ADAPTATION OF STATISTICAL MATCHING IN MICRO-REGIONAL ANALYSIS OF AGRICULTURAL PRODUCTION

PESTI CS., KÁPOSZTA J.*
Department of Farm Business Analysis, Research Institute for Agricultural Economics,
1093 Budapest, Zsil u. 3-5.
**Institute of Regional Economics and Rural Development, Szent István University,*
2103 Gödöllő, Páter K. u. 1.

KEYWORDS: micro-regional, statistical matching, administrative data, FADN

RUNNING HEAD: STATISTICAL MATCHING IN MICRO-REGIONAL ANALYSIS

ABSTRACT

Agricultural production and agricultural policy has many special dimensions. The production structure, income positions, and labour input have large regional disparities, the production intensity is not homogenous in space, and farms have different risk factors and market possibilities in the different regions of Hungary. Land use and production technology also varies largely, in many regions farming is competitive, highly specialized with big corporate farms, while other regions have small individual farms with mixed production structure and less concentration in land use.

There are no direct data for spatial analysis less aggregated than NUTS 3 level. Only the data of agricultural census and administrative database for direct payments are available at settlement and micro-regional level, but these databases do not provide information of farm income. The income statistics either cannot be disaggregated to micro-regional level (agricultural accounts) or are not representative at this level (FADN).

The administrative database of the Paying Agency contains the land use data and limited livestock numbers for all farms receiving direct payments. The FADN database contains a large accountancy dataset for a low number of farms. Statistical matching combines these two databases and provides a possibility for detailed regional analysis using estimated data.

INTRODUCTION

Agricultural production and agricultural policy has many special dimensions. The production structure, income positions, and labour input have large regional disparities, the production intensity is not homogenous in space, and farms have different risk factors and market possibilities in the different regions of Hungary. Land use and production technology also varies largely, in many regions farming is competitive, highly specialized with big corporate farms, while other regions have small individual farms with mixed production structure and less concentration in land use. The role of agriculture in local economy is also different, in many parts of the country agriculture has still a high proportion both in employment and in self-sufficiency of low income people, while in other parts of the country the role of agriculture is negligible compared to industrial and service sectors.

The spatial location and development of agricultural production is influenced by many factors. These factors can be divided into two groups: one depends on natural resources; the other depends on social and economic endowments. The first group involves the relief, climate, soil, land use; the second group involves the historical traditions of production, ownership, labour and capital input, machinery, spatially differentiated demand for crops and location of markets. One of important task of agricultural policy is to develop a production structure that adapts to natural resources and is capable to correspond to the common requirements of competitiveness, landscape management, soil protection and employment.

Regional analysis of agricultural production helps to find the best areas for intensive, competitive production systems with higher input proportion and what are the areas where the production needs to be diversified with extensive livestock farming and biomass energy production.

However, there are no direct data for spatial analysis less aggregated than NUTS 3 level. Only the data of agricultural census and administrative database for direct payments are available at settlement and micro-regional level, but these databases do not provide information of farm income. The income statistics either cannot be disaggregated to micro-regional level (agricultural accounts) or are not representative at this level (FADN).

One solution can be linking these databases. Vrolijk et. al. (2005) uses statistical matching for estimation of dairy farms in a municipality in the northern part of the Netherlands. They make an estimate for total revenues, total costs, net farm results, labour income. (These are the goal variables.) With data imputation it is possible to use the extra information from dairy farms in the larger region to make an estimation of the results of dairy farms in the specific municipality. The structural data for land use and livestock numbers of dairy farms in the municipality are provided from the census (population). The income data is provided by the dairy farms in FADN sample in the larger region (sample). In the estimation procedure a number of imputation variables are used:

- age;
- hectares grassland;
- hectares fodder crops;
- number of dairy cows;
- economic size (European Size Unit, ESU).

The imputation variables need to have a logical relationship to the goal variables. For each farm in the population of the municipality the three most similar farms in the FADN sample in the larger region are selected. They randomly choose one from the three farms and summarize the results for all farms in the municipality. After repeating the imputation several times, they check the standard error of statistical matching. They increase the confidentiality of imputation by making the estimation for the sample farms and show the differences between the real and estimated values.

The researchers of Teagasc University in Irelend develop a static spatial micro-simulation model. Kelly (2004) works out a method for re-weighting FADN sample farms to micro-regional census data. He calculates weight numbers for the farms in a micro-region so that the farm micro-data multiplied by the weights numbers cover the total utilized agricultural area and livestock numbers of the micro-region.

The method is further developed by Hynes et. al. (2006), whereas a static micro-simulation model is worked out, which covers the entire agricultural production in Ireland. The two most important data input for the model are the Irish National Farm Survey (FADN farms) and the Irish Census of Agriculture.

In general there are 1,200 farms in the National Farm Survey each year. The survey contains the farms' production, labour and income data but it is representative only at the national level, so directly it is not eligible for regional analysis.

The 2000 Census of Agriculture identified every operational farm, ca. 140.000. However, the register contains about 190,000 farms, it was expected that there would be only about 140,000 active farms. The data in census can be linked to municipality, but provide information only from land use, livestock numbers and labour.

The model combines the two databases using statistical matching techniques. For each farm in census the model finds the most similar farm in the FADN sample, for which all income and financing data are available. The confidentiality of the model is determined by the real similarity of the census farm and the chosen FADN sample farm. The model uses the common variables (imputation variables) for statistical matching which apply to farm size, farm structure and soil type. The common variables vary according to farm types. The model finds the best fitting imputation variables, so it needs high computing capacity, running the model for the whole Irish agriculture takes two days.

The micro-simulation model gives very detailed information of the income positions of agriculture, it can be disaggregated to municipality level, so it is not only interesting from the point of methodology, but its results can be used by decision makers.

MATERIALS AND METHODS

There are no the cost, income and labour input data in Hungary of all farms and their enterprises. Due to the lack of primary data, the output, income, cost and labour data of all farms *(population)* can be estimated on the basis of smaller but more detailed databases *(sample)*. Similar to Dorgai et. al. (2008), *statistical matching* techniques were used for this estimation, which is showed below in details. The concept *matching* means that for each element of the *population* an element of the *sample* is selected on the basis of similarity of the *common variables*. So the variables, only available in the sample can be estimated for each elements of the population (Figure 1.).

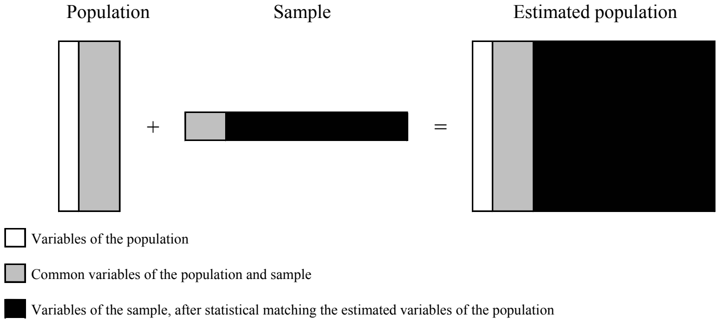
Population
Sample
Estimated population
+
=
Variables of the population
Common variables of the population and sample
Variables of the sample, after statistical matching the estimated variables of the population

Figure 1. Diagram of statistical matching

The following mathematical aspects are to be taken into consideration in choosing the common variables for statistical matching on the basis of Vrojlik (2004):

- The common variables need to be in *strong correlation* with the estimated variables for the population (only present in the sample). For example, if economic size (or utilized agricultural area) is used for estimating profit before tax, the estimation is only confidential, if there is a strong correlation between the two variables.
- The *distribution of the common variables* need to be similar in the sample and the population, otherwise the estimation is less reliable. So for each elements in the population there need to be at least on similar element in the sample. If there are extreme values (outliers) in the sample and/or in the population, they may be excluded from the analysis.
- Due to the mathematical model of matching it is no use adapting *more than five common variables* for estimation.
- Usually there are *no general solutions* for estimation, the common variables must be chosen according to the aim of the analysis.

The statistical matching selects for each element of the population that element of the sample, where the summarized distance of the common variables is minimal. This method was applied with success in agricultural economic studies with spatial focus (Dol, 1991; Vrolijk, 2004; Hynes et. al. 2006). The difference in methodology is the method for measuring the distance. The authors of this study applied the Euclidean distance calculation on the basis of Vrolijk (2004).

The distance function calculates the distance between a sample farm and a population farm:

$$D_{j,k}=\sum_{i=1}^{m} a_i \; S_{j,i}-S_{k,i}{}^{\beta_i}$$

in which:

$D_{j,k}$	Distance between sample unit j and population unit k
a_i	Weight constant of variable i
$S_{j,i}$	Normalised score of sample unit j on variable i
$S_{k,i}$	Normalised score of population unit k on variable i
j,k	Unit identifier
ß_i	Exponent of variable i
i	Variable identifier
m	Number of variables

The statistical matching procedure orders to each element of the population that element of the sample where the distance function takes its minimum value.

To find the appropriate databases the aim needs to be clarified and the variables for achieving the aim need to be defined. The aim is in our case estimating field crop farms' income at micro-regional level, the variables for this are the gross margins of the farms' enterprises.

Matching the Hungarian Census of Agriculture and FADN farm level data

Theoretically this matching would be the best for estimating the cost and income variables of the universe of farms, because the census contains the total production structure and labour use of the farms. Labour input is important because it refers to the farm's technological level. If a farm with a specific production structure uses less labour, it is assumed to have better production technology than other farms with the same structure using more labour. So the statistical matching model selects for each farm in the census the best fitting FADN sample farm. The common variables are the livestock numbers, the hectares of crops and the labour utilization, varying according to the types of farming. The census has more potential common variables than the administrative database, so the results would be more reliable.

Thus the practice does not allow using the census for population in statistical matching. The reason is the specific Hungarian methodology of the census. The individual farms are surveyed in micro-regions. In each micro-region all individual farms are surveyed, but the micro-regions cover only a part of Hungary. Therefore the census uses weight numbers for individual farms and census data is only representative at national and NUTS 2 level. Comparing the total arable land in micro-regions from census and from the administrative database shows large differences. So census is not representative at micro-regional or municipality level, but in the future there may be a possibility of utilizing the census for such purposes, supposing the improvement of data collection.

Matching the administrative database and FADN enterprise level data

This type of matching is capable of estimating the income of the enterprises of farms in the administrative database. The sum of the income of a farm's enterprises gives a *cal-*

culative farm income. The administrative database does not contain the total production structure of the farms, but only the enterprises receiving direct payments. These are crop production, milk production, sheep husbandry and bull fattening.

The statistical matching in this case means that for each enterprise of the farms in the administrative database the best fitting enterprise of a sample farm (with similar size and structure) was selected. The analysis was made for the field crop enterprises (10 different enterprises). The common variables were:

- Legal status (individual farm/corporate farm);
- Hectares of arable land;
- Average soil quality of arable land (in administrative database average soil quality in the municipality)
- Hectares of field crops (wheat, corn, barley, rye, oat, triticale, sunflower, rape, silage corn, alfalfa);

There are ca. 200,000 registered farms receiving direct payments, while FADN represents ca. 90,000 holdings exceeding 2 ESU. It is well known that a part of registered farms do not have agricultural activity, they only register for subsidies and another farm is cultivating their area. Besides, at least half of the registered holdings are below 2 ESU, so they could not be sample farms. Therefore the matching can be only confidential if we assume that there is practically no difference in the income of farms below 2 ESU and farms between 2 and 4 ESU.

Although data from 2007 is already available, 2006 data of the administrative database and FADN was used in the procedure. The reason was that 2007 was an irregular year, because draught was very different even at areas close to each other.

The statistical matching was computed with Statistics for Regional Studies (STARS) developed by LEI in the Netherlands and MySQL database system.

RESULTS

Gross margins of field crop enterprises of all registered farms were estimated by statistical matching. Figure 1 shows the gross margins of field crop production per hectare, summarized by farms and micro-regions.

Field crop production is the most profitable in South-Transdanubia, and in Békés and Hajdú-Bihar counties. On most of these areas the climate and soil favours for field crop production, and the farm structure and agricultural traditions presume high income. Production is the least profitable in the mountains, in the south-west areas of the Hungarian flatland and in the middle areas of the Tisza river. The reasons of the low incomes are mostly the less favoured natural resources. In the mountain areas erosion and thin soil layer, in the south-west flatland sandy soil with bad water capacity, in middle areas of Tisza salinization and flood reduce incomes.

Figure 2 shows land rent prices and Figure 3 shows soil quality.

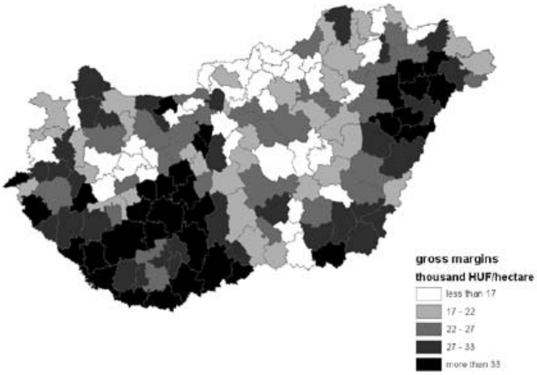

Figure 1: Gross margins of field crop production in the Hungarian micro-regions, 2006

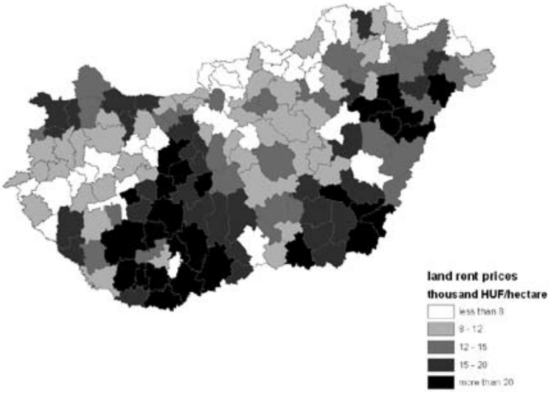

Figure 2: Land rent prices in the Hungarian micro-regions, 2004-2006 (on 2006 prices)

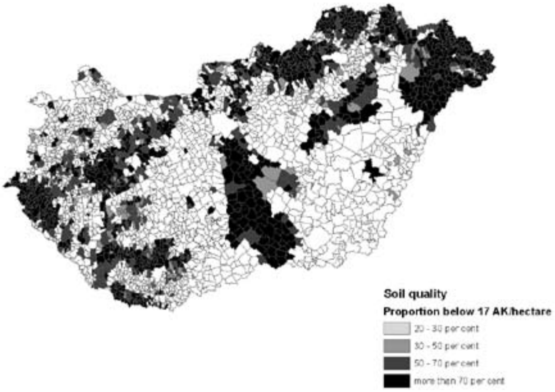

Figure 3: Soil quality in the Hungarian micro-regions

Comparing the two figures with the map of gross margins (Figure 1) shows that natural resources largely influence the income of field crop production and the land rent prices follow the potential income of production.

DISCUSSION

Statistical matching can be successfully applied for regional analysis of agricultural production. It is possible to analyse and forecast the financial and income positions of farms in the census or in the administrative database by spatial micro-simulation models.

Analysing the gross margins of field crop production supplemented by risk management and the analysis of production structure reveals that there is a need for investments improving yield and income safety (irrigation, water management) on large parts of the Hungarian flatland. While on mountainous and other less favoured areas extensive farming should be supported, like grazing livestock (sheep and bovine).

In the authors' opinion statistical matching can be utilized in regional analysis of agriculture for solving practical problems. An example is estimating the straw yields and income of farms in the neighbourhood of a potential biomass power station or mapping the potential suppliers near by a fruit processing plant. In this case statistical matching can be fine-tuned paying attention to local aspects. The selection of common variables, the method of distance calculation and the controlling procedure can be optimized to the specific problem.

REFERENCES

DORGAI L., LUDVIG K., MÁRKUSZ P., MOLNÁR A., PESTI CS., SZÉKELY E., TÓTH E., UDOVECZ G. (2008): The social, economic and environmental impacts of assumed decreasing of direct payments (First Review). Research Institute for Agricultural Economics, Agricultural Economics Studies, 137 pp.

HYNES S., MORRISSEY K., O'DONOGHUE C. (2006): Building a Static Farm Level Microsimulation Model: Statistically Matching the Irish National Farm Survey to the Irish Census of Agriculture. 46th Congress of the European Regional Science Association, http://www.ersa.org/ersaconfs/ersa06/papers/431.pdf

KELLY D. (2004): SMILE Static Simulator Software User Manual. Teagasc: Teagasc Athenry Publication.

VROLIJK H., DOL W., KUHLMAN T. (2005): Integration of small area estimation and mapping techniques. Tool for Regional Studies. LEI, The Hague 60 pp.

VROLIJK H. (2004): STARS: statistics for regional studies. Proceedings of Pacioli 11: New roads for farm accounting and FADN. Report 8.04.01. LEI, The Hague