Small area estimation for rural development statistics in Hungary

Mr. György LENGYEL*
Hungarian Central Statistical Office,
Budapest, Hungary – gyorgy.lengyel@ksh.hu

Abstract

Rural development statistics is one of the most important statistics in Europe. Gross Value Added (GVA) per capita is one of the key indicators of rural development statistics. Territorial level has an extremely high importance in rural development statistics. Most of the territorial units has 1 or more urban centres and their rural neighbourhood. The bigger the territorial unit is, the more urban centres it has. In a bigger territorial unit the inner inequalities are hidden by the indicators. That's why it would be very important to use the indicators at the most detailed level. However, the possibility to calculate many indicators is limited to regional level due to technical, methodological and financial reasons. This paper presents a possible method to estimate GVA per capita at small area level in Hungary. The basis for this estimation is the set of indicators available at settlement level, while this multiple regression model is built at county level (NUTS3). The first step is the selection of potential predictors: Analysing the correlation between the independent variables and the GVA and that of amongst them. The second step is building the model with stepwise method. The third step is comparing the predicted values with the original GVA at county level and making some changes according to the results (fine tuning). The fourth step is evaluating the models by analysing their coefficient of determination, and the mean of absolute deviations and standard deviation of deviations. The final step is the calculation of small area data using the best model. The results will be compared to other important rural development indicators, such as Change of total population and it will be visualized on maps.

Keywords: rural development statistics; GVA per capita; small area estimation.

1. Introduction

The rural development policy is the long-term strategic objective of the contribution to the competitiveness of agriculture, the sustainable management of natural resources and climate action and the balanced territorial development of rural areas. Statistical databases don't always contain the exact information needed for indicators that have been formulated based on policy needs. One of the main problems is the insufficient specification of geographical data. Rural development policy should be analysed at a sufficiently detailed territorial level in order to be able to describe different situations and to assess overall trends across the analysed area. This is obvious for environmental aspects, but it is also necessary for indicators related to diversification and the quality of life in rural areas.

Territorial level has an extremely high importance in rural development statistics. Most of the territorial units has 1 or more urban centres and their rural neighbourhood. The bigger the territorial unit is, the more urban centres it has. In a bigger territorial unit the inner inequalities are hidden by the indicators. That's why it would be very important to use the indicators at the most detailed level. However, the possibility to calculate many indicators is limited to regional level due to technical, methodological and financial reasons.

The lower the territorial breakdown is, the less data is available. On the other hand, several variables are needed to measure and evaluate the procedures and situations on the basis of a more detailed territorial breakdown. The variable, Gross Value Added (GVA) (as a relevant indicator) forms the subject of our investigation. The aim is to give a model – using analysis at country level – of the calculation of GVA at small regional level using available variables.

2. Methodology

GVA was analysed in 11 years (2002-2012) at county level (20), so the number of cases is 220. To be able to build a model, several different variables were collected (see table below). The considerations behind the range of variables were the followings:

- the variable has to be available at small regional level
- the variable has to be available at least on a yearly basis
- theoretically the variable has to have impact on the value of GVA.

SPSS for Windows software was used in the whole process of analyses. The aim was to create a multiple linear regression model to be able to calculate GVA at small regional level.

Two groups of variables were separated according to the projection base:

- variables based on the area of the county;
- variables based on the population of the county.

The very first method to get an overall picture of the dataset we have collected is the descriptive statistics. It was visible from the range of the variables and their relative standard error, that Budapest would mislead us, as it is an outlier from every point of view. At this stage Budapest has been eliminated from the dataset to ensure the consistency. So, the number of cases in the analysed dataset decreased to 209.

If the range of a variable (the difference between the maximum and minimum values) is relatively too small, then it would not be a sufficient predictor, because its value is almost the same in every territorial unit. The same type of situation is occurred considering the relative standard error: if it is too close to 0, then it does not show too much variability. In case of relative standard errors, the relatively too high values are not welcome in case of regression as well, because outliers can be suspected in the background. Hence, based on descriptive statistics, we can skip the variables having very small range and with very small or very high relative standard errors.

The steps of building the multiple regression model:

1. Selection of potential predictors: Comparing the correlation values of the variables to the GVA and to each other.
2. Building the model with stepwise method[1].
3. Comparing the predicted values to the original GVA at county level.
4. According to the results, some items/conditions of the model have to be changed.
5. Back to the first step.

3. Evaluation criteria

The comparison and evaluation of the models has to be carried out from more point of views, combining more aspects and indicators. 3 indicators had been analysed:

- R^2 (Coefficient of determination): it shows, how much the predictors explain of the GVA's development. The closer this value to 1 is, the more we cover the GVA's development.
- Mean of absolute deviations: The GVA values had been estimated according to the models, and the results were compared to their original pairs. The differences are expressed in the percentage of the original values. These percentages show – from one point of view – the conformance of the model. To get an overall picture, and to be able to compare the different models, their mean has to be calculated. However, as the differences can be positive and negative as well, their simple average would be misleading. The mean of their absolute value has to be calculated. The closer this value falls to 0, the better the model we use.

[1] Stepwise method means that at first step one predictor is taken into the model (the most correlating one), then in every step one additional, until the R^2 reaches its maximum. (It is also possible to ignore a variable in the procedure instead of a new variable.)

- Standard deviation of deviations: A regression model should meet the requirements of stability and should not give outlier estimates in any case. So, the stability of the model has to be checked before the final evaluation. Standard deviation of deviations is the proper variable to measure this stability. The smaller this value is, the more stability our model shows.

An additional method to check the reality and consistency of the models is to calculate and analyse the results given at the level of small regions (e.g. negative or impossibly small/high values). Another aspect of analyses is to aggregate the estimated GVA per small regional level into county level, and compare these numbers to the original values of the counties.

4. The model

According to the requirements mentioned above several publicly available datasets were downloaded from the data warehouse of the Hungarian Central Statistical Office. Based on these datasets 28 variables (see table 1.) were created using the population as projection base. Several other were developed using the area as projection base but they were eliminated after the first test. According to the correlation matrix, 17 more variables were eliminated due to their low relationship with the target variable. Additionally, 2 other variables (marked with 'm' in the table) were dropped out because of the high correlation within a 3-variables group.

Severeal regression model were tested with different entering methods supported by SPSS. The difference amongst the results was negligable. This paper contains only the result of the best model using stepwise method. The variables included in the model are marked in table 1.

Table 1. Selected variables

Name of variable	Unit	Selected based on correlation	Included in the model
Total length of stay/Permanent population	Nights/Person		
Length of stay of international tourists / Permanent population	%		
Length of stay of international tourists / Total length of stay	%	X	X
Operating enterprises / Permanent population	Piece/Person	X	X
People having personal income tax / Permanent population	%	X	
Personal income tax base / People having personal income tax	HUF /Person	X	X
Number of passenger cars / Permanent population	Piece/Person	X	
Number of passenger cars registered first time in Hungary / Permanent population	Piece/Person		
Number of telephones / Permanent population	Piece/Person		
Number of dwellings connected to cable television network / Permanent population	Piece/Person	X	X
Total unemployed / Permanent population	%		
Unemployed men / Permanent population	%		
Unemployed women / Permanent population	%		
Total long term unemployed / Total unemployed	%	m	
Number of long term unemployed men / Unemployed men	%	X	X
Number of long term unemployed women / Unemployed women	%	m	
Total school-leaver unemployed / Total unemployed	%		
Number of school-leaver unemployed men / Unemployed men	%		
Number of school-leaver unemployed women / Unemployed women	%		
Net migration / Permanent population	%	X	
Number of general practitioners / Permanent population	Piece/Person		
Number of family paediatricians / Permanent population	Piece/Person		
Number of hospital beds / / Permanent population	Piece/Person		
Number of nursery places / Permanent population	Piece/Person	X	X
Number of pharmacies / Permanent population	Piece/Person		
Number of kindergarten places / Permanent population	Piece/Person		
Number of petrol stations / Permanent population	Piece/Person		
Number of catering units / Permanent population	Piece/Person		
Gross Value Added / Permanent population	Thousand HUF/Person		

4. Results

Based on the possible objective evaluation criteria, the best estimations of GVA per population at county and at small regional level was given by this model of 6 variables. One more correction was implemented in the practical use of the calculation: the results given by the model were adjusted at small regional level, to satisfy the rule that the sum of GVA of small regions gives us the county level GVA value.

The parameters of the model are presented in table 2.

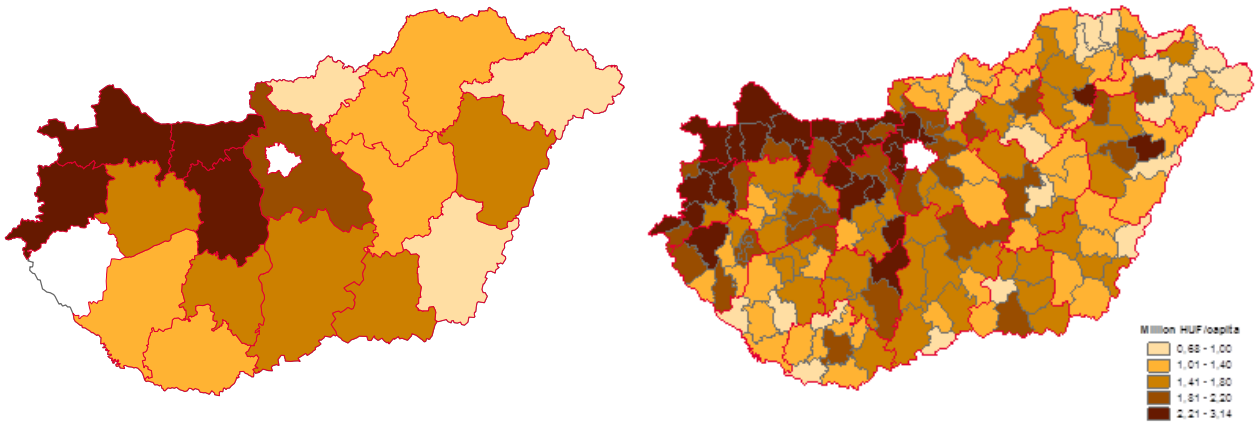
Table 2. Coefficients and statistics of the best model

Variable	Coefficient (B)	t	Significance
Constant	305,033	1,946	0,053
Number of nursery places / Permanent population	39,727	2,527	0,012
Personal income tax base / People having personal income tax	0,967	20,237	0,000
Number of long term unemployed men / Unemployed men	-1963,312	-10,119	0,000
Length of stay of international tourists / Total length of stay	538,361	4,868	0,000
Operating enterprises / Permanent population	5,710	2,909	0,004
Number of dwellings connected to cable television network / Permanent population	0,683	2,349	0,020

R^2 (Coefficient of determination)	0,859
Mean of absolute deviations	0,073
Standard deviation of deviations	0,097
Mean of absolute deviation of small regional aggregates at county level	7,2%

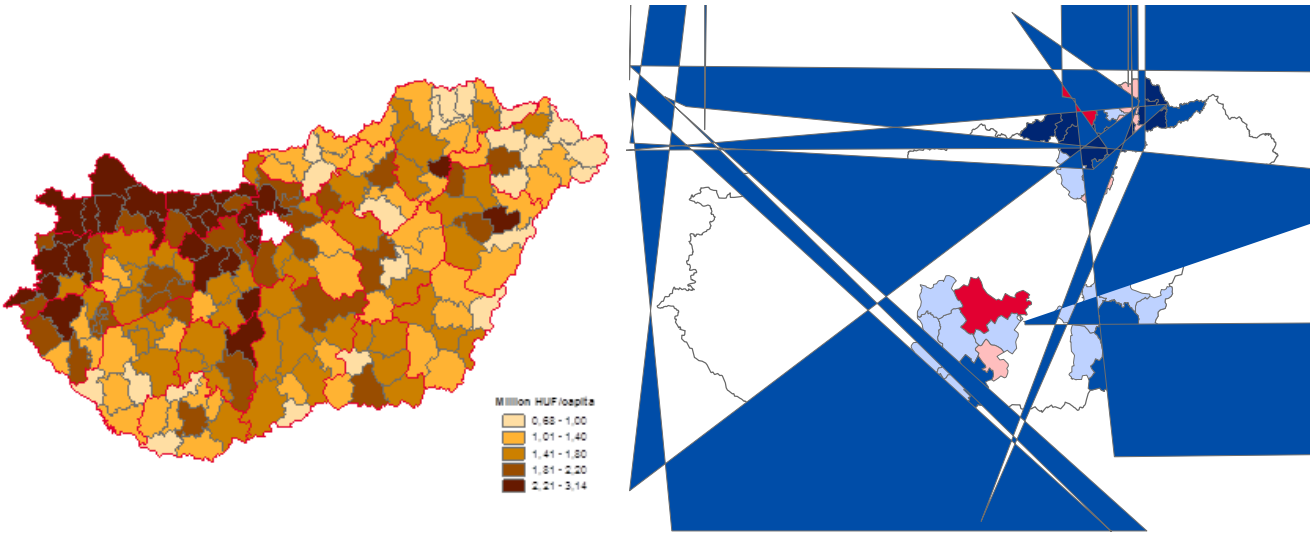
Based on the results some maps are designed to demonstrate the significance of the lower level data. On the first pair of graphs it is clearly visible that the higher territorial breakdown (NUTS3) does not form homogeneous group from the point of view of GVA. In general, we can say that the North-West part of the country, and the neighbourhood of the capital possess the highest values. From other point of view, the bigger cities, especially county capitals show the highest values, while their surroundings are less productive in most of the cases. This is why such an indicator has to be analysed at small regional level: to be able to compare the real situation in the different parts of the country.

Graph 1. Gross Value Added per capita in Hungary, county and small area level



The result of the estimation is compared to the indicator ‘Change of total population’ and the comparison is presented on graph 2.

Graph 2. Gross Value Added per capita and Change of total population at small area level



The maps show that immigration is significant mostly in those areas where the per capita GVA is higher.

5. Conclusions

With this small investigation of possible models for GVA at more detailed territorial breakdown, the aim was to give an idea of such a method in theory and in practice and to present that good results can be achieved statistically.

In case of a highly complex variable as gross value added, the modelling has to be worked out continuously in every year, analysing the predictors with the help of longer or updated time series. Fine tuning can be managed even on advanced statistical level, e.g. with the use of non-linear models.

References