–Altman* [2004], *Mantel* [1966], *Schoenfeld* [1981]). The need for applying some adjustment on the *p*-values of repeated logrank tests is now increasingly recognized (*Altman et al*. [1994], *Heinzl–Tempfer* [2001], *Williams et al*. [2006]). Such an adjustment can be achieved by the following formula for an adjustment of the minimal *P*-value (p_{min}) valid for large sample sizes, to allow for the multiple testing thanks to Lausen and Schumacher ([1992], [1996]), *Miller* and *Siegmund* [1982], and *Heinzl* [2000]:

$$P_{cor} = \varphi(z)\left(z - \frac{1}{z}\right) ln\left(\frac{(1-\varepsilon)^2}{\varepsilon^2}\right) + \frac{4\varphi(z)}{z}. \qquad /5/$$

Here P_{cor} denotes the adjusted (corrected) minimum *P*-value of the logrank statistic, φ is the standard normal density, z is the $\left[1-(P_{min}/2)\right]$-quantile of the standard normal distribution and ε is defined as follows. The minimum *P*-value approach requires the choice of a selection interval. It is defined by the ε and $(1-\varepsilon)$-quantile of the observed values of the continuous covariate $(0<\varepsilon<0.5)$. Values outside the selection interval are not considered as potential cutpoints. In CPA we also seek potential cutpoints between C_ε and $C_{1-\varepsilon}$ percentiles, so the meaning of ε in CPA is the same.

By means of formula /5/, we compared the power of the above adjustment rule for $\varepsilon = 0.01$, 0.05 and 0.10 with CPA for a set of different nominal *p*-values, allowing for as many as 10 cutpoints ($k = 10$) in CPA. Results summarized in Table 5 show that CPA is much more efficient in detecting possible differences of the two distributions to be compared than the one defined by formula /5/. This is reflected by the fact that the P_{cor} values are substantially higher – in most cases more than twice as large – than the corresponding P_{adj} values of CPA for unadjusted alpha values less than 0.05. Due to this, the cutpoints of CPA can be more easily significant despite their strictly controlled Type I error level. However, we agree with Heinzl and *Tempfer* [2001] that without any biological (clinical, psychological, etc.) indications for the actual existence of a cutpoint even the correct application of the minimum *P*-value approach as well as CPA has to be considered methodologically questionable.

Table 5

Comparison of three corrected p-values and the adjusted p-value of CPA based on the Bonferroni method for different unadjusted nominal p-values

Unadjusted *p*-value	$P_{cor}(\varepsilon = 0.01)$	$P_{cor}(\varepsilon = 0.05)$	$P_{cor}(\varepsilon = 0.10)$	$P_{cor}(k = 10)$
0.10	1.0000	0.8806	0.7208	1.0000
0.05	0.8980	0.6183	0.4916	0.5000
0.01	0.3132	0.2087	0.1615	0.1000
0.005	0.1859	0.1231	0.0946	0.0500
0.001	0.0509	0.0334	0.0255	0.0100
0.0005	0.0285	0.0186	0.0142	0.0050
0.0001	0.0071	0.0046	0.0035	0.0010

Note. *P*-values are based on based on formula /5/ (P_{cor} with $\varepsilon = 0.01$, 0.05, and 0.10).

A special type of the minimum *P*-value approach is the following method. Two groups are compared by means of a quantitative variable the same way as in CPA, looking for an "optimal" cutpoint. In this approach a cutpoint is regarded as optimal if the usual chi-square statistic computed from a 2×2 table based on the frequencies below and above the cutpoint in the two groups is maximal. Miller and Siegmund [1982] investigated the asymptotic distribution of this maximally selected chi-square statistic and provided tail probabilities and critical values for its significance for different nominal alpha levels and selection interval defined by ε and $(1-\varepsilon)$ the same way as mentioned formerly (see Tables 1 and 2 in *Miller–Siegmund* [1982]). Since the computation of these values is built on the same formula (see formula /8/ in *Miller–Siegmund* [1982] that appears in /5/), the superiority of CPA over this approach in terms of power still remains. With the same method, *Koziol* [1991] provided better critical values and tail probabilities based on the exact finite-sample distribution theory, *Betensky* and *Rabinowitz* [1999] generalized the asymptotic distribution of the maximally selected chi-square statistic for the multi-group case, and *Boulesteix* [2006] generalized the results of Koziol to any ordinally scaled dependent variable in the two-group-comparison case.

The P_{cor} corrected *p*-values of the minimum *P*-value approach refer to the significance of the most significant cutpoint that discriminates the two groups based on the continuous dependent variable, whereas the P_{adj} adjusted *p*-values of the CPA approach refer to the significance of all *k* cutpoints identified in CPA. As we could see formerly, in the practically relevant cases, when the corrected/adjusted *p* values are close to significance (this is true when the unadjusted *p*-values are less than or

equal to 0.01; see Table 5), the adjusted *p*-values of CPA are always substantially smaller than those of the corrected *p*-values of the minimum *P*-value approach and for this reason they can detect differences between the two distributions to be compared with a higher efficiency. This is completely true for dependent variables where the number of different values do not exceed 10 (in this case CPA compares the two distributions with all possible cutpoints). However, if the dependent variable is really continuous and has large many different values, it may well happen that the selected set of cutpoints in CPA does not include the value of the dependent variable which discriminates most significantly the two groups (distributions). To have some information about how often this unlucky situation may arise, we carried out the following empirical investigation.

From an archival data set including 811 Rorschach-protocols that served as the basis for the construction of the Hungarian Rorschach Standard (*Vargha* [1989]), we selected 236 quantitative variables of elementary Rorschach scores and computed indices. The elementary scores were relative frequencies of different Rorschach responses (referring to the location, determinant, content category, popularity, or originality of the response, etc.). These relative frequencies were computed by dividing the number of occurrences of different Rorschach-items with the number of total response number. As an example, the value of the Anat% Rorschach-variable was obtained for a specific person by dividing the number of anatomical responses occurring in the protocol by the total number of responses. More than 70 percent of these Rorschach-variables was practically continuous, having more than 10 different values. Out of the 811 protocols, 363 originated from mentally normal persons (MN), while the other 448 from institutionalized non-psychotic patients (INP).

In order to assess merits and weaknesses of CPA, the two groups (MN and INP) were compared for each of the 236 Rorschach-variables:

1. performing a CPA (with parameters $k = 10$ and $\varepsilon = 0.01$);

2. identifying the best discriminating point, that is the cutpoint within the middle $(\varepsilon; 1-\varepsilon)$ part of the scale of the dependent variable for which the tail probability of a 2×2 chi-square test (or the two-sided *p*-value of the Fisher-exact test if the minimal expected cell frequency in the 2×2 table does not exceed 20) is the smallest;

3. performing the Kolmogorov–Smirnov two-sample test for a global comparison of the two distributions.

The results of the performed analyses can be summarized as follows.

4. For 157 out of the 236 dependent variables there were more then 10 different values within the middle $(\varepsilon; 1-\varepsilon)$ part of the scale of the dependent variable. Out of these 157 variables, the smallest adjusted

P-value was less than or equal to 0.10 for 68 variables, and in 53 cases (78%) out of the 68 variables the mostly significant cutpoint was identical with the CPA cutpoint for which the tail probability of the 2×2 chi-square test was minimal. This means that the set of the *k* cutpoints of CPA contained the best discriminating point of the dependent variable in the large majority of cases. However, if one wants to decrease the risk of missing a relevant cutpoint, the value of *k* can be increased even above 10. Based on data of Table 5, one can conclude that even an increase by 50 per cent can keep the advantage of the CPA method over the alternative methods (for example maximizing chi-square).

5. For a comparison of the efficiencies of CPA and the Kolmogorov–Smirnov two-sample test, we cross-tabulated the significances of the Kolmogorov–Smirnov test and the most significant cutpoint (adjusted probability) of CPA for the 236 Rorschach-variables. (See Table 6.) From Table 6 it seems to be evident that CPA highly outperforms the Kolmogorov–Smirnov in terms of power. It occurred only four times out of 236 cases that the Kolmogorov–Smirnov test was significant (3 times at 10 per cent and once at 5 per cent level), but the CPA was not, whereas the opposite situation occurred in 38 cases. In addition, in 31 (out of the 38) cases, the CPA was significant at least two levels stronger than the Kolmogorov–Smirnov test (the opposite situation never occurred)

Table 6

Cross-tabulation of the significances of the Kolmogorov–Smirnov test and the most significant cutpoint of the cutpoint analysis for 236 different quantitative Rorschach variables

Kolmogorov–Smirnov test	CP					
	$p > 0.10$	$p < 0.10$	$p < 0.05$	$p < 0.01$	$p < 0.001$	*Total*
$p > 0.10$	137	18	12	5	3	*175*
$p < 0.10$	3	7	2	5	0	*17*
$p < 0.05$	0	1	6	7	6	*20*
$p < 0.01$	0	0	0	6	5	*11*
$p < 0.001$	0	0	0	0	13	*13*
Total	*140*	*26*	*20*	*23*	*27*	*236*

CPA can be regarded as a multiple test like post-hoc analyses in ANOVA. The *k* cutpoints are selected based on the distributional characteristics of the pooled sam-

ple, independently from the differences between the two groups. The application of the well-known Bonferroni method guaranties that if any of the cutpoints is significant at an adjusted alpha level, then the probability of Type I error (of the null hypothesis of the equality of the two distributions binarized in this cutpoint) will not exceed α. Hence, if any cutpoint of CPA is significant, the two distributions can confidently be declared different from each other. Since CPA gave significant results in many more cases than the Kolmogorov–Smirnov test, it seems in this context to be more appropriate for detecting differences. It is also important to add that CPA not only detects efficiently the inequality of the two distributions but identifies also the cutpoints where the differences are most salient.

It should be noted that if the same data set is used both to dichotomize variables by CPA and to estimate a regression model with the dichotomized variables as predictors, the corresponding regression coefficients will obviously be biased. In such cases we suggest that, if the sample size is large enough, a portion (say 2/3 of the sample) be used for exploration and the rest for the verification of the model. If the sample is relatively small, the results of CPA need confirmation in an independent study. We assert, however, that, at a minimum, CPA is a simple but effective way of deriving hypotheses to be confirmed in future studies.

To sum up, CPA is a new technique and software for finding efficiently truly significant dichotomizing points in a quantitative variable that maximizes the association to another dichotomous variable, which might otherwise be hidden if one were to use conventional statistical approaches.

In the present article, only two special cases were treated but we believe that the reasoning employed merits consideration also in other ones. The most obvious extensions are to the case where the grouping variable is not dichotomous and to the one where the relationship is studied between two variables while controlling for a third. ROPstat can handle the multigroup case of CPA whose nice empirical illustration can be found in *Borbély–Vargha* [2010].

Appendix

The description of the computer program

We implemented CPA in the group comparison module of ROPstat, a new statistical program package. It is a user-friendly statistical software that is rich in robust techniques and procedures with ordinally scaled variables, and includes a number of procedures for pattern and person oriented analysis.

Its free demo version can be downloaded from the site www.ropstat.com by clicking on the text "Download and test DEMO version in English". A description of the package can also be found there. A CPA is carried out in the following way.

1. Run the downloaded setup program of ROPstat. As a result, a folder called "c:_vargha\ropstat" will be created and a program called "ropstat.exe" will be installed in it.

2. Run ropstat.exe.

3. Open an input data file by means of icon "Open". ROPstat has a special option accepting SPSS portable files imported from SPSS in *.por format and tab-delimited files imported from Excel in *.txt format. The DEMO version of ROPstat accepts at most 5 variables and 500 cases but otherwise performs complete statistical analyses.

4. After loading a data file, click on the menu point "Statistical analyses", and within it the submenu points "Comparing groups or variables" and "One-way comparison of independent samples".

5. In the appearing program window, put the given continuous variable (X) from the list of variables to the box of "Dependent variables", and a grouping variable having two code values (or defined by two intervals in the variable characteristics window of the data sheet) to the box of "Grouping variable".

6. In the box of "Scale type"', change the scale type of the dependent variable from "interval" to "ordinal", and change the option of "Detailed comparison of distributions" in this program window from "No" to "Yes".

7. When you click on the icon "Run" in the bottom of the program window, a list of the following results will appear in a text window:

a) Nonparametric group comparison with the classical Mann–Whitney test.

b) Two robust alternatives of the Mann–Whitney test (Brunner–Munzel and corrected Fligner–Policello tests).

c) Detailed point-wise comparison of the two distributions (CPA). The two groups are compared by a 2×2 chi-square test if the minimal expected cell frequency exceeds 20, otherwise by the Fisher-exact test. Here $k = min(10, \text{number of different values of } X)$.

d) Kolmogorov–Smirnov's two-sample test for a global comparison of the two distributions.

8. In the CPA part of the output, the rightmost column with the header "Adjusted p" will contain the adjusted p values for the point-wise comparisons. If such a p value is less than 0.05, the corresponding score of the test variable can be regarded as a significant cutpoint.

References

ALTMAN, D. G. – LAUSEN, B. – SAUERBREI, W. – SCHUMACHER, M. [1994]: Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors. *Journal of the National Cancer Institute*. Vol. 86. No. 11. pp. 829–835.

BETENSKY, R. A. – RABINOWITZ, D. [1999]: Maximally Selected Chi-Square Statistics for kx2 Tables. *Biometrics*. Vol. 55. No. 1. pp. 317–320.

BLAND, J. M. – ALTMAN, D. G. [2004]: The Logrank Test. *British Medical Journal*. Vol. 328. No. 7447. p. 1073.

BORBÉLY, A. – VARGHA, A. [2010]: Az *l* variabilitása öt foglalkozási csoportban – Kutatások a Budapesti Szociolingvisztikai Interjú beszélt nyelvi korpuszban. *Magyar Nyelv*. Vol. 106. No. 4. pp. 455–470.

BOULESTEIX, A. L. [2006]: Maximally Selected Chi-Square Statistics for Ordinal Variables. *Biometrical Journal*. Vol. 48. No. 3. pp. 451–462.

COHEN, J. [1983]: The Cost of Dichotomization. *Applied Psychological Measurement*. Vol. 7. No. 3. pp. 249–253.

DECOSTER, J. – ISELIN, A-M. R. – GALLUCCI, M. [2009]: A Conceptual and Empirical Examination of Justifications for Dichotomization. *Psychological Methods*. Vol. 14. No. 4. pp. 349–366.

DELONG, E. R. – DELONG, D. M. – CLARKE-PEARSON, D. L. [1988]: Comparing the Area Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*. Vol. 44. No. 9. pp. 837–845.

FARRINGTON, D. P. – LOEBER, R. [2000]: Some Benefits of Dichotomization in Psychiatric and Criminological Research. *Criminal Behaviour and Mental Health*. Vol. 10. No. 2. pp. 100–122.

FITZSIMONS, G. J. [2008]: Death to Dichotomizing. *Journal of Consumer Research*. Vol. 35. No. 1. pp. 5–8.

HEINZL, H. [2000]: Dangers of Using "Optimal" Cutpoints in the Evaluation of Cyclical Prognostic Factors. In: *Ferligoj*, *A*. *– Mrvar, A*. (eds): *New Approaches in Applied Statistics*. Metodološki zvezki, 16. FDV. Ljubljana.

HEINZL, H, – TEMPFER, C. [2001]: A Cautionary Note on Segmenting a Cyclical Covariate by Minimum P-Value Search. *Computational Statistics & Data Analysis*. Vol. 35. Issue 4. pp. 451–461.

KOZIOL, J. A. [1991]: On Maximally Selected Chi-Square Statistics. *Biometrics*. Vol. 47. No. 4. pp. 1557–1561.

LAUSEN, B. – SCHUMACHER, M. [1992]: Maximally Selected Rank Statistics. *Biometrics*. Vol. 48. No. 3. pp. 73–85.

LAUSEN, B. – SCHUMACHER, M. [1996]: Evaluating the Effect of Optimized Cutoff Values in the Assessment of Prognostic Factors. *Computational Statistics & Data Analysis*. Vol. 21. Issue 3. pp. 307–326.

MANTEL, N. [1966]: Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration. *Cancer Chemotherapy Reports*. Vol. 50. No. 3. pp. 163–170.

MAXWELL, S. E. – DELANEY, H. D. [1993]: Bivariate Median Splits and Spurious Statistical Significance. *Psychological Bulletin*. Vol. 113. No. 1. pp. 181–190.

MAXWELL, S. E. – DELANEY, H. D. [2004]: *Designing Experiments and Analyzing Data: A Model Comparison Perspective*. Lawrence Erlbaum Associates. Mahwah.

MICCERI, T. [1989]: The Unicorn, the Normal Curve, and Other Improbable Creatures. *Psychological Bulletin*. Vol. 105. No. 1. pp. 156–166.

MILLER, R. – SIEGMUND, D. [1982]: Maximally Selected Chi Square Statistics. *Biometrics*. Vol. 38. No. 4. pp. 1011–1016.

MÓZES, T. – VARGHA, A. [2007]: A születési sorrend és a személyiség összefüggései. In: *Bagdy*, *E.* – *Mirnics*, *Zs*. – *Vargha*, *A*. (eds.): *Egyén–Pár–Család*. *Tanulmányok a pszichodiagnosztikai tesztadaptációs és tesztfejlesztő kutatások köréből*. pp. 249–270. Animula. Budapest.

OLÁH, A. [1985]: A Californiai Pszichológiai Kérdőív hazai alkalmazásával kapcsolatos tapasztalatok. In: *Hunyady*, *G*. (ed.): *Pszichológiai Tanulmányok*. *XVI*. pp. 53–101.

OVERALL, J. E. [1968]: Standard Psychiatric Symptom Description: The Factor Construct Rating Scale (FCRS). *Triangle*: *Sandoz Journal of Medical Sciences*. Vol. 8. No. 5. pp. 178–186.

PARKER, G. [1989]: The Parental Bonding Instrument: Psychometric Properties Reviewed. *Psychiatric Developments*. Vol. 7. No. 4. pp. 317–335.

PARKER, G. [1990]: The Parental Bonding Instrument: A Decade of Research. *Social Psychiatry and Psychiatric Epidemiology*. Vol. 25. No. 6. pp. 281–282.

PETHŐ, B. [2001]: *Klassifikation, Verlauf und Residuale Dimension der Endogenen Psychosen*. Platon Verlag Budapest. Universitätsverlag. Ulm.

ROCKLAND, I. H. – POLLIN, W. [1965]: Quantification of Psychiatric Mental Status. *Archives of General Psychiatry*. Vol. 12. No. 1. pp. 23–28.

SCHOENFELD, D. [1981]: The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions. *Biometrika*. Vol. 68. No. 1. pp. 316–319.

VARGHA, A. [1989]: *A Magyar Rorschach Standard táblázatai*. Schoolbook Publisher. Budapest.

VARGHA, A. – DELANEY, H. D. [1998]: The Kruskal-Wallis Test and Stochastic Homogeneity. *Journal of Educational and Behavioral Statistics*. Vol. 23. No. 2. pp. 170–192.

VARGHA, A. – DELANEY, H. D. [2000]: A Critique and Improvement of the CL Common Language Effect Size Statistic of McGraw and Wong. *Journal of Educational and Behavioral Statistics*. Vol. 25. No. 2. pp. 101–132.

WILLIAMS, B. – MANDREKAR, J. N. – MANDREKAR, S. J. – CHA, S. S. – FURTH, A. F. [2006]: *Finding Optimal Cutpoints for Continuous Covariates with Binary and Time-to-Event Outcomes*. Technical Report Series No. 79. Department of Health Sciences Research, Mayo Clinic. Rochester.

Debt Dynamics and Sustainability*

Csaba G. Tóth
Research Fellow
Századvég Economic Research Ltd.,
PhD Student
University of Sciences of Debrecen
E-mail: toth@szazadveg-eco.hu

Hungary's government debt and its sustainability is a current issue from many different aspects. Applying debt-dynamic analyses, in the first part of this paper, the author examines what periods the last twelve years can be split into with regard to economic policy. His most important conclusion is that in spite of the similar rate of increase in the gross debt of the general government as a proportion of GDP in the periods from 2002 to 2006 and from 2007 to 2010, the reasons are markedly different. After that, from the several tests of debt sustainability, he makes first an analysis based on fiscal reaction function, at the aid of which tries to quantify the correction mechanisms of the Hungarian economic policy in the last two decades and to compare them with international examples. There are several reasons for the analysis of the difference between the real interest and the real growth rate. On the one hand, it shows that Hungary's government debt left the sustainable path in 2001/2002. On the other hand, according to its estimations on the primary gap, with the difference between the real interest and the real growth rate unchanged, the general government deficit of around 4 percent of GDP, typical in the past few years, may be enough to stabilise the debt ratio. However, in order to reduce the gross government debt to GDP ratio below 60 percent within the next 5–10 years, the balance has to be improved by 1 or 2 percentage points.

KEYWORD:
Government debt.
Sustainability.

* The study was made as part of the scientific researches assisting the preparation of the National Strategy for Sustainable Development on behalf of the National Council on Sustainable Development.

Hereby, the author expresses his thanks to *András Balatoni, László Hunyadi, László Muraközy, Endre Szolnoki, András Viszkievicz* and the anonymous reader for their useful proposals. Naturally, solely the author is responsible for errors or mistakes.

The article is the translated version of *Tóth* [2011] published in Hungarian.

The world economic crisis and especially the related financing crisis proved the extraordinary problems caused by indebtedness if there is no more trust and sources on the money market dry up. Although the reasons are rather various,[1] it is still true that a lot of countries – that no one had supposed before the crisis to do so – announced that they were unable to finance their economy from the market. Hungary became one of the first members of this group, which in itself is a reason for the scrutiny of the issue. However, this is not the only such factor. Among our competitors in the region, the gross government debt ratio to GDP (hereafter: debt ratio) is the highest in Hungary, and this is also true for the twenty years since the change of regime. Through interest payments this means a heavy burden on the general government, pressure for the withdrawal of sources with regard to the economy, and enhances the vulnerability of the country, which may lead to a financing crisis – as observed in autumn 2008.

The papers devoted to the issue and published to date can be classified into two groups. A part of them endeavoured to explore the specific features of budget policy. *Karsai* [2006] was among the first to draw attention to the parallel movement of budget and election cycles, which has since become an axiom in professional and public discussions. *Ohnsorge-Szabó* and *Romhányi* [2007] tried to point out the role played by the different items of expenditure in the budget expansion between 2000 and 2006. *Orbán* and *Szapáry* [2006] outlined frightfully accurately half a decade ago the dangers threatening the Hungarian economy from the side of fiscal policy. The different approaches contributed significantly to the exploration of the problem. *Győrffy* [2005] examined deficit budgeting from the point of view of the institutional system, while *Muraközy* [2008] drew attention to the effects of heritage from the past.

As for this study, however, the other group of the published papers – which may be narrower than the previous one – is more important and it examined budget policy through government debt and (also) its sustainability[2]. Out of them the work of *Czeti–Hoffmann* [2006] is worth to be highlighted, since they were among the first who endeavoured to apply debt dynamics tools to explore what factors contributed to the change of Hungary's gross government debt between 1995 and 2005 and to what extent. The article published by *Pápa* and *Valentinyi* [2008] is notable from different aspects. On the one hand, it very concisely summarizes the results of former researches, on the other

[1] See for example the writing of *Obstfeld–Rogoff* [2009] or *Stein* [2011].

[2] The expression „sustainability of government debt", widely spread in the literature, actually refers to the sustainability of indebtedness. The related researches (also) examine whether or not a certain level of indebtedness can be considered sustainable based on different aspects.

hand, in addition to drawing attention to the difficulties of analyses of sustainability, it suggests by means of some relatively simple methods also applicable in practice that there may be serious problems concerning the sustainability of Hungary's government debt. The article of *Ábel* and *Kóbor* [2011] was already made after the crisis, and they examined first of all the role of uncertainty in the change of government debt and in the light of this, the aspects worth to be kept in mind if economic policy decision-makers wish to set an upper limit for government debt as a proportion of GDP.

A common feature of the formerly mentioned three works is that each of them includes debt dynamics analysis, however, in the latter two papers the authors made only a brief account – in a few lines – of the methodology and the results. After presenting in short the theories of government debt, we endeavour to make up this deficiency in the second part of our study. By disaggregating the change of government debt, we look for an answer to what uniform sections the past twelve years can be split into as regards economic policy. In the next part the most widespread methods for studying the sustainability of government debt are considered, starting with the analysis based on reaction function, passing over to the different examinations relying on the difference between the real interest and the real growth rate. Each of these aims on the one hand at evaluating past processes regarding sustainability and on the other hand at trying to draw conclusions relevant for the future too.

1. Sustainable government debt

A very up-to-date dimension[3] of the researches on government debt examines sustainability, and therefore is interwoven with the issue of fiscal sustainability. However, before the accurate definition of this latter term it should be underlined that the sustainability of budget is determined by future budget policy, thus, sustainability in the narrower sense of the word is not measurable (*Pápa–Valentinyi* [2008]).

The many different definitions of fiscal sustainability are built on the concept of solvency. This is most often referred to by economists as the capacity of government to be always able to meet current obligations of repayment, without a request for rescheduling or any other similar external aid (*Burnside* [2005]). On this basis there is a relative professional consensus on the definition according to which a budget policy is sustainable if it does not threaten the solvency of the country in the future either (*Croce–Juan-Ramon* [2003]). However, a more detailed description is given by *Agnello* and *Sousa* [2009], who underline in addition that a budget deficit, which very often accom-

[3] See *Török* [2011].

panies unsustainable government debt, threatens the welfare state, since it firstly hinders the efficient allocation of resources, secondly affects sensitively the next generation through growing government debt, thirdly increases inflation and its volatility. The approach made by *De Castro* and *De Cos* [2002] is also related to the possible threats, and they drew attention to unsustainable fiscal policy sooner or later leading to the rise of interest rates, which in turn hinders economic growth (see *Reinhart–Rogoff* [2010] and *Presbitero* [2010]). The presented definitions and descriptions are summarised probably in the best way by *Buiter* [2004], who classifies the consequences of unsustainable fiscal policy into three groups: *1*. the state can spend less money and has to collect more taxes than planned earlier on, *2*. the threat of inflation, and *3*. the threat of sovereign default grow. In connection with the sustainability of debt, the concept of budget limit also occurs frequently (see *Buiter* [1985] or *Blanchard* [1990]), according to which the present value of revenues to be realized in the future should be equal to the present value of the public debt. It is important to see, however, that in itself it is not a condition of sustainability, since it is met in case of a later adjustment too. Nevertheless, sustainability prevails exactly if present processes do not lead to insolvency even without intervention. Namely, if a budget policy is not sustainable, then the question is not whether it will be broken but in what way. The state will either implement the correction on its own or the market will do that for it.

2. Debt dynamics analysis

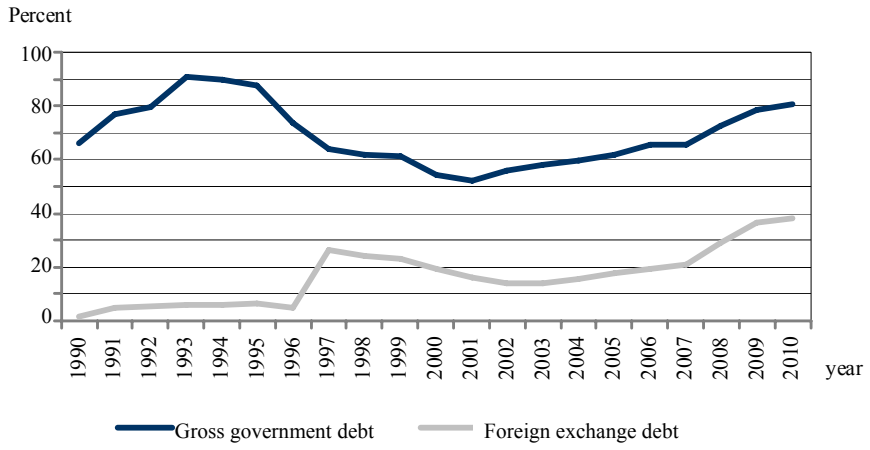
To judge the sustainability of an economic process and to select the most sustainable path of the possible scenarios, it is indispensable to explore and scrutinise connections. In the case of government debt[4], this (also) means that the examination of the past period must be an integral part of the analysis. If one can quantify and through this understand what factors and to what extent influenced government debt, and what are the most marked characteristics of these processes, then one can draw relevant conclusions. Accordingly, we first give a general outline of government debt, the main trends and characteristics thereof. Following this, after the selection and summary of the appropriate methodology, we examine processes concerning debt dynamics, that is, disaggregate the contribution of the different factors to the change of government debt.

At the time of the change of regime, at the end of 1990, Hungary's government debt relative to GDP was 66.2 percent. According to official data, the proportion of liabilities accounted for in foreign exchanges was negligible within this, but their role

[4] In the following, consolidated gross government debt accounted for according to the methodology of ESA is meant by government debt.

was actually significant already at the time. Namely, at the beginning of the nineties the costs of indebtedness in foreign exchanges along with unlimited financing by the central bank were not recorded directly in the budget, "only" worsened the profit of the National Bank of Hungary (NBH), therefore the comparability of data is ensured only for the period after the debt transformation implemented at the end of 1996 (*Barabás–Hamecz–Neményi* [1998]). As for liabilities, the first section is clearly the period between 1990 and 1995: the debt ratio then rose by over 20 percentage points. After the budget adjustment announced in 1995 the debt ratio decreased continuously for six years, and reached a low in 2001 at 52.2 percent. In parallel, the proportion of foreign exchange debt recorded in budget accounts soared, which is due to the separation of direct public financing from the central bank in line with EU requirements, in other words, government debt was changed.

Figure 1. Gross government debt as a percentage of GDP



Source: NBH.

Gross government debt as a proportion of GDP grew continuously from 2002, and reached 80.2 percent in the last year of the examined period.[5] Though the proportion of foreign exchange debt within this hardly changed in the first half of the decade, varying between 25 and 30 percent, it increased to almost two-fold in the last three years examined. The main reason for this is that as an impact of the financial crisis the Hungarian state was not able to satisfy the financing needs of general government from the market, thus it had to take up foreign exchange loans from the triad of the International Monetary Fund, the European Union and the World Bank.

[5] The period examined by the study lasts until 31 December 2010, so the use of wealth in private pension funds for debt repayment and any other event which has occurred since this date are not covered.

Since the change of regime, the structure of the debt stock has also been transformed. In the middle of the nineties the proportion of loans was over 65 percent, while their share gradually fell to 10 percent by 2007. Parallelly, the role of securities, typically government bonds, increased. Within the issued securities, it was solely the value of long-term government bonds that grew, which indicates that with market economy evolving – in crisis-free periods – the Hungarian state can have longer-term liabilities. This is also apparent from the combined proportion of short-term loans and securities not exceeding 20 percent of total government debt in any of the last twenty years. The role of loans increased again in the last three years, and as an effect of borrowings from international financial organisations, their proportion within total debt stock reached again 28 percent by the end of the examined period.[6]

The point of debt dynamics analyses is to decompose the effect of the different factors on the change of government debt according to the methodology chosen in line with the aim of the examination. Hereinafter we will follow the methodology also applied by *Ra–Rhee* [2005]. For this, one first has to state the general formula of nominal debt, separating liabilities in HUF and in foreign exchanges.

$$D_t = -PB_t + \left(1 + id_{t-1}\right) ID_{t-1} + \left(1 + ix_{t-1}\right)\left(1 + \varepsilon_t\right) XD_{t-1} + OD_t, \qquad /1/$$

where

D_t – gross government debt in HUF at the end of period t;

PB_t – primary balance of general government in period t, not including interest payments;

ID_{t-1} – government debt in HUF at the end of period $t-1$;

XD_{t-1} – government debt in foreign exchanges at the end of period $t-1$;

id_{t-1} – interest rate on government debt in HUF in period $t-1$;

ix_{t-1} – interest rate on government debt in foreign exchanges in period $t-1$;

ε_t – nominal depreciation at the end of period t;

OD_t – other items (such as privatisation) in period t.

Let us suppose that a_{t-1} is the proportion of government debt recorded in foreign exchanges within total government debt in period $t-1$. In this case one can state for the interest rate that

[6] For more on Hungary's government debt see the works of *Antal* [2006]; *Muraközy* [2004]; *Mellár* [1997], [2002]; *Kun* [1996], and *Czike* [2010].

$$1+i_{t-1}=\left(1+id_{t-1}\right)\left(1-a_{t-1}\right)+\left(1+ix_{t-1}\right)a_{t-1}\,, \qquad /2/$$

where i_{t-1} – interest rate on government debt in period $t-1$;

Accordingly, government debt may also be given in a simplified form, and it is worth to further transform it:

$$D_t=-PB_t+\left(1+id_{t-1}\right)\left(1-a_{t-1}\right)D_{t-1}+\left(1+ix_{t-1}\right)\left(1+\varepsilon_t\right)a_{t-1}D_{t-1}+OD_t\,, \qquad /3/$$

$$D_t=-PB_t+\left[\left(1+id_{t-1}\right)\left(1-a_{t-1}\right)+\left(1+ix_{t-1}\right)\left(1+\varepsilon_t\right)a_{t-1}\right]D_{t-1}+OD_t\,, \qquad /4/$$

$$D_t=-PB_t+\left[\left(1+id_{t-1}\right)\left(1-a_{t-1}\right)+\left(1+ix_{t-1}\right)a_{t-1}+\left(1+ix_{t-1}\right)\varepsilon_t a_{t-1}\right]D_{t-1}+OD_t\,, \qquad /5/$$

$$D_t=-PB_t+\left[\left(1+i_{t-1}\right)+\left(1+ix_{t-1}\right)\varepsilon_t a_{t-1}\right]D_{t-1}+OD_t. \qquad /6/$$

After this let us divide the equation with nominal GDP in period t (Y_t), while marking debt relative to GDP, the primary balance and other items by *d*, *pb* and *od* respectively.

$$d_t=-pb_t+\left[\left(1+i_{t-1}\right)+\left(1+ix_{t-1}\right)\varepsilon_t a_{t-1}\right]\frac{D_{t-1}}{Y_{t-1}}\frac{Y_{t-1}}{Y_t}+od_t\,, \qquad /7/$$

$$d_t=-pb_t+\left[\left(1+i_{t-1}\right)+\left(1+ix_{t-1}\right)\varepsilon_t a_{t-1}\right]d_{t-1}\frac{Y_{t-1}}{Y_t}+od_t\,. \qquad /8/$$

Let us replace the $\frac{Y_{t-1}}{Y_t}$ growth by a new formula: $\frac{1}{\left(1+g_t\right)\left(1+\pi_t\right)}$, where *g* is real growth rate, and π is inflation. After this, debt can be broken down in the following manner:

$$d_t=-pb_t+\frac{\left(1+i_{t-1}\right)+\left(1+ix_{t-1}\right)\varepsilon_t a_{t-1}}{\left(1+g_t\right)\left(1+\pi_t\right)}d_{t-1}+od_t. \qquad /9/$$

To quantify the change let us deduct debt in period *t*–1 from debt in period *t*:

$$d_t-d_{t-1}=-pb_t+\left\{\frac{\left(1+i_{t-1}\right)+\left(1+ix_{t-1}\right)\varepsilon_t a_{t-1}}{\left(1+g_t\right)\left(1+\pi_t\right)}-1\right\}d_{t-1}+od_t= \qquad /10/$$

$$= -pb_t + \left\{ \frac{(1+i_{t-1}) + (1+ix_{t-1})\varepsilon_t a_{t-1} - (1+g_t)(1+\pi_t)}{(1+g_t)(1+\pi_t)} \right\} d_{t-1} + od_t = \qquad /11/$$

$$= -pb_t + \left\{ \frac{i_{t-1} - \pi_t(1+g_t) + (-g_t) + (1+ix_{t-1})\varepsilon_t a_{t-1}}{(1+g_t)(1+\pi_t)} \right\} d_{t-1} + od_t = \qquad /12/$$

$$= -pb_t + \left[\frac{i_{t-1}}{(1+g_t)(1+\pi_t)} + \frac{-\pi_t(1+g_t)}{(1+g_t)(1+\pi_t)} + \right.$$
$$\left. + \frac{-g_t}{(1+g_t)(1+\pi_t)} + \frac{(1+ix_{t-1})\varepsilon_t a_{t-1}}{(1+g_t)(1+\pi_t)} \right] d_{t-1} + od_t \ . \qquad /13/$$

With the aid of the equation the change of debt may be accurately disaggregated. The impact of the different factors may be quantified according to the following equations:

$-pb_t$ – primary balance;

$\left(\frac{-g_t}{(1+g_t)(1+\pi_z)} \right) d_{t-1}$ – real growth;

$\left(\frac{-\pi_t}{1+\pi_t} \right) d_{t-1}$ – inflation;

$\left(\frac{i_{t-1}}{(1+g_t)(1+\pi_t)} \right) d_{t-1}$ – nominal interest;

$\left(\frac{(1+ix_{t-1})\varepsilon_t a_{t-1}}{(1+g_t)(1+\pi_t)} \right) d_{t-1}$ – change in exchange rate;

od_t – other items influencing debt.

In the following, the change of gross consolidated nominal (Maastricht) government debt in the period from 31 December 1998 to 31 December 2010 is analysed on the basis of the presented methodology. The development of the different items is shown in the Appendix. The choice of the starting date of the examined period is partly explained by the availability of data and the fact that the exchange of debt between the NBH and the general government was completed at the end of 1996. This means that the use of budget data on earlier years would lead to serious distortions, while the correction of official data is hindered by methodological limitations. It is important to draw attention to

another aspect as well to the procedure we applied. As we cannot adjust general government statistics by the balance of the NBH,[7] the economic management of the central bank is only revealed in the disaggregation through the primary gap, in the year of accounting. This has the most important role in quantifying the impact of the change in the exchange rate: in case the exchange rate weakens, the value of the government debt accounted in foreign exchanges increases, in parallel, however, the value of the foreign exchange reserves of the central bank rises. Equation /13/ quantifies solely the former impact, the growth of the foreign exchange reserves is revealed in the primary balance through the profit of the central bank.

During the period examined, between 1999 and 2010, the government debt of the country rose by 18.5 percentage points relative to GDP. When looking at the twelve years as a whole, 80 percent of the increment turns out to be due to fiscal policy, the primary balance of the general government. The effect of the real interest rate could be almost fully offset by economic growth despite the crisis, so the difference between the real interest and the real growth rate increased the debt by 3 percentage points over twelve years. On the whole, the impact of the change in the exchange rate and of other items cannot be referred to as substantial either, especially in the light of the fact that they almost wholly counterbalance each other. However, the aggregated examination of the examined period disguises the most important connections. Namely, if you have a closer look, three periods with very different characteristics can be seen based on the data after disaggregating the different factors influencing the debt.

Table 1

Effect of different items on government debt as a percent of GDP
(percent)

Items	Periods			
	1999–2001	2002–2006	2007–2010	1999–2010
Starting debt	62.0	52.2	65.6	62.0
Closing debt	52.2	65.6	80.2	80.2
Change in debt	–9.8	13.3	14.7	18.3
Primary balance	–4.1	19.3	–0.1	15.1
Nominal interest	18.0	22.0	18.3	58.3
Inflation	–16.2	–14.0	–11.4	–41.6
Real interest	1.8	8.1	6.9	16.7
Economic growth	–6.6	–10.2	3.1	–13.7
Change in exchange rate	–0.7	0.2	2.7	2.2
Other items	–0.2	–4.0	2.2	–2.0

[7] As was done by *Czeti–Hoffmann* [2006].

1. Between 1999 and 2001 the government debt was reduced substantially as a proportion of GDP, from 62.0 to 52.2 percent. Nearly the half of this resulted from the improvement of the primary balance, while the other half from the impact of the difference between the real interest and the real growth rate. Namely, though the interest level was high, the impact of the real interest rate – because of the almost similarly high inflation – did not even reach 2 percentage points, while the economy was increasing dynamically all through the period, and the effect of the change in the exchange rate and of other items was not significant.

2. The debt rate grew at a robust rate, by 13.4 percentage points between 2002 and 2006, and reached 65.6 percent by the end of the period. The expansive general government had such a serious role in this that the primary budget balance in itself increased the indebtedness of the country by 19.3 percentage points. Compared to the earlier period, the difference between the real interest and the real growth rate decreased the debt ratio only to a much lower extent, by 2.1 percentage points. As the rate of economic growth remained high and balanced, this was caused exclusively by the fact that not independently from the hectically changing inflation, the level of the nominal interest went down only slowly. However, the aggregate impact of other items changed favourably, which was primarily owing to the one-time marked rise[8] of revenues from privatisation, while the impact of the change in the exchange rate was not considerable.

3. In the years between 2007 and 2010 the debt ratio grew at a similar rate to that in the previous period, which reached 80.2 percent relative to GDP by the end of 2010. However, the reasons for indebtedness are markedly different. The current balance of the general government did not contribute at all in itself to the growth of debt. Instead, the changed sign of the difference between the real interest and the real growth rate increased the debt ratio by 10 percentage points, which stemmed first of all from the dramatic decline of growth, with a special regard to the considerable recession in 2009. The weakening of the exchange rate, consistent with the financial crisis, also played a part in indebtedness. In addition, however, there exists another important factor. Although other items increased the debt ratio by "only" 2.2 percentage points in total, this item was 4.9 percentage points in 2009, and primarily resulted from the fact that the state placed the unused part of borrowings from international financial organisations on foreign exchange ac-

[8] The revenue from the sales of the Budapest Airport approximated 2 percent as a proportion of GDP.

counts kept in the National Bank of Hungary, thus raising the official foreign exchange reserves of the country.

3. Sustainability analyses

Public finance within that the sustainability of government debt has been a popular research issue for decades. Its main reason is that this aspect of evaluation has always been an important dimension of judging the different trends or action programmes in economic policy. The popularity of the issue is justified by two additional important factors. On the one hand, in parallel with globalisation and economic integration, access to capital means less and less a real limitation for the general government. On the other hand, excessive indebtedness is a long-term harmful process that can be remedied at the lower cost the earlier the intervention is made. Therefore, it is a vital interest of decision-makers too to be aware of the long-term impacts of the economic policy they pursue, and to see the points where processes are heading.

Because of all this – and especially since the beginning of the financial crisis – there is a high interest in sustainability analyses concerning government debt. However, it is important to stress that since the change of government debt depends on several different factors whose relation is very various even with one another, in the majority of cases it is impossible to judge absolutely surely which process is sustainable and which is not. Therefore, sustainability analyses, too, usually follow a structure where past processes are examined and conclusions on sustainability are made based on one or two selected aspects.

In the following, the sustainability of Hungary's government debt is analysed on the basis of two very widespread methods which make part of most such analyses (*Callen et al.* [2003]). One of them is the "response function analysis" measuring sustainability through the flexibility of fiscal policy, while the other extrapolates ex-post processes by examining the difference between the real interest and the real growth rate, at the same time as "freezing" (external) factors outside the direct remit of fiscal policy.

3.1. Fiscal reaction function analysis

One of the most widespread types of sustainability tests was first used by *Bohn* [1998], who analysed US budget data with the method of reaction function analysis. The essence of the procedure is to investigate the connection between two (or more) variables. One has to be a fiscal instrument which indicates the changes in economic

policy, while the other has to reflect fiscal goals. In case of researches testing government debt sustainability, to maintain the stability of government debt is an obvious goal, while the other (fiscal) variable in the relation is primary balance. Namely, a number of people already studied the impact of fiscal policy on government debt, concerning either debt dynamics analyses (*Hall–Sargent* [2010] and *Bognetti–Ragazzi* [2009)] or the effect of budget policy on interests (*Ardagna–Caselli–Lane* [2004], *Baldacci–Kumar* [2010]), or its role in successful debt reduction (*Reinhart–Rogoff–Savastano* [2003], *Baldacci–Gupta–Mulas-Granados* [2010], *Nickel–Rother–Zimmermann* [2010]). Bohn, however, drew attention to the fact that it is not only the primary balance that may influence government debt (as was presented in the previous chapter), the effect may also be mutual, and is very much consistent with sustainability. Namely, if a government reacts quickly and efficiently to the change of government debt through the primary balance, then it practically averts the danger of government debt becoming unsustainable. Accordingly, in case of fiscal reaction function analyses, government debt (and the fiscal policy behind) is considered as sustainable if past evidence proves that the position of the budget improves in response to the increase of government debt, and prevents indebtedness; while unsustainability occurs if budget policy is inflexible to the development of debt ratio.

Let us examine on this basis Hungary's figures for the last twenty years. It has already turned out from the foregoing as well that in the first half of the nineties the primary deficit was relatively large along with a high level of debt (*P. Kiss* [1998]).

Following a budget adjustment, the balance improved substantially, and parallelly the debt level fell until 2002, then the deficit was very considerable again during four years and the indebtedness of the country grew. In the last three years of the examined period, though the primary balance was in surplus again, the growth rate of the debt ratio did not decline. After this let us state the equation of regression estimation:[9]

$$pb_t = \beta_0 + \beta_1 d_{t-1} + \beta_2 pb_{t-1} + \varepsilon \,. \qquad /14/$$

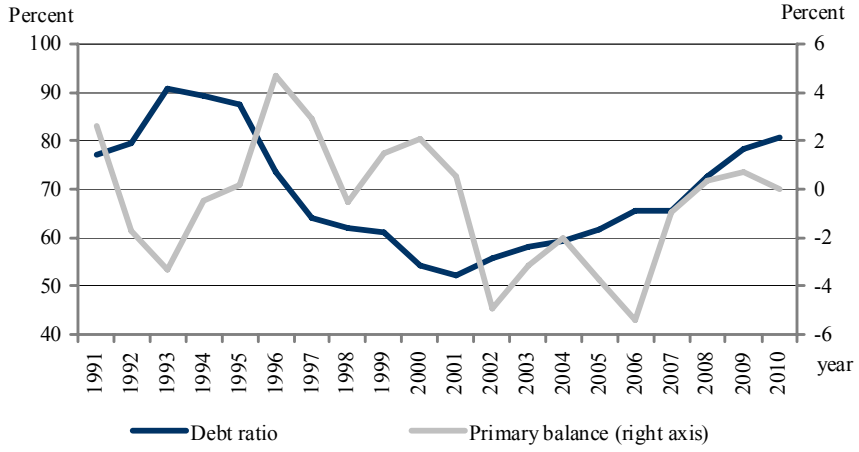
Primary balance (pb_t) is the dependent variable, and explanatory variables include government debt measured at the end of the previous period (d_{t-1}) and the primary balance of the previous period (pb_{t-1}). In the regression calculation both the debt and the primary balance were measured relative to GDP.[10]

$$pb_t = \underset{(-2.322)}{-5.993} + \underset{(2.086)}{0.0804}\, d_{t-1} + \underset{(4.107)}{0.4353}\, pb_{t-1} \,. \qquad /15/$$

[9] Details of the estimation are available in the Appendix.
[10] *t* statistics in parentheses.

This equation reveals that in short term, in the first year, a 10 percentage-point change of government debt as a proportion of GDP changes by 0.8 percentage point the primary balance relative to GDP, with all other factors unaltered. As explanatory variables in equation /15/ include pb_{t-1}, short- and long-term effects are different. The estimated coefficient of pb_{t-1} is a kind of smoothness parameter, in other words, it shows the speed of primary balance in adapting to the change of government debt. Roughly 56 percent of total adaptation occurs in the first year. Total adaptation is easy to calculate from this: $0.0804/0.5647 = 0.1424$. That is, if government debt relative to GDP is up (down) by 10 percentage points over a year, then primary balance as a proportion of GDP improves (deteriorates) by 1.4 percentage points in the long run. (Detailed results are available in the table in the Appendix.)

Figure 2. Government debt and balance as a percentage of GDP



Source: NBH, Ministry for National Economy, *IMF* [2007].

A significant part of fiscal reaction function analyses attempts to eliminate in the next step the effect of economic cycles from that of debt on the balance (see *Izak* [2009] or *Greiner–Koeller–Semmler* [2004]). The essence of this is that budget management is separated from the direct and automatic negative or positive impact of the cyclical changes in the economy on the balance, which gives a more accurate picture on the behaviour of fiscal policy. We can do this in the first step by extending the set of explanatory variables with the indicator of output gap. To achieve this, at the aid of the output gap (og_t) estimated by HP filter, equation /14/ is adjusted in the following:

$$pb_t = \beta_0 + \beta_1 d_{t-1} + \beta_2 pb_{t-1} + \beta_3 og_t + \varepsilon. \qquad /16/$$

The results indicate that by eliminating the impact of economic cycles from the model, in the case of the debt, both significance and the value of the coefficient strengthen, though this latter is rather low even so. As for the impact of the lagged level of primary balance (pb_{t-1}), its significance increased compared to the previous estimation, and the value of the coefficient rose as well. Although the effect of economic cycle was no significant, all in all the explanatory power of the model, that is, adjusted R^2 slightly grew. (See the table in the Appendix.)

$$pb_t = \underset{(-2.891)}{-10.2416} + \underset{(2.763)}{0.1486}\, d_{t-1} + \underset{(4.537)}{0.6035}\, pb_{t-1} + \underset{(1.232)}{0.2670}\, og_t \ . \qquad /17/$$

Referring to the former example, as 40 percent of total adaptation occurs in the first year, primary balance relative to GDP improves (deteriorates) by 1.4 percentage points in the first year and by 3.7 percentage points in the long term when government debt as a proportion of GDP increases (falls) by 10 percentage points.

Concerning the estimation method, however, a difficult-to-handle problem of endogeneity is caused by the fact that the relation between output gap and primary balance is supposed not to be "unidirectional", which distorts results. The solution may be to eliminate the impact of cycles not by expanding the scope of explanatory variables with output gap but by changing dependent variable by introducing cyclically adjusted primary balance.

Namely, the effect of cycles may be eliminated in the easiest way by explaining cyclically adjusted primary balance instead of the primary balance used formerly. Though there are more and more discussions on the problems of its use (see *Lewis* [2010] or *Darvas–Kostyleva* [2011]), the aim of the indicator is exactly to give a picture of budget balance – which is independent from the change of economic cycles – or its expected change with the potential decrease of output gap. However, investigation is restricted by the fact that time series on cyclically adjusted primary balance are available only from 1996. The new equation can be stated in the following way:

$$pb\_cic_t = \beta_0 + \beta_1 d_{t-1} + \beta_2\, pb\_cic_{t-1} + \varepsilon, \qquad /18/$$

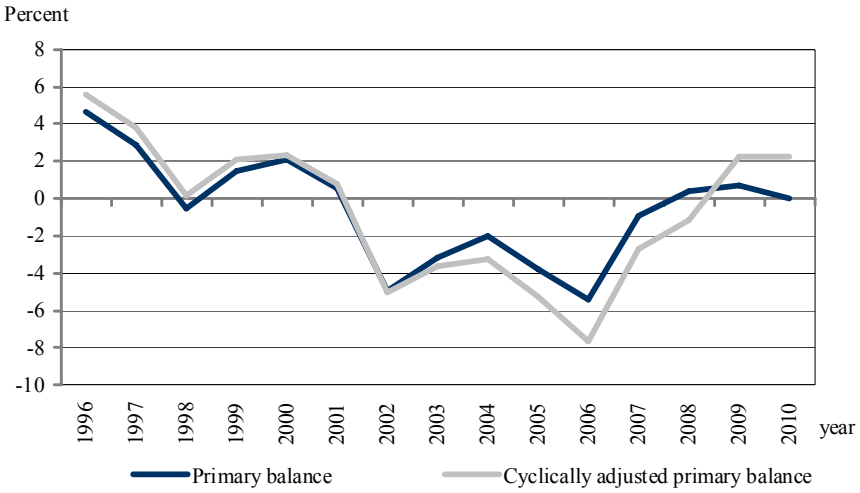
where cyclically adjusted primary balance (pb_cyc_t) is the dependent variable, and the value of cyclically adjusted primary balance in period t–1 (pb_cyc_{t-1}) is the new explanatory one.

$$pb\_cic_t = \underset{(-3.664)}{-13.0973} + \underset{(3.777)}{0.1975}\, d_{t-1} + \underset{(3.896)}{0.5432}\, pb\_cic_{t-1} \ . \qquad /19/$$

The result achieved this way refers to the fact that by eliminating economic cycles, the explanatory power of the model can be enhanced (see Appendix). Concern-

ing the essence of the fiscal reaction function analysis, the significance of government debt increased and regression coefficient rose too. As 46 percent of total adaptation occurs in the first year, cyclically adjusted primary balance relative to GDP improves (deteriorates) by 2 percentage points in the first year and by 4.3 percentage points in the long term when government debt as a proportion of GDP increases (falls) by 10 percentage points.

Figure 3. Deficit indicators of general government as a percentage of GDP



Source: The annual macro-economic database (AMECO), Ministry for National Economy.

The results achieved are in harmony with the findings of former studies. Earlier on *Izak* [2009] made the fiscal reaction function analysis of all the 10 central and eastern European countries that joined the European Union in 2004 or 2007. In connection with Hungary's figures they also judged the effect of government debt to be significant, while the regression coefficient could be somewhat higher (–0,2359) than the results we achieved since the examined periods do not exactly correspond to each other.[11,12] Although in their later investigation *Câmpeanu* and *Stoian* [2010] did not find the explanatory power of the first difference of Hungary's government debt to be significant in the change of primary balance, it seems to be justified in the light of the fact that they limited their test on the period be-

[11] The second term in equations /15/, /17/, and /19/.

[12] An interesting finding of the previously mentioned study is that a significant relation between the balance and the debt can only be recorded in Hungary out of the examined countries. This may result from the fact that government debt is the highest in Hungary in the region, and until it reaches a critical level, it is unnecessary to respond to the growth of debt by fiscal austerity.

tween 2000 and 2008, in the largest part of which period government debt increased along with a high deficit.

3.2. Analysing difference between real interest and real growth rate

Another frequent type of sustainability analyses is built on analysing the difference between real interest and real growth (see for example *Callen et al*. [2003] and *Lewis* [2010]). These types of work are based on the relation that the simplest formulas describing the change of government debt include the starting value of government debt as well as real interest, growth, and primary balance.

$$\Delta d = \frac{r-g}{1+g} d_{t-1} - pb_t \ . \qquad /20/$$

Hereinafter the difference between real interest and real growth is defined as follows:

$$u = \frac{r-g}{1+g}. \qquad /21/$$

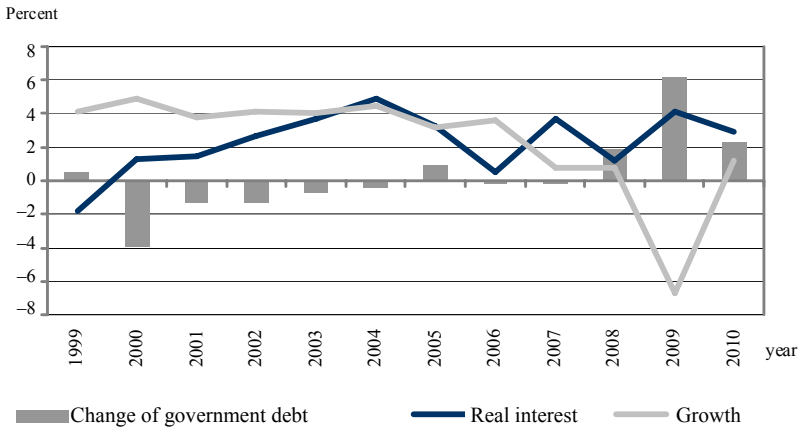
One of the most important features of the relation between the difference between real interest and real growth, primary balance and government debt is that fixing the former two leads to the equilibrium point of government debt. As detailed by *Mellár* [2002] too, four versions are distinguished depending on whether primary balance and the difference between real interest and real growth are positive or negative. If the latter is positive, that is, real interest exceeds real growth, the equilibrium point is negative in the case of a budget deficit and positive in the case of a budget surplus, but is not stable in either cases. If the difference between real interest and real growth is negative, the equilibrium point is stable in any case. In the case of a budget deficit, its value is higher, while in the case of a budget surplus, lower than zero.[13]

After presenting the way the value, and especially the sign, of the difference between real interest and real growth influences the development of government debt, the resulting relations are used to test sustainability. In the first step let us have a look on how the difference between real interest and real growth changed in Hun-

[13] Equation /20/ is a first-order difference equation. If the one-period lagged level of government debt relative to GDP is added to both sides, the coefficient of the lagged dependent variable will be the value of the difference between real interest and real growth increased by one. The stability characteristic of the fixed point depends on how the absolute value of this parameter is related to one. Accordingly, if the difference between real interest and real growth falls in the interval of (–2.0), then the fixed point of the dynamic system is stable; if it falls outside, then the fixed point is unstable. As in general no negative autocorrelation is recorded for government debt, only the sign of the difference between real interest and real growth is usually examined.

gary. It is important to underline that in our case real interest is quantified in the following way: $r_{t-1} = i_{t-1} - \pi_t$, which indicates that, as opposed to inflation and similarly to nominal interest, real interest "looks ahead" too, therefore (the change of) the debt in period *t* depends on the real interest in period $t-1$ (as well).

Figure 4. Structure of difference between real interest and real growth and its effect on government debt as a proportion of GDP



Note. The change of debt shows exclusively the impact of the difference between real interest and real growth.

Source: Own calculations based on HCSO data and AMECO.

At the beginning of the examined period the rate of economic growth exceeded substantially real interest. This means that the difference itself between real interest and real growth decreased government debt, which would not have grown under such circumstances even along with a relatively significant primary deficit.

At the beginning of the 2000s this trend faded away, and apart from a few exceptional years the difference between real interest and real growth reduced government debt to a small extent. This was, however, the case only until the beginning of the economic crisis: the recession suffered in the year 2009 raised the difference between real interest and real growth considerably, which alone increased government debt relative to GDP by almost 6 percentage points.

All in all, therefore, one can say about the last twelve years that the average value of the difference between real interest and real growth was close to zero (0.002), and its average impact on the increase of debt was extraordinarily small as well (0.32 percentage points as a proportion of GDP). The picture is certainly more favourable if the past period is examined without the last three years hit by the crisis. In this case it can be stated that the negative value (–0.013) of the difference between real interest and real

growth lowered government debt by 0.73 percentage point as a proportion of GDP per year on average. The results obtained are extremely sensitive to the way of calculating real interest, so preferably it is worth to highlight the general picture that the difference between real interest and real growth changed to a low extent but favourably in the years before the crisis, while it contributed significantly to indebtedness during the crisis.

After quantifying the difference between real interest and real growth, let us come back to the investigation of sustainability. In the first step it is worth relying on the work of *Blanchard* [1990], who introduced the concept of primary budget gap. The essence of the procedure is to assign the primary balance that would stabilise government debt to the past data of the difference between real interest and real growth, and to deduct that from the current balance. If the real balance is better than the calculated one, that is, the gap is positive, then government debt is sustainable, while vice versa further interventions are needed to ensure the sustainability of government debt.

Table 2

Calculation of primary budget gap

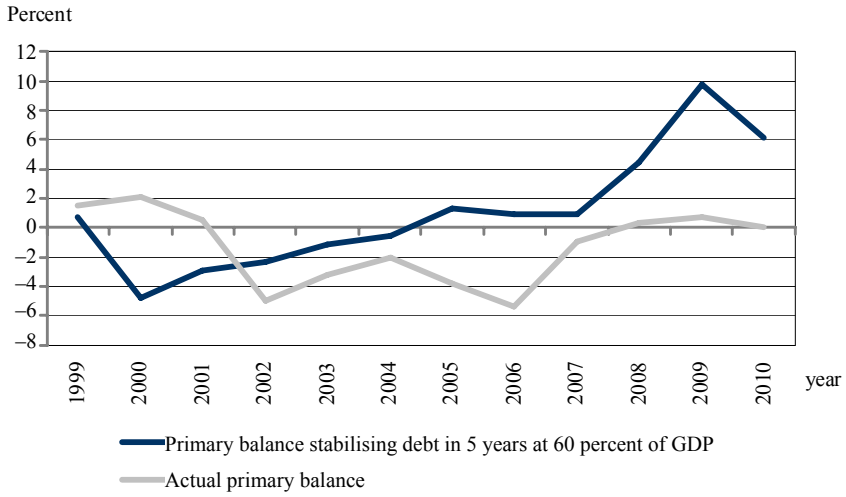
Averaged period	Starting government debt (as a percentage of GDP)	Difference between real interest and real growth	Calculated primary balance	Actual primary balance (2010)	Primary budget gap
			(as a percentage of GDP)		
1999–2010	80.2	0.002	0.2	0	–0.2
1999–2007	80.2	–0.013	–1.0	0	1.0

Source: Own calculation.

Based on Hungary's figures for the past twelve years, using the average difference between real interest and real growth, a primary surplus of 0.2 percent as a proportion of GDP is needed to stabilise the government debt of 80.2 percent measured at the end of 2010. This is more or less the same as the primary balance in 2010, which means that the budget deficit of around 4 percent recorded for the last few years is roughly enough to stabilise government debt. Although economic crises return from time to time – at varying intervals, it is worth performing the analysis also by leaving out of consideration the last three years, mostly hit by the crisis, when calculating the average difference between real interest and real growth. In this case the primary balance necessary to stabilise the present government debt is –1.0 percent, in other words, allows for a deficit, which means that if the total deficit is less than 5 percent, then government debt is already sustainable according to the approach of Blanchard.

In the case of the Hungarian fiscal policy, however, the stabilisation of government debt relative to GDP cannot be considered as a satisfactory objective. This results on the one hand from our obligation to the European Union, on the other hand from the substantial burden on the budget represented by annual interest payments, accounting for nearly 10 percent of tax revenues. In the following, therefore, the criterion of sustainability will be to reduce government debt below 60 percent of GDP, and to prevent it from growing above this level.

Figure 5. Calculated and actual primary balance as a percentage of GDP



Source: Own calculation.

We should first examine what primary balance would have been needed in the past twelve years – with the difference between real interest and real growth in the particular period – to stabilise government debt at 60 percent in five years. The change of exchange rate and other items affecting government debt relative to GDP are left out of consideration in this test, and the simplified version of the equation used by *Pápa–Valentinyi* [2008] is applied to obtain the calculated primary balance:

$$\overline{pb} = \left(\frac{1+r}{1+g} - 1\right) \frac{\left(\frac{1+r}{1+g}\right)^n - \frac{b^*}{b}}{\left(\frac{1+r}{1+g}\right)^n - 1} b, \qquad /22/$$

where $\overline{pb}$ is the primary balance with which government can stabilise government debt at b^* level as a proportion of GDP in period n. In our case the primary balance

that would stabilise government debt at 60 percent of GDP in five years with fixed difference between real interest and real growth in the particular year is calculated for every year.

The results reveal that from the point of view of sustainability, the turning-point was in 2001/2002. From that time on until the end of the examined period, the primary balance was lower than needed each year, and although the position of the budget improved considerably in the last four years, the value of the difference between real interest and real growth rose because of the recession, so the surplus needed to reach a government debt of 60 percent of GDP increased too.

In the next step we will examine what primary budget balance should be achieved along with the differences measured between real interest and real growth in the past periods, depending on the number of years in which we wish to see government debt go below 60 percent of GDP. The most appropriate method for this is the procedure applied by *Burnside* [2005], which is based on the following formula:

$$x_t = u\frac{(1+u)^J b_t - b^*}{(1+u)^J - 1}, \qquad /23/$$

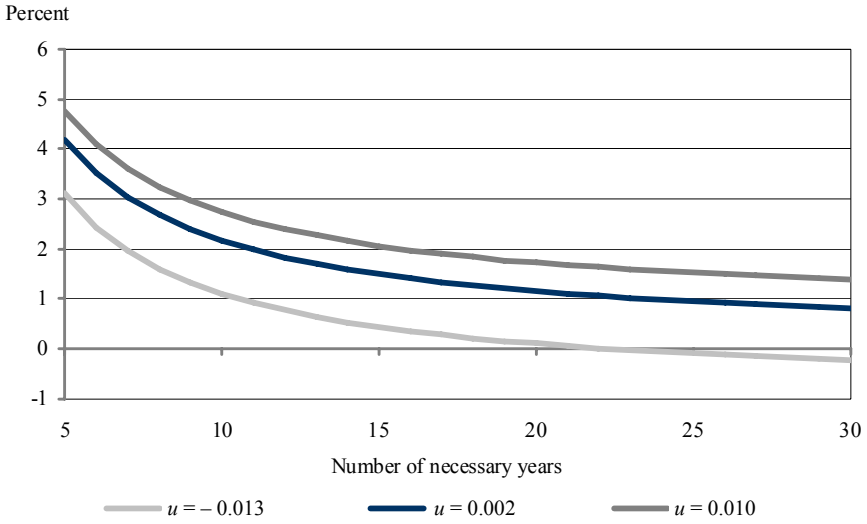
where

$$u = \frac{r-g}{1+g}, \text{ and} \qquad /24/$$

b_t in the equation is current government debt as a proportion of GDP, b^* is the government debt to be reached in period *J*, and the primary balance needed for this is *x*. *r* continues to mean real interest, while *g* real growth. Equation /23/ was calculated with Hungary's data using three different *u*-s. In the most favourable case the difference between real interest and real growth ($u = -0.013$) was the same as the average difference between real interest and real growth calculated for the period lasting from 1999 to 2007, that is, the impact of the crisis was eliminated from past figures. Then a budget surplus of 3.1 percent of GDP should be reached if government debt relative to GDP is to be reduced in five years from 80.2 percent at the end of 2010 to 60 percent, if ten years are allowed, then the necessary primary balance is only 1.1 percent, while in the case of twenty years, 0.1 percent.

In the baseline scenario, the difference between real interest and real growth is the same as the average for the last twelve years ($u = 0.002$). In this case a budget surplus of 4.2 percent of GDP should be reached to lower government debt relative to GDP to 60 percent in five years, while if the Hungarian state wishes to reach the required level in ten or twenty years, the primary surplus of general government can be 2.2 or 1.2 percent of GDP, respectively.

Figure 6. Primary balances needed to reach government debt of 60 percent



Note. The horizontal axis represents the number of years planned to see government debt reach 60 percent of GDP, and the vertical axis the primary balance needed for this.

Source: Own calculation.

In the least favourable scenario it was assumed that the real interest would exceed the rate of growth by 1 percent on average in the future. Then a primary surplus of 4.8 percent is needed to reach the 60 percent level as a proportion of GDP in five years, while if one wants to achieve the same in ten or twenty years, then general government should have a primary surplus of 2.7 or 1.7 percent, respectively.

As the primary balance of general government was 0 in 2010 and the surplus is planned to be 0.7 percent of GDP in 2011, it can be stated that government debt will be decreased to 60 percent of GDP in nineteen and twelve years respectively – depending on whether the balance of 2010 or 2011 is taken into account – according to the most favourable scenario, while more than thirty years will be needed to reach the level aimed at according to both the baseline and the pessimistic scenario.

The results of our investigation are in line with the findings of earlier works in this area (see for example, *Aizenman–Pasricha* [2010]). Though *Aristovnik–Bercic* [2007] obtained a wider negative gap, it can be explained by their estimation based on figures for 2004, when the position of the Hungarian budget was worse than in 2010. In his study *Lewis* [2010] confirms our finding, too, that no large adjustment is needed to maintain the level of debt, while the position of the budget should be improved significantly to reduce government debt below 60 percent in a relatively short time.

4. Conclusions

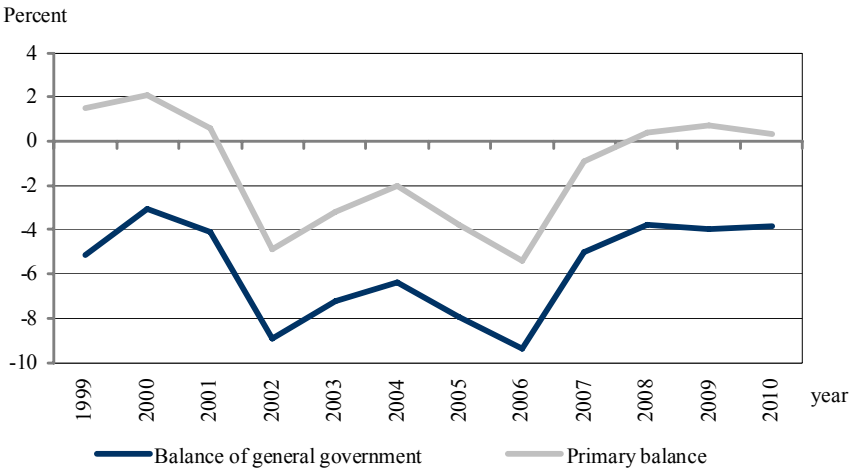
Regarding Hungary's government debt, the last twelve years can be split into three different periods. Primary balance and the difference between real interest and real growth both contributed to government debt decreasing as a proportion of GDP by nearly 10 percentage points until 2001 from 62.0 to 52.2 percent. In the next five years, debt grew by more than 13 percentage points, which was solely due to fiscal policy, additional factors (difference between real interest and real growth, other items) even lowered indebtedness. Although for other reasons, government debt in the last four years of the examined period increased further as a proportion of GDP, by 14.9 percentage points to 80.2 percent. This was caused primarily by the fall of GDP, but the growth of foreign exchange reserves as well as the impact of the change of exchange rate contributed to the further rise of debt level.

Concerning the sustainability of government debt there are two important findings in the light of the results of the fiscal reaction function analysis. On the one hand, relation can be detected between government debt and primary balance, that is, a kind of correction mechanism can be explored in the fiscal policy of the last two decades, which contributes in any case to the sustainability of the process. On the other hand, this correction mechanism (the size of the regression coefficient) is rather weak both in the short and the long term. Value 1 means total correction, in other words, that the increase of debt is fully compensated for by the improvement of primary balance, while the results of the tests of Hungary's data with different parameters range between 0.08 and 0.43.

Based on the investigation of the difference between real interest and real growth, it can be stated that all in all this difference did not play a significant role in indebtedness in the last one and a half decades. If the period between 2008 and 2010, hit by the economic crisis, is left out of consideration, then the average difference between real interest and real growth even reduced debt to a small extent. It can be said for the future that along with the average value of the differences measured in the past period between real interest and real growth, the primary balance of 2010, close to equilibrium, is enough to prevent debt from increasing further, but with government debt falling to 60 percent of GDP in approximately ten years considered as the criterion of sustainability, the primary balance for 2010 should be improved by an additional 1-2 percentage points.

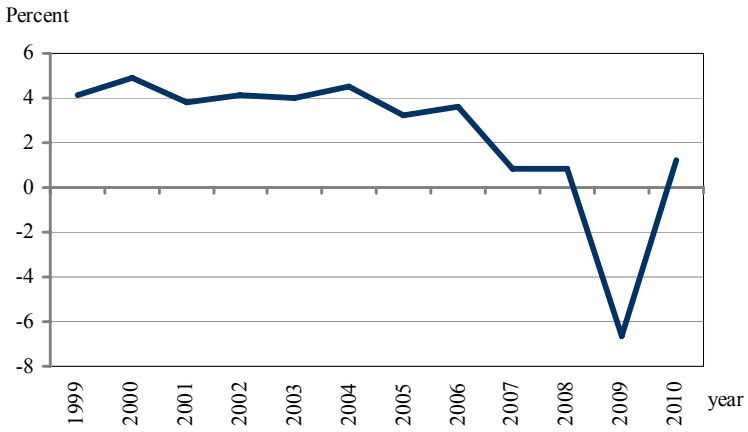
Appendix

Figure A1. Budget balance
(as a percentage of GDP)



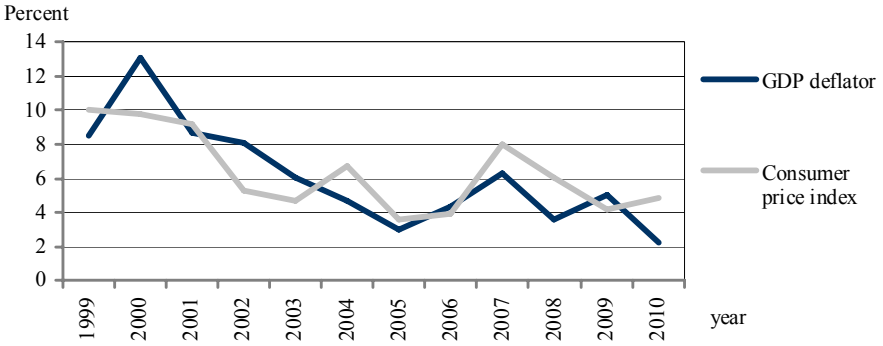
Source: Report on Excessive Deficit Procedure, April 2011.

Figure A2. Volume index of GDP



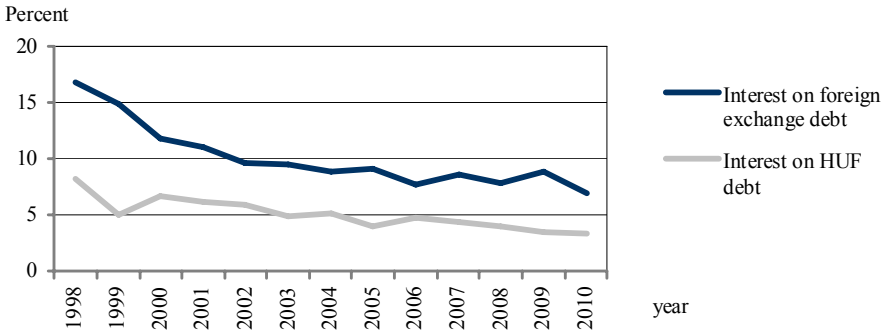
Source: HCSO.

Figure A3. Inflation



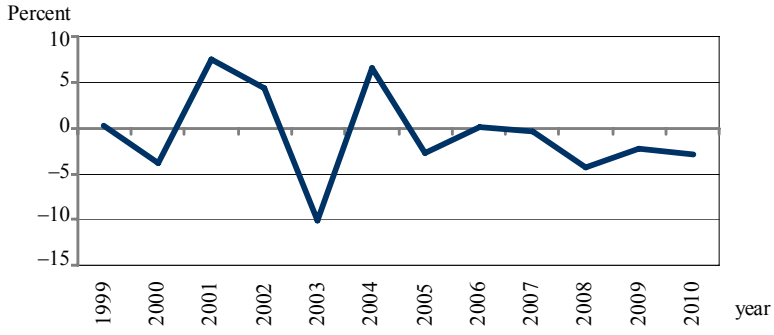
Source: NBH, HCSO.

Figure A4. Interest on HUF and foreign exchange debt



Source: Own calculation.

*Figure A5. Change in exchange rate of HUF against EUR**



* Data refer to end of year, nominal exchange rate; positive change means appreciation.
Source: NBH.

Estimation results for fiscal reaction functions

Variable	Model 1 pb_t	Model 2 d_pb_t	Model 3 pb_cic_t
c	–5.9929** (–2.322)	–10.2416** (–2.891)	–13.097** (–3.664)
d_{t-1}	0.0804* (2.086)	0.1486** (2.763)	0.1975*** (3.777)
pb_{t-1}	0.4353*** (4.107)	0.6035*** (4.537)	
og_t		0.2670 (1.232)	
pb_cic_{t-1}			0.5432*** (3.896)
N	19	19	14
R^2	0.37	0.42	0.66
Adjusted R^2	0.29	0.30	0.60
Akaike value	86.9	87.5	64.6

Note. Because of the HAC (heteroskedasticity and autocorrelation consistent) weight matrix applied in the estimation, *t* statistics are robust, even in the presence of heteroscedasticity and autocorrelation.

References

ÁBEL, I. – KÓBOR, Á. [2011]: Növekedés, deficit és adósság – fenntartható keretben. *Közgazdasági Szemle*. Vol. LVIII. No. 6. pp. 511–528.

AGNELLO, L. – SOUSA, R. M. [2009]: *The Determinants of Public Deficit Volatility*. European Central Bank Working Paper. No. 1042. Frankfurt am Main.

AIZENMAN, J. – PASRICHA, G. [2010]: *Fiscal Fragility: What the Past May Say about the Future*. National Bureau of Economic Research Working Paper Series. No.16478. Cambridge.

ANTAL, J. [2006]: *Külső adósságdinamika*. MNB-tanulmányok. No. 51. Hungarian National Bank. Budapest.

ARDAGNA, S. – CASELLI, F. – LANE, T. [2004]: *Fiscal Discipline and the Cost of Public Debt Service*: *Some Estimates for OECD Countries.* European Central Bank Working Paper. No. 411. Frankfurt am Main.

ARISTOVNIK, A. – BERCIC, B. [2007]: Fiscal Sustainability in Selected Transition Countries. *Journal of Economics*. Vol. 55. No. 7. pp. 659–675.

BALDACCI, E. – GUPTA, S. – MULAS-GRANADOS, C. [2010]: *Restoring Debt Sustainability After Crises: Implications for the Fiscal Mix.* International Monetary Fund Working Paper. No. 232. Washington, D.C.

BALDACCI, E. – KUMAR, M. S. [2010]: *Fiscal Deficits, Public Debt, and Sovereign Bond Yields*. International Monetary Fund Working Paper. No. 184. Washington, D.C.

BARABÁS, GY. – HAMECZ, I. – NEMÉNYI, J. [1998]: A költségvetés finanszírozási rendszerének átalakítása és az eladósodás megfékezése II. *Közgazdasági Szemle*. Vol. XLV. No. 9. pp. 789–802.

BLANCHARD, O. [1990]: *Suggestions for a New Set of Fiscal Indicators*. OECD Department of Economics and Statistics Working Paper. No. 79. Paris.

BOGNETTI, G. – RAGAZZI, G. [2009]: *EU New Member Countries*: *Public Sector Accounts and Convergence Criteria*. Economics of European Integration Working Paper. No. 20. Milano.

BOHN, H. [1998]: The Behavior of U.S. Public Debt and Deficits. *The Quarterly Journal of Economics*. Vol. 113. No. 3. pp. 949–963.

BUITER, W. H. [1985]: A Guide to Public Sector Debt and Deficits. *Economic Policy*. Vol. 1. No. 1. pp. 13–79.

BUITER, W. H. [2004]: *Fiscal Sustainability. Paper presented at The Egyptian Center for Economic Studies.* http://www.nber.org/~wbuiter/egypt.pdf.

BURNSIDE, C. (ed.) [2005]: *Fiscal Sustainability in Theory and Practice*: *A Handbook*. The World Bank Publications. Washington, D.C.

CALLEN, T. – TERRONES, M. – DEBRUN, X. – DANIEL, J. – ALLARD, C. [2003]: Public Debt in Emerging Markets: Is It Too High? In: *International Monetary Fund*: *World Economic Outlook*. Washington, D.C. pp. 113–152.

CÂMPEANU, E. – STOIAN, A. [2010]: Fiscal Policy Reaction in the Short Term for Assessing Fiscal Sustainability in the Long Run in Central and Eastern European Countries. *Czech Journal of Economics and Finance*. Vol. 60. No. 6. pp. 501–518.

CROCE, E. – JUAN-RAMON, H. V. [2003]: *Assessing Fiscal Sustainability: A Cross-Country Comparison.* International Monetary Fund Working Paper. No. 145. Washington, D.C.

CZETI, T. – HOFFMANN, M. [2006]: *A magyar államadósság dinamikája: elemzés és szimulációk*. MNB-tanulmányok. No. 50. Hungarian National Bank. Budapest.

CZIKE, A. O. [2010]: Az állampapír-piaci referenciahozamok a makrogazdaság tükrében. *Hitelintézeti Szemle*. Vol. 9. No. 1. pp. 85–105.

DARVAS, ZS. – KOSTYLEVA, V. [2011]: *The Fiscal and Monetary Institutions of CESEE Countries*. Bruegel Working Paper. No. 2. Brussels.

DE CASTRO, F. – DE COS, P. H. [2002]: On the Sustainability of the Spanish Public Budget Performance. *Revista de Economía Pública*. Vol. 160. No.1. pp. 9–27.

GREINER, A. – KOELLER, U. – SEMMLER, W. [2004]: *Debt Sustainability in the European Monetary Union*: *Theory and Empirical Evidence for Selected Countries.* Center for Empirical Macroeconomics Working Paper. No. 71. University of Bielefeld. Bielefeld.

GYŐRFFY, D. [2005]: Társadalmi bizalom és költségvetési hiány. *Közgazdasági Szemle*. Vol. LIV. No. 3. pp. 274–290.

HALL, G. J. – SARGENT, T. J. [2010]: *Interest Rate Risk and Other Determinants of Post-WWII U.S. Government Debt/GDP Dynamics.* National Bureau of Economic Research Working Paper Series. No. 15702. Cambridge.

IMF (INTERNATIONAL MONETARY FUND) [1997]: *IMF Staff Country Report.* No. 104. Washington, D.C.

IZAK, V. [2009]: *Primary Balance, Public Debt and Fiscal Variables in Postsocialist Members of the European Union*. Prague Economic Papers. No. 2. pp. 114–130.

KARSAI, G. [2006]: Ciklus és trend a magyar gazdaságban 1995–2000 között. *Közgazdasági Szemle*. Vol. LIII. No. 6. pp. 509–525.

KUN, J. [1996]: A seigniorage és az államadósság terhei I–II. *Közgazdasági Szemle*. Vol. XLIII. No. 9. pp. 783–804. and No. 10. pp. 891–904.

LEWIS, J. [2010]: *How Has the Financial Crisis Affected the Eurozone Accession Outlook in Central and Eastern Europe*. De Nederlandsche Bank Working Paper. No. 253. Amsterdam.

MELLÁR, T. [1997]: Egyensúly és/vagy növekedés. *Közgazdasági Szemle*. Vol. XLIV. No. 6. pp. 474–487.

MELLÁR, T. [2002]: Néhány megjegyzés az adósságdinamikához. *Közgazdasági Szemle*. Vol. XLIX. No. 9. pp. 725–740.

MURAKÖZY, L. [2004]: *Már megint egy rendszerváltás – Történelmi tanulságok és tanulatlanságok*. Competitio Könyvek 2. University of Debrecen. Debrecen.

MURAKÖZY, L. [2008]: Magyarország felemelkedése és hanyatlása. *Közgazdasági Szemle*. Vol. LV. No. 2. pp. 149–168.

NICKEL, C. – ROTHER, P. – ZIMMERMANN, L. [2010]: *Major Public Debt Reductions Lessons from the Past, Lessons for the Future*. European Central Bank Working Paper. No. 1241. Frankfurt am Main.

OBSTFELD, M – ROGOFF, K. [2009]: *Global Imbalances and the Financial Crisis*: *Products of Common Causes*. Paper prepared for the Federal Reserve Bank of San Francisco Asia Economic Policy Conference. 18–20 October. Santa Barbara.

OHNSORGE-SZABÓ, L. – ROMHÁNYI, B. [2007]: Hogy jutottunk ide: magyar költségvetés, 2000–2006. *Pénzügyi Szemle*. Vol. LII. No. 2. pp. 239–285.

ORBÁN, G. – SZAPÁRY, GY. [2006]: Magyar fiskális politika: quo vadis? *Közgazdasági Szemle*. Vol. LIII. No. 4. pp. 293–309.

P. KISS, G. [1998]: *Az államháztartás szerepe Magyarországon.* MNB-füzetek. No. 4. Hungarian National Bank. Budapest.

PÁPA, L. – VALENTINYI, Á. [2008]: Költségvetési fenntarthatóság. *Közgazdasági Szemle*. Vol. LV. No. 5. pp. 395–426.

PRESBITERO, A. F. [2010]: *Total Public Debt and Economic Growth in Developing Countries*. Money and Finance Research Group Working Paper. No. 44. Ancona.

RA, S. – RHEE, C. Y. [2005]: *Managing the Debt*: *An Assessment of Nepal's Public Debt Sustainability*. Nepal Resident Mission Working Paper. No. 6. Asian Development Bank. Mandaluyong City.

REINHART, C. M. – ROGOFF, K. S. – SAVASTANO, M. A. [2003]: Debt Intolerance. *Brookings Papers on Economic Activity*. No. 1. pp. 1–74.

REINHART, C. M. – ROGOFF, K. S. [2010]: Growth in a Time of Debt. *American Economic Review*. Vol. 100. No. 2. pp. 573–578.

STEIN, J. L. [2011]: *The Diversity of Debt Crises in Europe*. CESIFO Working Paper. No. 3348. Munich.

TÓTH, G. CS. [2011]: Adósságdinamika és fenntarthatóság. *Statisztikai Szemle*. Vol. 89. No. 12. pp. 1242–1268.

TÖRÖK, Á. [2011]: Költségvetési fenntarthatóság és átláthatóság – Fiscal Policy Councils: Why Do We Need Them and What Makes Them Effective? *Közgazdasági Szemle.* Vol. LVIII. No. 4. pp. 368–373.

Short Introduction to the Generalized Method of Moments*

Peter Zsohar
PhD Student
Central European University
E-mail:
zsohar_peter@ceu-budapest.edu

The generalized method of moments (GMM) is the centrepiece of semiparametric estimation frameworks. After putting GMM into context and familiarizing the reader with the main principles behind the method, we discuss the estimation procedure and the properties of the GMM estimator in details. We also provide a short survey of recent research areas in the field. To facilitate understanding, most concepts are illustrated by simple examples.

KEYWORDS:
GMM.
Semiparametric Estimation.

* Acknowledgements: The author wants to thank both *Laszlo Hunyadi*, Editor-In-Chief of the *Hungarian Statistical Review* and *Sergei Lychagin* referee, Assistant Professor of the Central European University for their helpful comments which have improved the study. All remaining errors are the author's.

Econometric analysis begins with some economic phenomenon that is of interest to us that we intend to analyse. First we turn to economic theory to see what insights it can offer. It postulates an explanation in some sort of conditions that describe the phenomena in terms of the key economic variables and model parameters. However, to answer specific questions, we have to quantify the parameters involved. We would like to adopt an estimation method whose implementation does not require the imposition of additional restrictions to the data generating process beyond those implied by the economic model. If it turns out that just for the purpose of getting these estimates we have to place further restrictions and make more assumptions and these are found to be unjustified by theory or inappropriate for the data then we run the risk that the invalidity will undermine all our subsequent inferences about the phenomenon of our interest. We would like to use a method of statistical estimation that fits well with exactly the kind of information we are getting out of our economic models. But what form does that information take? Very often restrictions implied by economic theory take the form what we will refer to as population moment conditions. The generalized method of moments (GMM) is a statistical method that combines observed economic data with the information in population moment conditions to produce estimates of the unknown parameters of this economic model. Once we have those parameters, we can go back to perform inference about the basic question that is of interest to us. Shortly we will see that GMM is very well tailored exactly to the kind of information we are getting out from our economic models.

The purpose of this article is to provide an introduction to the GMM framework and to give a rough picture of current on-going issues in the field. There are excellent textbooks and reference books available on the topic which are more precise and elaborate in all aspects like *Mátyás* [1999] or *Hall* [2005]. We will heavily rely on them and the interested reader is encouraged to study them. Our treatment misses many details but all simplifications were made to facilitate easy understanding.

After introducing the principle of the method of moments in Section 2, we show how to generalize the idea into GMM in Section 3. In Section 4 we discuss the properties of the GMM estimator. The estimation procedure is described in Section 5, while Section 6 provides a short description of testing in the GMM framework. We will also address briefly the question of moment selection in Section 7. After a short survey of the recent research in Section 8, Section 9 concludes.

1. The method of moments principle

The population moment conditions will play a crucial role in the discussion so it is worth going back to the primitives to understand the mechanics of GMM.

The raw uncentered moments are easy to compute and they reveal important aspects of a distribution. For example, the first four moments tell us about the population mean, variance, skewness and kurtosis. Using them we can immediately place restrictions according to our theory on the location, scale or shape of the distribution without specifying a full model or distribution.

Once we have some information on the population, the question remains how to use the sample to estimate the parameters of interest. In general, sample statistics each have a counterpart in the population, for example, the correspondence between the sample mean and the population expected value. The natural next step in the analysis is to use this analogy to justify using the sample moments as bases of estimators of the population parameters. This was the original idea in *Karl Pearson*'s work [1893], [1894], [1895] in the late 19th century.

The Pearson family of distributions is a very flexible mathematical representation that has several important and frequently used distributions among its members depending on the parameterization you choose. Pearson's problem was to select an appropriate member of the family for a given dataset.

Example 1 – Simple method of moments estimator

To show a very simple example, assume that the population distribution has unknown mean μ and variance equal to one. In this case, the population moment condition states that $E[x_i] = \mu$. If $\{x_i : i = 1,2,\ldots,n\}$ is an independent and identically distributed sample from the distribution described formerly, then the sample average $\bar{x} = \frac{1}{n}\sum_{i=i}^{n} x_i$ is the sample analogue to the population mean $E[x_i]$. By utilizing this analogy principle, the method of moments (MM) estimator for $E[x_i] = \mu$ is simply given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu}_n$.

Basically we had to work out the first moment, then to replace it with the sample analogue and to solve the equation for the unknown parameter. What remains to be established is whether this approach is the best, or even a good way to use the sample data to infer the characteristics of the population.[1] Our intuition suggests that the bet-

[1] We will return to this subject in Section 4 discussing the properties of the GMM estimator.

ter the approximation is for the population quantity by the sample quantity, the better the estimates will be.

To make a step further, it is time to introduce some more general definitions.

Definition 1 – Method of moments estimator

Suppose that we have an observed sample $\{x_i : i = 1,2,\ldots,n\}$ from which we want to estimate an unknown parameter vector $\theta \in \mathrm{R}^p$ with true value θ_0. Let $f(x_i,\theta)$ be a continuous and continuously differentiable $\mathrm{R}^p \to \mathrm{R}^q$ function of θ, and let $E[f(x_i,\theta)]$ exist and be finite for all i and θ. Then the population moment conditions are that $E[f(x_i,\theta_0)] = 0$. The corresponding sample moments are given by

$$f_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(x_i,\theta).$$

The method of moments estimator of θ_0 based on the population moments $E[f(x_i,\theta)]$ is the solution to the system of equations $f_n(\theta) = 0$.

Note that if $q = p$, then for an unknown parameter vector θ the population moment conditions $E[f(x_i,\theta)] = 0$ represent a set of p equations for p unknowns. Solving these moment equations would give the value of θ which satisfies the population moment conditions and this would be the true value θ_0. Our intuition suggests that if the sample moments provide good estimates of the population moments, we might expect that the estimator $\hat{\theta}$ that solves the sample moment conditions $f_n(\hat{\theta}) = 0$ would provide a good estimate of the true value θ_0 that solves the population moment conditions $E[f(x_i,\theta_0)] = 0$.

Now we present some common models in terms of the MM terminology.

Example 2 – Ordinary least squares (OLS)

Consider the linear regression model

$$y_i = x_i'\beta_0 + u_i,$$

where x_i is the vector of p covariates, β_0 is the true value of the p unknown parameters in β, and u_i is an exogenous error term. In this case our population moment condition $E\left[f\left(x_i, \theta_0\right)\right] = 0$ translates to $E\left[x_i u_i\right] = E\left[x_i \left(y_i - x_i'\beta_0\right)\right] = 0$. Then the sample moment conditions are given by

$$\frac{1}{n}\sum_{i=1}^{n} x_i \hat{u}_i = \frac{1}{n}\sum_{i=1}^{n} x_i \left(y_i - x_i'\hat{\beta}\right) = 0 .$$

Thus the MM estimator of β_0 is given by $\hat{\beta}$ that solves this system of p linear equations and is equivalent to the standard OLS estimator.

Example 3 – Instrumental variables (IV)

If in Example 2 we allow u_i to be correlated with the covariates in x_i, we can state the population moment conditions in terms of the exogeneity assumption on the p instruments. Our population moment conditions are given by $E\left[z_i u_i\right] = E\left[z_i \left(y_i - x_i'\beta_0\right)\right] = 0$ and the sample moment conditions are

$$\frac{1}{n}\sum_{i=1}^{n} z_i \left(y_i - x_i'\hat{\beta}\right) = 0 .$$

Just like previously, the MM estimator of β_0 is given by $\hat{\beta}$ that solves this system of p linear equations and this result shows that the standard IV estimator is also an MM type estimator.

Note that as long as the exogeneity of the error term and the instrument can be justified by economic reasoning, these examples do not impose any additional restrictions on the population that is not implied by some theory.

Example 4 – Maximum likelihood (ML)

In case we have a fully specified model, the sample log-likelihood is $\frac{1}{n}\sum_{i=1}^{n} l\left(\theta \middle| x_i\right)$. The first order conditions for the maximization of the log-likelihood function are then

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial l\left(\theta\middle|x_i\right)}{\partial\theta}\bigg|_{\theta=\hat{\theta}}=0\,.$$

These first order conditions can be regarded as a set of sample moment conditions so the maximum likelihood estimator can be given an MM interpretation as well.

So far we have considered cases where the number of moment conditions q was equal to the number of unknown parameters p. Assuming functionally independent moment equations, the resulting system of equations provided by the moment conditions can be solved to obtain the MM estimator. In the case of $q < p$ there is insufficient information and the model is not identified. If $q > p$, the model is over-identified, and in most cases, we are not able to solve the system of equations. However, estimation still can proceed and the next section will show the proper way to follow.

2. The GMM Estimator

We shall recall that population moment conditions represent information implied by some theory. It is quite natural that we want to use the most information available.[2] Unfortunately the MM estimator cannot incorporate more moments than parameters.[3]

Example 5 – Motivation for GMM

Consider again Example 1. Notice that our estimation was based solely on the first raw moment of the distribution. Now suppose that we believe to know that the sample at hand is a result of n independent draws from a Poisson distribution with parameter λ. Thus the new (additional) population moment condition based on the second raw moment is $E\left[x_i^2\right]-\lambda^2-\lambda=0$. The MM estimator of λ should satisfy the system of equations based on the sample moments

[2] Resisting the temptation to impose additional assumptions that might be unjustified by theory.

[3] However, there are still many possible actions one might think of like using all different sets of moments and then averaging the estimates, etc. but here this is not the road taken.

$$\begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i - \hat{\lambda} \\ \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2\right) - \hat{\lambda}^2 - \hat{\lambda} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Now we have two moment conditions and one unknown parameter which means that we do not have a general solution for $\hat{\lambda}$.

We could use only p number of moments to estimate the parameters but by dismissing the $q - p\,(>0)$ additional moments, we would loose the information contained in those conditions. The remedy for this situation was introduced to the econometrics literature by *Hansen* [1982] in his famous article and it is called GMM. The idea behind GMM estimation is that once it is impossible to solve the system of equations provided by the sample moment conditions, we can still have an estimate of θ that brings the sample moments as close to zero as possible.[4] Note that in the population still all moment conditions hold and the problem arises because we have a finite sample.

Definition 2 – Generalized method of moments estimator

Suppose that the conditions in Definition 1 are met and we have an observed sample $\{x_i : i = 1,2,\ldots,n\}$ from which we want to estimate an unknown parameter vector $\theta \in \Theta \subseteq \mathbb{R}^p$ with true value θ_0. Let $E\left[f\left(x_i, \theta\right)\right]$ be a set of q population moments and $f_n\left(\theta\right)$ the corresponding sample counterparts. Define the criterion function $Q_n\left(\theta\right)$ as

$$Q_n\left(\theta\right) = f_n\left(\theta\right)' W_n f_n\left(\theta\right),$$

where W_n, the weighting matrix, converges to a positive definite matrix W as n grows large. Then the GMM estimator of θ_0 is given by

$$\hat{\theta} = \underset{\theta \in \Theta}{arg\,min}\, Q_n\left(\theta\right).$$

[4] There are, of course, some statistical antecedents to GMM. The method of minimum Chi-square by *Neyman* and *Pearson, E.* [1928] deals with the general question how to estimate parameters when having more moment conditions than unknown parameters. However, they did not work with population moment conditions explicitly, the general idea was basically the same.

Basically the GMM estimator is a way of exploiting information from our general form of population moment conditions. When the number of moment conditions equals the number of unknown parameters GMM = MM. When $q > p$ then the GMM estimator is the value of θ closest to solving the sample moment conditions and $Q_n(\theta)$ is the measure of closeness to zero.

It might be useful to have a look at two practical applications from the literature that result in over-identifying moment conditions.

Example 6 by *Hansen* and *Singleton* [1982]

In their classical paper they analysed the movement of assets over time in a consumption-based capital asset pricing model. In a somewhat simpler version of their non-linear rational expectations model, the representative agent maximizes expected discounted lifetime utility

$$E\left[\sum_{\tau=0}^{\infty}\beta^{\tau}U\left(c_{t+\tau}\right)\middle|\Omega_t\right]$$

subject to the budget constraint

$$c_t + p_t q_t \le r_t q_{t-1} + w_t \qquad \forall t\,,$$

where c_t is per period consumption, p_t, q_t, r_t are relative price, quantity and return on the asset with one period maturity, w_t is real wage and Ω_t is the information set of the agent in period *t*. Hansen and Singleton use a constant relative risk aversion utility function $U(c) = \left(c^{\gamma} - 1\right)\Big/\gamma$ so the first order conditions to this optimization problem are

$$E\left[\beta\left(\frac{c_{t+1}}{c_t}\right)^{\gamma}\frac{r_{t+1}}{p_t} - 1\middle|\Omega_t\right] = 0\,.$$

This looks pretty much like a population moment condition but the problem is that we have two parameters to estimate (β,γ) and only one moment condition. However, by an iterated conditional expectations argument for any vector $z_t \in \Omega_t$ the Euler-equation becomes

$$E\left[\left(\beta\left(\frac{c_{t+1}}{c_t}\right)^{\gamma}\frac{r_{t+1}}{p_t}-1\right)z_t\right]=0\,,$$

so in theory the model is identified by using any variables that are known to the agent in period t, such as lagged values of r_t/p_{t-1} or c_t/c_{t-1} and can be estimated consistently with GMM.[5] In contrast, maximum likelihood estimation of this model would involve exactly specifying conditional distributions of the variables and a lot of numerical integration which is computationally burdensome.

In the structural model from the previous example we were originating population moment conditions from what we were referring to as economic theory. However, sometimes "economic theory" means just some plausible assumptions based on intuition or other reasoning. Next we show an example which is based on a much less structural model, and moment conditions come from the exogeneity assumption on the instrumental variables.

Example 7 by *Angrist* and *Krueger* [1991]

The authors investigate the number of years spent in education and the subsequent earning potentials of individuals. They were interested in the impact of compulsory schooling laws in the US and estimated the following equation:

$$ln(w_i)=\beta_0+\beta_1 ed_i+controls+u_i\,.$$

The parameter of interest was β_1, the semi-elasticity of wage with respect to education. Estimating this linear equation by OLS could be biased and inconsistent as ed_i is probably correlated with individual factors in the regression error term u_i such as individual costs and potential benefits of schooling or other options outside the schooling system, most of which are unobserved by the researcher. Using the structure of compulsory school attendance laws at that time in the US they were able to argue that (in addition to the controls) dummy variables indicating the quarter of birth for each individual could be used to in-

[5] Note that the original variables in the model need not be stationary as taking consequent ratios makes the series stationary.

strument for the years spent in education. Their exogeneity assumption implies that the following population moment conditions hold:

$$E\left[z_i \left(ln\left(w_i \right) - \beta_0 - \beta_1 ed_i - controls \right) \right] = 0 ,$$

where the vector of instruments z_i contains the exogenous variables from the original model supplemented by the quarter of birth dummies. Note that there are more moment conditions then parameters and we could estimate the model by GMM.

3. Properties of the GMM Estimator

Under some sufficient conditions the GMM estimator as given in Definition 2 is consistent and asymptotically normally distributed. In the following we will discuss these properties and the sufficient conditions in somewhat more detail.

Population moment conditions provide information about the unknown parameters. The quality and the utilization method of this information are crucial in several aspects. First, a natural question arises about the sufficiency of the information contained in the moment conditions whether it is enough for the estimation to be "successful". This leads us to the issue of identification.

Assumption 1 – Identification

In the following, we present the necessary conditions for identification.

– *Order condition*: As we have already seen if $q < p$, the model is not identified and we are unable to estimate the parameters. So we need $q \geq p$.

– *Rank condition*: Once we have enough moment conditions, it is still crucial that among those moments should be at least *p* functionally independent ones which are satisfied if the expectation of the $q \times p$ Jacobian of the moment equations evaluated at θ_0 has rank (at least) *p*.

– *Uniqueness*: If we think of $E\left[f\left(x_i, \theta \right) \right]$ as a function of θ, then for successful estimation $E\left[f\left(x_i, \theta \right) \right] = 0$ has to be a unique property of θ_0. It means that θ_0 should be the only parameter vector which satisfies the population moment conditions.

We also need to establish a connection between the population moments and their sample counterparts. This will ensure that in the limit, the true parameter vector will be the one that solves the sample moment equations.

Assumption 2 – Convergence of sample moments

If the data generating process is assumed to meet the conditions for some kind of law of large numbers to apply, we may assume that the sample moments converge in probability to their expectation. That is $f_n(\theta_0)\left(=\frac{1}{n}\sum_{i=1}^{n} f(x_i,\theta_0)\right)$ converges to $E\left[f_n(\theta_0)\right]=E\left[f(x_i,\theta_0)\right]=0.$

Note that we have basic laws of large numbers only for independent observations. For a more general case, with dependent or correlated observations, we would assume that the sequence of observations $f(x_i,\theta)$ constitutes an ergodic q-dimensional process. Assumptions 1 and 2 together with the conditions from Definitions 1 and 2 establish that the parameter vector will be estimable.

Now we make a statistical assumption that allows us to establish the properties of asymptotic distribution of the GMM estimator.

Assumption 3 – Distribution of Sample Moments

We assume that the sample moments obey a central limit theorem. This assumes that the moments have a finite asymptotic covariance matrix, $(1/n)F$, so that

$$\sqrt{n}\, f_n(\theta_0) \xrightarrow{d} N(0,F).$$

Again, if the observations are not independent, it is necessary to make some assumptions about the data so that we could apply an appropriate central limit theorem.

Theorem 1 – GMM is consistent and asymptotically normal

Under the preceding assumptions, the GMM estimator is consistent and asymptotically normally distributed with asymptotic covariance matrix V_{GMM} defined as

$$V_{GMM}=\frac{1}{n}\left[G(\theta_0)'WG(\theta_0)\right]^{-1}G(\theta_0)'WFWG(\theta_0)\left[G(\theta_0)'WG(\theta_0)\right]^{-1},$$

where G is a $(q \times p)$ matrix defined as

$$G(\theta) = E\left[\frac{\partial f(x,\theta)}{\partial \theta}\right] = E\begin{bmatrix} \frac{\partial f_1(x,\theta)}{\partial \theta_1} & \cdots & \frac{\partial f_1(x,\theta)}{\partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_q(x,\theta)}{\partial \theta_1} & \cdots & \frac{\partial f_q(x,\theta)}{\partial \theta_p} \end{bmatrix},$$

that is, $G(\theta_0)$ is the expected value of the Jacobian of the population moment functions evaluated at the true parameter value θ_0.

The point is that in general the variance of the GMM estimator depends on the choice of W_n, so we can extract the most information from the moment conditions by choosing an appropriate weighting matrix. By analysing the quadratic form in the GMM criterion function

$$Q_n(\theta) = f_n(\theta)' W_n f_n(\theta),$$

we see that setting $W_n = I$ gives us the sample moments' error sum of squares.

In fact, we could use a diagonal weighting matrix $W_n = \langle w \rangle$ to minimize the weighted sum of squared errors. This is a natural idea as some moments might be more volatile than others and, thus, it makes sense to normalize the errors in the moments by their variance.

However, the elements of $f_n(\theta)$ are freely correlated. Suppose the asymptotic covariance of the sample moments normalized by the root of the sample size is $Asy.Var\left[\sqrt{n} f_n(\theta_0)\right] = F$. Then the choice of $W_n = F^{-1}$ weights all elements of the criterion function appropriately so should be optimal based on the same idea that motivates generalized least squares.

Theorem 2 – Optimal weighting matrix

For a given set of moment conditions with the optimal choice of the weighting matrix $W_n = F^{-1}$, the GMM estimator is asymptotically efficient with covariance matrix

$$V_{GMM,optimal} = \frac{1}{n}\left[G(\theta_0)' F^{-1} G(\theta_0)\right]^{-1}.$$

It is important to emphasize that the efficiency result is valid only for a given set of moment conditions. That is, GMM is asymptotically efficient in the class of consistent and asymptotically normal estimators that do not use any additional information on top of that is contained in the moment conditions. The traditional ML utilizes a fully specified distribution so the two estimators are incomparable in a sense that they rely on different information sets. However, as we saw earlier in Example 4 if the moment conditions are the same as the first order conditions of ML estimation then the two estimators are numerically equal.

Especially, if the model is correctly specified and the underlying distribution is one from the exponential family, we can use the sufficient statistics as bases for moment conditions. In these cases GMM is efficient in a sense that it attains the Cramer–Rao lower bound asymptotically. The problem with this theoretical case is that it is unoperational as GMM's main strength is not specifying an exact distribution.

4. Estimation

After having discussed the properties of the GMM estimator, it is time to turn to some more practical issues like estimation. The question is how do we get those numbers when we have the data.

In the exactly identified case when $q = p$, GMM works the same as MM and there is no need for optimization as the system of moment conditions can be solved for the unknown parameters.

Example 8 – Exactly identified case

Consider the Poisson model from Example 5. Recall that we had two moment conditions for the single unknown parameter λ. Suppose we have a sample of $n = 20$ observations. Now we are going to estimate λ based on both moment conditions separately. The estimators relying on the first two raw moments are

$$\hat{\lambda}_{first} = \frac{1}{20}\sum_{i=1}^{20} x_i = 3.55\,,$$

$$\hat{\lambda}_{second} = \frac{-1 + \sqrt{1 + 4\frac{1}{20}\sum_{i=1}^{20} x_i^2}}{2} = 3.3859\,.$$

In order to utilize both moments at once, we need to compute the GMM estimator. Recall that in the over-identified case when $q > p$ the asymptotic variance of the GMM estimator, V_{GMM} depends on the weighting matrix. We want to get the most information out of our moment conditions thus we would like to use the optimal weighting matrix that minimizes V_{GMM} . As we discussed earlier, this would be F^{-1}. Logic suggests that first we should estimate the optimal weighting matrix so that we could use it in the criterion function to estimate θ_0 efficiently. The problem is that to get an estimator of F^{-1}, we already need an estimate of θ_0 .

We can resolve this circularity by adopting a multi-step procedure.

1. We can choose a sub-optimal weighting matrix, say *I,* and minimize the simple sum of squared errors in the moments $Q_n(\theta) = f_n(\theta)' f_n(\theta)$. This will deliver a preliminary but consistent estimate of θ_0 which can be used then to estimate F and thus F^{-1} consistently.

2. With the optimal weighting matrix estimate at hand, we can minimize the new criterion function $Q_n(\theta) = f_n(\theta)' \hat{F}^{-1} f_n(\theta)$, and estimate θ_0 efficiently.

This is the so-called two-step GMM estimator which is consistent and efficient.

Example 9 – GMM estimation

We now continue with the Poisson example. In the first step we have to minimize the criterion function with using I as a weighting matrix:

$$Q_n(\theta) = f_n(\theta)' f_n(\theta) = \begin{bmatrix} \frac{1}{20}\sum_{i=1}^{20} x_i - \lambda \\ \frac{1}{20}\sum_{i=1}^{20}\left(x_i^2\right) - \lambda^2 - \lambda \end{bmatrix}' \begin{bmatrix} \frac{1}{20}\sum_{i=1}^{20} x_i - \lambda \\ \frac{1}{20}\sum_{i=1}^{20}\left(x_i^2\right) - \lambda^2 - \lambda \end{bmatrix}.$$

To facilitate computation, we started the optimization routine from the MM estimate based on the first raw moment $\hat{\lambda}_{first} = 3.55$. The first-step GMM estimate is

$$\hat{\lambda}_1 = 3.3885$$

which can be used to estimate F as follows:

$$\hat{F}_n = \frac{1}{n}\sum_{i=1}^{n} f\left(x_i,\hat{\theta}_1\right) f\left(x_i,\hat{\theta}_1\right)' = \frac{1}{20}\sum_{i=1}^{20}\begin{bmatrix} x_i-\hat{\lambda}_1 \\ x_i^2-\hat{\lambda}_1^2-\hat{\lambda}_1 \end{bmatrix}\begin{bmatrix} x_i-\hat{\lambda}_1 \\ x_i^2-\hat{\lambda}_1^2-\hat{\lambda}_1 \end{bmatrix}'.$$

Substituting in for $\hat{\lambda}_1$ and inverting $\hat{F}$ the estimated optimal weighting matrix is

$$\hat{W}_{optimal} = \hat{F}_n^{-1} = \begin{bmatrix} 10.5333 & -1.4504 \\ -1.4504 & 0.2085 \end{bmatrix}.$$

In the second step we have to minimize the new criterion

$$Q_n(\theta) = f_n(\theta)' \hat{F}^{-1} f_n(\theta).$$

The optimization routine was started from the first-step GMM estimate. Solving the minimization problem, for the second-step GMM estimator of λ we get

$$\hat{\lambda}_{GMM} = 3.2651.$$

We still have to estimate the variance of the estimator, V_{GMM}. First we recompute $\hat{F}_n^{-1}$ with $\hat{\lambda}_{GMM}$ at hand exactly as we did previously. Then we also have to estimate G, the matrix of the derivatives. G is the expected value of the Jacobian but notice that in our case the derivatives do not depend directly on the data so it can be estimated simply as

$$\hat{G} = \begin{bmatrix} -1 \\ -2\hat{\lambda}_{GMM} & -1 \end{bmatrix}.$$

Now we can compute the estimated variance of $\hat{\lambda}_{GMM}$ as

$$\hat{V}_{GMM} = \frac{1}{n}\left[\hat{G}'\hat{F}_n^{-1}\hat{G}\right]^{-1} = 0.0973.$$

Notice that the MM estimate based on the first raw moment equals the ML estimate for which we can estimate the asymptotic variance from the Cramer–Rao lower bound as $\hat{\lambda}/n$.

The Table compares the results.

Comparison of estimators

Results	ML	Two-step GMM
λ	3.55	3.2651
Standard error	0.4213	0.3119

In fact, we could continue this multi-step procedure to obtain the so-called iterated GMM estimator. *Hansen*, *Heaton* and *Yaron* [1996] suggest a method where the dependence of the weighting matrix on the unknown parameters is acknowledged and taken care of during the optimization procedure. Their approach became known as the continuously updated GMM. There is fairly compelling evidence to suggest that there are gains to iteration in terms of finite sample performance of the estimator but in most cases the two-step estimator is applied.

Given the mathematical form of the moment conditions, in some cases we can solve the optimization problem analytically and get a closed form representation of the estimates in terms of the data which will speed up the computations. However, in other cases such as with nonlinear models, we have to use numerical optimization routines. The problem with the widely used Newton–Raphson and other practical numerical optimization methods is that global optimization is not guaranteed. The GMM estimator is defined as a global minimizer of a GMM criterion function, and the proof of its asymptotic properties depends on this assumption. Therefore, the use of a local optimization method can result in an estimator that is not necessarily consistent or asymptotically normal.

Care should be taken with nonconvex problems where the existence of possibly multiple local minima may cause problems. With starting the optimization algorithm from several initial values spread out in the parameter space, one might be able to find the global minimum. However, it should be noted that the multi-start algorithm does not necessarily find he global optimum and is computationally intensive.

There are of course much more advanced numerical techniques and there is a freely available and fairly user friendly GMM toolbox for MATLAB by *Kyriakoulis* [2004].

An alternative solution in such cases is the use of Monte Carlo simulation methods to compute an otherwise intractable criterion function. The method of simulated moments approach is the simulated counterpart of the traditional GMM procedure and is applicable when the theoretical moments cannot be computed analytically. An extensive survey of recent (mostly theoretical) results in the subject can be found in *Liesenfeld–Breitung* [1999].

5. Testing in the GMM framework

Most of the times there are three broad inference questions that are of interest to us:

– Is the model correctly specified?
– Does the model satisfy certain particular restrictions?
– Which model appears to be more consistent with the data?

The first question is particularly important. Recall that the population moment conditions were deduced from an underlying economic model and all our inference is going to be based on them. As our estimate is relying on the information contained in the moment conditions, it is crucial whether the original model is consistent with the data or whether it appears to be a good representation of the data.

If the hypothesis of the model that led to the moment equations in the first place is incorrect, at least some of the sample moment restrictions will be systematically violated. This conclusion provides the basis for a test of the over-identifying restrictions and if we have more moments than parameters, we have scope for testing that. There is a very simple to compute statistic to use as an over-identifying restrictions test (the so-called J test) which is just the sample size times the value of the GMM criterion function evaluated at the second step GMM estimator

$$nQ_n\left(\hat{\theta}\right)=\left[\sqrt{n}\,f_n\left(\hat{\theta}\right)\right]'\left(Est.Asy.Var\left[\sqrt{n}\,f_n\left(\theta_0\right)\right]\right)^{-1}\left[\sqrt{n}\,f_n\left(\hat{\theta}\right)\right].$$

Notice that this is a Wald statistic and under the null

$$H_0{:}E\left[\,f\left(x_i,\theta_0\right)\right]=0\,,$$

and it has a large sample Chi-squared distribution with $q-p$ degrees of freedom. However, the over-identifying restrictions test can be computed only in case of $q>p$, as in the exactly identified model the criterion function is zero. The reason for the importance and the popularity of this test is that it really examines the heart, the crux of GMM, and it is easy to calculate, as it is an obvious by-product of the estimation procedure. The statistic is ubiquitously reported in all applications involving GMM estimation just as reporting the log of the likelihood function in ML estimation. We would like to stress that it is very important to do some kind of misspecification test as in misspecified models the properties of the GMM estimator are substantially different which is likely to make all subsequent inferences misleading.

Example 10 – Test for over-identifying restrictions
Consider the Poisson model from Example 9. Now we have

$$J_n = nQ_n\left(\hat{\lambda}_{GMM}\right) = 20 \times 0.2694 = 5.388 .$$

As $J_n \sim \chi^2[1]$, there is only very weak evidence in favour of the population moment conditions so the model can be rejected.[6]

The second inference question asks whether the model satisfies certain additional restrictions implied by economic and statistical theory that we could impose and what might tell us about economic behaviour. Fortunately all the well-known likelihood-based testing procedures have their GMM counterparts with very similar implementations. The GMM-based LR test is computed by using nQ_n instead of $\ln L$ in the test statistic. The GMM-based Wald statistic is computed identically to the likelihood-based one by using the GMM estimates instead of the ML estimates. The LM test is derived by the same logic applied to the derivatives of the GMM criterion function.

The third question is model selection. The previously mentioned tests are applicable for nested models but selection from a set of non-nested models would require specifying the distribution of the data generating process.

Those interested in details should read the extensive discussion in Chapter 5 of *Hall* [2005].

6. Choice of moment conditions

So far we have covered how we can exploit information from our moment conditions in an efficient way but we haven't mentioned what is the best set of moment conditions to be used. It turns out that there is quite a straightforward answer to this, although it won't be very useful in terms of practical work.

Maximum likelihood is the asymptotically efficient estimator in the class of consistent and asymptotically normal estimators. Recall that we have already shown that ML is an MM type estimator based on the score function. Thus, if we use the derivatives of the log-likelihood function as moments, we will get an efficient estimator. Unfortunately this is not feasible as in most economic settings the population distribution is

[6] However, the small sample size and the discrete nature of the Poisson distribution should raise some concerns.

unknown. Making an additional assumption on the underlying distribution places restrictions on the economic variables involved that might be unjustified by economic theory and that is exactly the kind of thing that GMM was designed to help us avoid.

But what if we have more moment conditions than parameters? We might expect that more information never hurts but it turns out that sometimes in fact it doesn't help either. There are two main approaches in the literature to moment selection. One suggests optimal moment condition selection based on asymptotic theory among the class of generalized instrumental variables. Unfortunately in many settings it is infeasible just like with the score function. The other strand of literature emphasizes practical data based moment selection introducing different selection criteria. Some results suggest that they may help avoid situations where the asymptotic approximation of finite sample behaviour is poor. For a detailed summary of recent results please see *Hall* [2005].

7. Actively researched topics

All our reasoning and inferences so far were based on large sample theory. Two important questions arise:

– How well does this theory approximate finite sample behaviour in the kind of places where we want to apply GMM?

– Can we identify factors and aspects of model specification that appear to affect the quality of this approximation?

Numerous studies try to address these issues in the literature. There are analytical approaches based on higher order asymptotics and simulation supported studies applied to generated artificial data from structures to which we typically fit our economic theories. A detailed discussion and summary of the topic can be found in *Podivinsky* [1999] or in Chapter 6 of *Hall* [2005]. *Harris* and *Mátyás* [2004] provide an extensive comparative analysis of different IV and GMM estimators, focusing on their small sample properties. These studies assess how well the methods perform and the findings are perhaps not that surprising. Loosely speaking, sometimes GMM works well but sometimes it does not. The main factors that were found to affect the quality of asymptotic approximation are:

– form of moment conditions $f\left(x_i,\theta\right)$. Basically, the more nonlinearity is involved, the less good the approximation is;

– degree of over-identification $\left(q-p\right)$;

– interrelation between the elements of moment conditions;
– quality of identification.

How can we improve on the quality of inference then? There are three main strands of responses in the literature, two of which stay in the GMM framework and one suggesting another method.

– If we would like to stick to first order asymptotic theory, the method of moment selection has to be revised. There are procedures for selecting those moments that contribute to parameter estimation and retain the ones that help and discard those that don't add new information.

– There are alternative considerations to develop a large sample theory and try to use this alternative asymptotic framework to come up with inference procedures:

♦ weak identification – to tackle the case of uninformative moment conditions;

♦ artificial resampling techniques, especially bootstrap – to improve the accuracy of critical points used in tests;

♦ alternative theory of many moment conditions asymptotics for cases when $q >> p$.

– Step outside GMM. The problems arise because of the structure of GMM estimation so propose the generalized empirical likelihood class of estimators which contains the so-called continuously updated GMM and other empirical likelihood-based estimators.

8. Concluding Remarks

The econometrics literature offers the researcher a broad variety of estimation methods differing in the amount of information they use, ranging from fully parameterized likelihood-based techniques to pure nonparametric methods and a rich variety in between. Choosing one appropriately is a respectful task as a correctly specified parametric model provides much better quality estimates than methods that assume little more than mere association between variables at one another. However, this efficiency comes at a cost of possibly false restrictions. From another standpoint, semi- and non

parametric methods are much more robust to variations in the underlying data generating process and still may provide consistent estimates without imposing additional assumptions. We have discussed that GMM is more robust to model specification than ML as it requires less information. This explains the increasing popularity of semi-parametric estimation frameworks like GMM, as they allow to incorporate only as much restriction as economic theory implies.

To state it differently, the GMM estimator is built on more general (assumed) characteristics of the population than in the classical likelihood-based framework, as it requires fewer and weaker assumptions.

References

ANGRIST, J. – KRUEGER, D. [1991]: Does Compulsory School Attendance Affect Schooling and Earnings? *Quarterly Journal of Economics*. Vol. 106. No. 4. pp. 979–1014.

HALL, A. R. [2005]: *Generalized Method of Moments*. Advanced Texts in Econometrics. Oxford University Press. Oxford.

HANSEN, L. P. [1982]: Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*. Vol. 50. No. 4. pp. 1029–1054.

HANSEN, L. P. – HEATON, J. – YARON, A. [1996]: Finite-Sample Properties of Some Alternative GMM Estimators. *Journal of Business and Economic Statistics*. Vol. 14. No. 3. pp. 262–280.

HANSEN, L. P. – SINGLETON, K. J. [1982]: Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *Econometrica*. Vol. 50. No. 5. pp. 1269–1286.

HARRIS, M. N. – MÁTYÁS, L. [2004]: A Comparative Analysis of Different IV and GMM Estimators of Dynamic Panel Data Models. *International Statistical Review*. Rev. 72. No. 3. pp. 397–408.

KYRIAKOULIS, K. [2004]: *Gmm Toolbox for Matlab*. http://www.kostaskyriakoulis.com/gmmgui.html

LIESENFELD, R. – BREITUNG, J. [1999]: Simulation Based Method of Moments. In: *Mátyás, L.* (ed.): *Generalized Method of Moments Estimation*. Themes in Modern Econometrics. Cambridge University Press. Cambridge. pp. 275–300.

MÁTYÁS, L. (ed.) [1999]: *Generalized Method of Moments Estimation*. Themes in Modern Econometrics. Cambridge University Press. Cambridge.

NEYMAN, J. – PEARSON, E. [1928]: On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference. Part 2. *Biometrika*. Vol. 20. No. 3–4. pp. 263–294.

PEARSON, K. [1893]: Asymmetrical Frequency Curves. *Nature*. Vol. 49. Issue 1253. p. 6.

PEARSON, K. [1894]: *Contributions to the Mathematical Theory of Evolution*. Philosophical Transactions of the Royal Society of London. Royal Society of London. London.

PEARSON, K. [1895]: *Contributions to the Mathematical Theory of Evolution II*: *Skew Variation*. Philosophical Transactions of the Royal Society of London. Royal Society of London. London.

PODIVINSKY, J. M. [1999]: Finite Sample Properties of GMM Estimators and Tests. In: *Mátyás, L.* (ed.): *Generalized Method of Moments Estimation*. Themes in Modern Econometrics. Cambridge University Press. Cambridge.

Statistical Yearbook of Agriculture, 2011

The Statistical Yearbook of Agriculture supports the work of foreign users with its bilingual (Hungarian and English) form: it provides comprehensive information about the status of the agricultural sector, shows the share of the sector in the national economy, furthermore, presents data on employment in and production, exports, imports and consumption of agriculture and food industry. It contains detailed data on the personal, technical and financial conditions of agricultural production, on plant cultivation and animal husbandry, on the management of agricultural enterprises, product sales, and prices. The main data of crop and animal production broken down by geographical units are published in a separate chapter. The yearbook also presents the agricultural production in the world as well as in EU27 and other major countries.

HUF 7400

Our publications can be bought
at the HCSO Bookshop
H-1024 Budapest II., Fényes Elek u. 14-18.
Phone: +36-(1) 345-6283; or
can be ordered from the HCSO Information Service
Mailing address: H-1024 Budapest II., Keleti Károly u. 5-7.
Phone: +36-(1) 345-6560 Fax: +36-(1) 345-6788
E-mail: info@ksh.hu
Internet: http://www.ksh.hu

Regional Statistical Yearbook of Hungary, 2011 (with CD-ROM)

The Regional Statistical Yearbook of Hungary is a rich treasury of territorial data. It contains detailed information on NUTS level 1 and NUTS level 2 regions as well as counties. The situation of statistical micro-regions and settlement groups, and the main data on the settlement network are portrayed in the yearbook through the most important social and economic indicators. The CD-ROM contains maps, and Excel tables enable users to make further calculations with the data.

HUF 7500

Our publications can be bought
at the HCSO Bookshop
H-1024 Budapest II., Fényes Elek u. 14-18.
Phone: +36-(1) 345-6283; or
can be ordered from the HCSO Information Service
Mailing address: H-1024 Budapest II., Keleti Károly u. 5-7.
Phone: +36-(1) 345-6560 Fax: +36-(1) 345-6788
E-mail: info@ksh.hu
Internet: http://www.ksh.hu