RECREATING LOCATION FROM NON-SPATIAL DATA –SAMPLE SIZE REQUIREMENTS TO REPRODUCE THE LOCATIONS OF FARMS IN THE EUROPEAN FARM ACCOUNTANCY DATA NETWORK

Martin Damgaard
IAMO (Leibniz-Institute of Agricultural Development in Central and Eastern Europe), Germany
Email: Damgaard@iamo.de

Abstract

Individual farm accountancy data sources such as the European Farm Accountancy Data Network (FADN) include no specific information on the spatial location of farms. However, spatial characteristics and site conditions determine the farms' production potential and its influence on the surrounding environment. Spatially explicit models that make use of the FADN data need to be able to recreate a landscape including the location of the farms in a plausible way.
This paper investigates the minimum sample size of farm locations required to insure the ability to reproduce a reliable map of a given region. This is done by analysing relative locations between all the 1871 farms present in the Danish river Gudenå watershed. As we have detailed information about each of the farms we can categorize the farms in groups in a way similar to what one would be able to do with farms from a FADN sample. By utilising the rich information that the FADN sample contains to create a multidimensional spatial set of requirements that the farms on average have to meet it is possible to reduce the number of available locations to a minimum. This investigation is divided into the following two-step procedure: First the variability of an individual farms spatial relationship is investigated with regard to variation in sample size and composition. Secondly is the average values investigated with regard to variation in sample size and composition.

Keywords: FADN, spatial location, methodology

Introduction

To recreate a reliable representation of the complex reality is one of the fundamental challenges in creating empirically founded models. Numerous models are based on abstract representations of the underlying system and do not need the empirical foundation for investigating the characteristics of the object of study. However once the findings from the models are used for policy recommendations, realistic and empirical founded models are preferred.

Obtaining sufficiently empirical data for large regional models through personal field studies are seldom possible. Most models are instead relying on available data from databases or other collectively gathered information. The accuracy of these data differs however. Many of the most adequate economic data are collected by the local authorities indirectly through the assessment of taxes or similar administrative issues. This means however, that the most reliable data are at times restricted to insure personal privacy. The European Farm Accountancy Data Network (FADN) is one of these large but restricted data collections.

Every year a large sample of farm accounts is collected in each of the member states in the European Union. From this base sample a number of so-called "representative" farms are found. Each with an extrapolation factor constructed in such a way that the farms provide a representative sample for the commercial farms in a given region. The extrapolation factor incorporates the regional characteristics, the economic size and type of farming found in the whole collection. The term "representative" as well as the

accuracy of the methodology is up to debate within the scientific community (Beer et al. 2001; Meier, 2004; Meier, 2005) however will not be questioned here.

The sensitive nature of the micro-economic data within the FADN sample means, that the data comes with no other specific geographical reference than which region/ country the collective sample represents. However the spatial nature of agricultural production means that both the farms' production potential as well as its impact on the surrounding environment makes it vital for a potential modelling application based on FADN-data to recreate the plausible spatial locations of the farms in the sample.

A few attempts based on indirect statistics have previously been published (Fais et al., 2005; Fais & Nino, 2004). One of the most ambitious attempts is undoubtedly the work done by the Seamless project (Elbersen et al. 2006). The methodology developed here is also making use of statistics and remotely sensed data. However the restricted nature of the FADN data sample makes it difficult to validate the findings. The present analysis is therefore taking a novel approach. Rather than working directly with the FADN-sample and thereby not knowing the underlying reality that the sample describes, this study is using a sample of 1871 farms located in the Danish watershed to river Gudenå. Both the exact location as well as production data for all the 1871 individual farms are known with similar categories as offered in the FADN sample, with the exception of the economic data present in the FADN sample.

Throughout the rest of this paper we are assuming that the "representative" farms found in the FADN sample and their extrapolation factors create a perfectly fitting description of the 1871 farms found in the river Gudenå watershed. Although this assumption is rather unrealistic it is similar to the normal confidence one has to have in the FADN sample, when no other information is available. This perfect sample consists of the production data in our database of the 1871 farms, with the exception of the geographical references.

Our task is to investigate the sample size of farm locations required to insure the ability to reproduce a reliable map of the region.

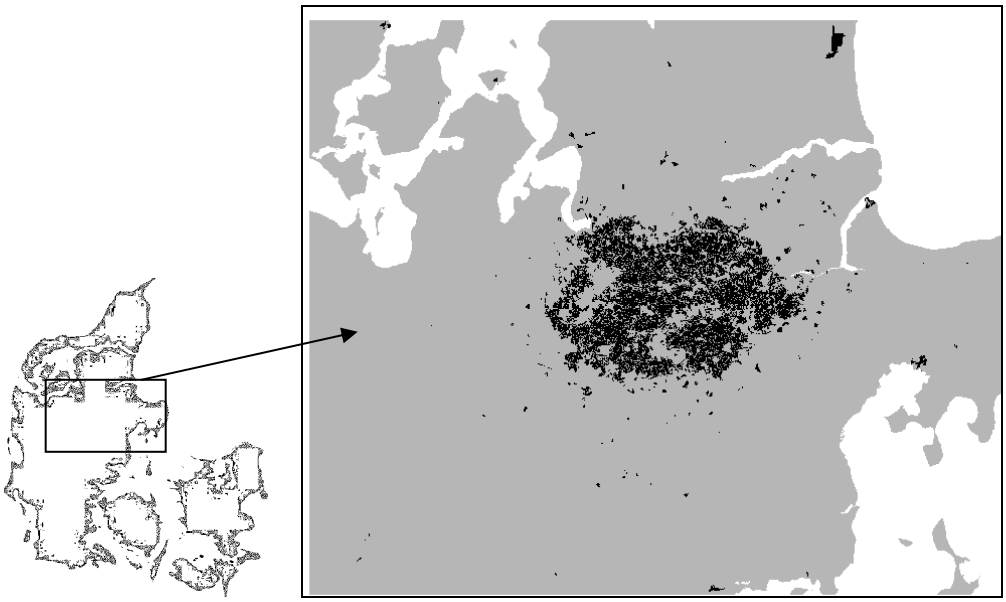
This is done by analysing the relative location between all the 1871 farms present in the Danish river Gudenå watershed. As we have detailed information about each of the farms we can categorise the farms in groups similar to what one would be able to do with farms from a FADN sample. By utilising the rich information that the FADN sample contains to create a multidimensional spatial set of requirements (such as the distance to the nearest dairy farm or to the 2nd nearest farm between 0 ha. and 20 ha.) that the farms on average have to achieve it is possible to reduce available locations down to a minimum. In this case we utilize data of the farms size, production type and number of animals units.

The rest of the paper is structured as follows: in section 2 an introduction to the study area as well as introduction to the empirical data is given. The difficulties in relying only on remotely sensed data and landscape characteristics, such as topographical maps, soil maps and road system maps for this particular case are presented. In section 3 the outline of the analysis is presented. In section 4 the results are presented and in the following section further applications as well as the difficulties in constructing the maps practically is shortly discussed. Finally a conclusion is made.

Introduction to the study area and the empirical data

The valleys of "Nørreå" and "Gudenå" are located in the central part of Jutland between three major cities: Aarhus, Viborg and Randers. The area covers over 76600 ha. 1871 farms on 72089 ha of arable land and 5089 ha of grassland on an average size of 41 ha are for most of them (62%) performing field crop farming. The other farms are then quite equally distributed among dairy farming (11%), grazing livestock farming (6%), granivores (14%) and mixed farming (7%).

Figure 1. Map of the study area. The dark marked area shows the involved farms and their fields. All fields belonging to farmers in the area are included even if the location of the field is outside the watershed. Note that due to Danish area requirements fields very far from the farmstead can still be favourably owned.



The study area was chosen partly due to the data availability and partly due to the landscape characteristics. In contrast to a large number of other areas is this region lacking strong spatial indicators by which the available space for farm locations could be deduced. This becomes apparent when one compares the Danish river Gudenå watershed region with other regions where the landscape characteristics can help in locating the farms through e.g. the topography.

The outline of the analysis

The location of a farm in space can be defined as an individual event independently of all other farms or structures in the vicinity. However such an analytical framework would not only reduce the historical process in creating the present agricultural structure out of the empirical data it would at the same time also reduce a large part of the knowledge we have of the present farms.
Even the freedom of action of the present farms will to a varying degree be determined by its history to its actual state as well as by the actions and history of other agents in the area. So although it would be reckless to claim that the location of a given farm will directly tell us much about the neighbouring farms it can still reveal some elements of an indirect relationship between the farms. Often local experts will be able to locate a given farm type to a small part of the region simply because farms are not randomly distributed in space but tend to cluster around certain areas. This means that we should be able to utilise this information, when we are going to recreate the distribution of farms in a given region. In the case of FADN farms however the difficulty is that we always start off with a sample seldom know what characterizes this particular selection. Therefore this investigation is conducted in such a way, that influence of both the sample size as well as the composition of the sample is analysed.
The incomplete knowledge one has in working with FADN data makes some kind of up-scaling or extrapolation of the location of all farms from the initial sample unavoidable.
We will here make use of a similar framework of thought as used in resampling techniques, such as jackknife or bootstrap as we investigate the possible level of error such extrapolations could lead to. At the same time will we utilise the rich information that the FADN sample contains to create a multidimensional spatial set of requirements that the farms on average have to achieve and thereby

exploit the possibility to reduce the available farm locations down to a minimum. The procedure will therefore draw upon interrelationships between the farms rather than the spatial characteristics of the individual farm. It is often beneficial to include such spatial characteristics. For reasons of simplification will these characteristics not be included in the following.

Here only the interrelationship between the farms spatial location is utilized as the location of the farmstead is viewed as a network of interrelated points in space. The network is represented as graphs that consist of a set of *vertices* (or nodes) connected by a set of *edges* (links). Here the vertices represent the farmsteads, and the links between the points represents the Euclidean distance between those farms. Each farmstead holds information about the farm size and the production system. This information is used to categorize a given farms relationship to the 1870 other farms (such as the 2nd nearest dairy farm or the nearest farm between 51 ha. and 100 ha.). Therefore the edges are directed lines, as the interrelationship is not symmetrical. This means that the investigated network consist of 3498770 (or 1871*1870) links.

The investigation of the network is divided into the following two-step procedure: First the variability of an individual farms spatial relationship is investigated with regard to variation in sample size and composition. Secondly is the average values investigated with regard to variation in sample size and composition.

To understand the chosen procedure it is important to remember that our enterprise is to investigate the minimum sample size of farm locations required to reproduce a reliable map of a given region. Therefore we will mimic the situation, where one is collecting data in the field by varying the sample size and this has been repeated with different order of the farms at least ten times. The latter is done as we can't be certain in what order the farms are chosen if one is collecting the data in the field. Though the number of different selections of farms from a combinatory point of view hardly scratches in the surface of possible orderings, the sample size will still provide us with some insights into the variation one normally will encounter.

The Analysis

We look at an individual farm by investigating the variability in the statistical properties in its relative location to all the other farms due to sample size and composition. This is done by taking approximately 10% of all the farms and for each of these farms calculating the Euclidean distance to all the 1871 farms in the region. For each of the 188 selected farms has the most commonly used descriptive statistics (including: mean, median, standard error, 95% confidence level, standard deviation) been calculated for sample sizes varying from 11 farms (the selected farm and 10 additional farms) and up to all the 1871 farms. This is done with an interval of 10 farms. In addition it is done for 11 different successions of the farms. The values are calculated based on the distance and no further categories have been made. The reliability of the values for each individual farm can be assessed through this calculation. This is important as the further analysis eliminate the uncertainty each individual farm constitutes in an incomplete sample. This uncertainty will however unavoidably be included in a sample solely building upon FADN data.

In figure 2. a plot of the relative deviation of the mean as a function of the sample size is presented. In figure 4 a plot of the relative deviation of the median as a function of the sample size presented. In the case of the relative deviation of the mean are the first plot supplemented by an additional plot (figure 3.) of the frequency by which the different relative deviations occur. Please note the scale of the frequency plot, as the scales are not made with equal intervals.

Figure 2. The relative deviation of the mean as a function of the sample size. Own calculations

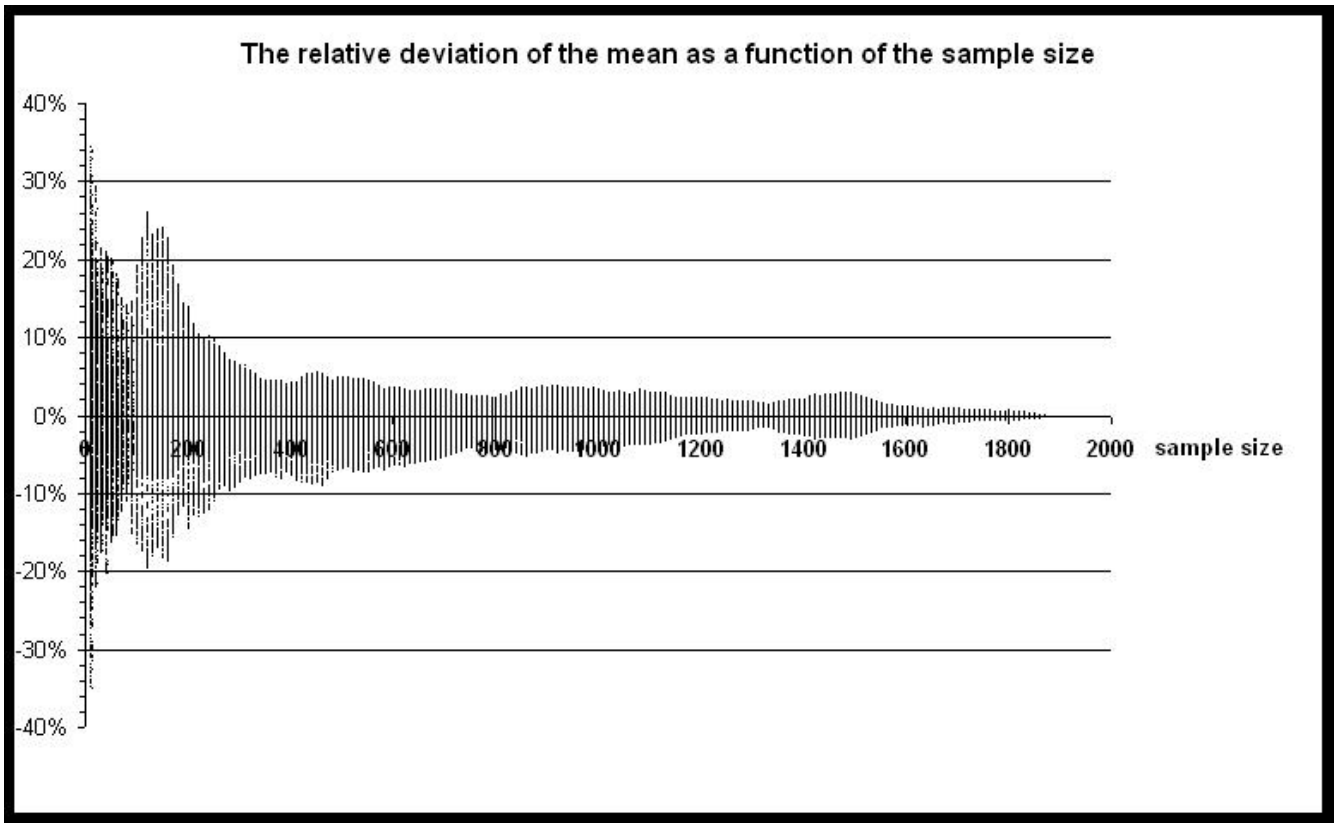


Figure 3. The frequency of the values of relative deviation of the mean. Own calculations.

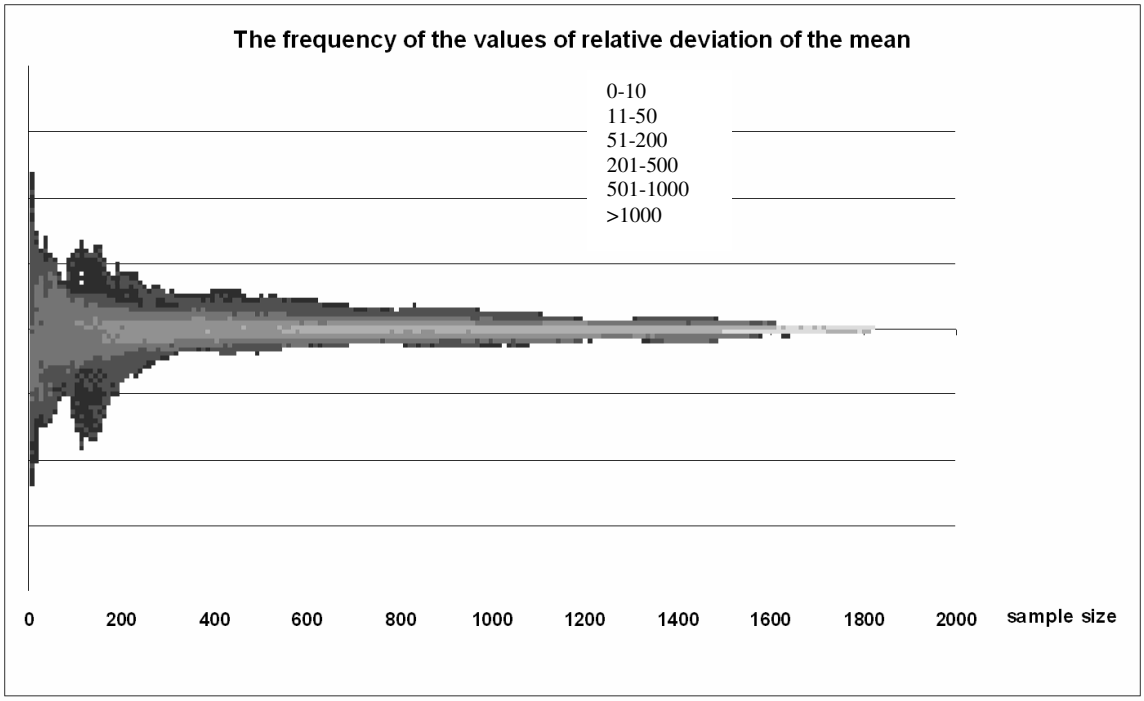
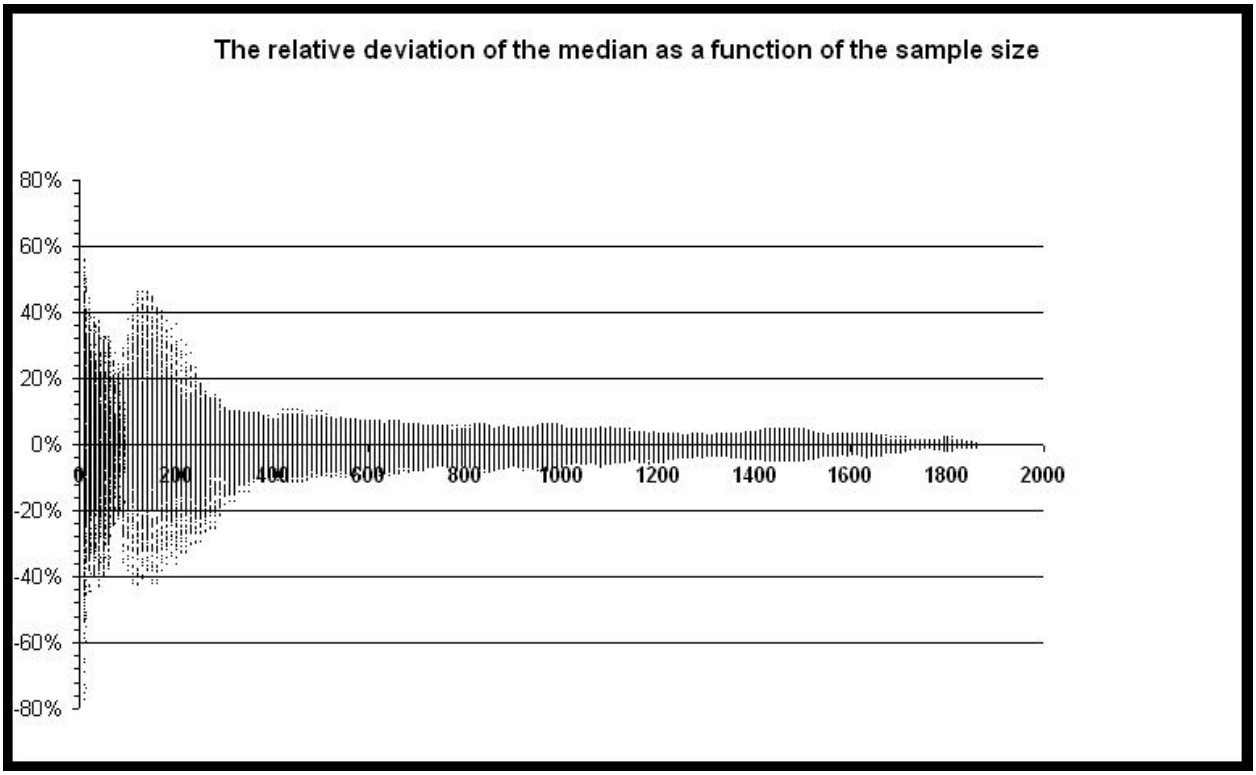


Figure 4: The relative deviation of the median as a function of the sample size. Own calculations.



Please note the difference in scales used in the plot for the relative deviation of the mean and the plot of the relative deviation of the median. Looking at figures 2-4 one can see that once the sample size is around 20% of the full sample (≈400 in this case) the individual farm values are generally reliable. Even earlier are the majority of values within a 10% span in the case of the mean. The median values are naturally fluctuating within a larger span, however otherwise show similar structural characteristic. When working with samples of less than 10% of all the farms the fluctuations within both the mean and median values are so large, that one hardly can trust ones findings to any significant degree.

The first part of the investigation has shown the reliability of the values for each individual farm, while varying the sample size and composition. In the real world this variability would be a part of the uncertainty entering into the average values now to be investigated. Here it would however only blur our findings. The entire network is therefore used in the second part of the investigation. Each of the 1871 farms knows now the Euclidean distance to all others. That means that the distance each individual farm contributes with is founded on perfect information.

Table 1: List of categories used in this study

Categories
All farms
0-20ha farms
21-50ha farms
51-100ha farms
101-200ha farms
More than 200ha farms
Plant production farms
1-50 animal unities
More than 50 animal unities
Pork
Dairy

The variations in the average values are only due to the size and composition of the selected sample. The further procedure is making use of the included production related data. As we know the production category for both the farm working as our point of reference as well as all the other farms we have created a 2-D matrix with the categories seen in table 1 on each side. In the case of the point of reference only the categories that the particular farm fulfils are in use. For all the other farms the scheme is expanded by the subcategories distance to the 1^{st} , 2^{nd}, 3rd , 4^{th} and 5^{th} nearest farm of the category as well as the average distance.

Below are two examples (figure 5-6 and 7-8) of what the ten different successions of farms produce. The two chosen examples are the distance to the nearest farm (figure 5-6) and the average distance to all other farms (figure 7-8).

In figure 5 and 7 are the nominal values presented. The percentile deviation from the full sample are presented in figure 6 and 8. The examples reveal mainly two general characteristics. First of all one can see the modifications that the selections produce. Secondly and more importantly is that the precision of course depend upon the number of farms falling into a given category. Only a fraction of the farms will influence the value of the nearest farm, where as all other farms will affect the average value. This simple fact makes a large number of the possible categories, one can make for a given region, questionable for the purpose considered here. If only a few farms fall into a given category the fluctuations for this group will simply be too large for one to rely on the results. However instead of dismissing such findings altogether the different categories should be supplemented with a weight factor expressing the reliability. Such a weight factor can of course only be an estimate and may be based on studies similar to this one.

Figure 5: The deviation of the distance to the nearest farm of all other farms for ten different successions of farms.

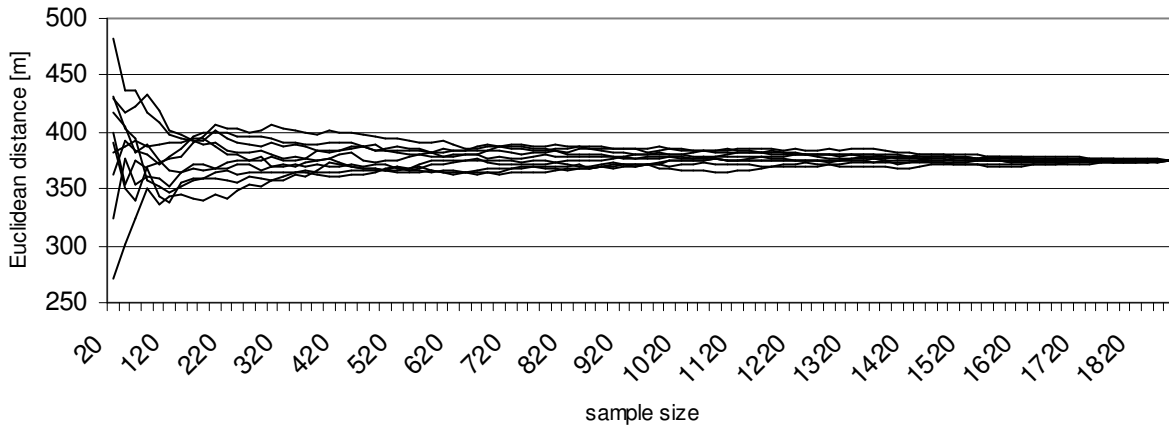


Figure 6: The relative deviation of the distance to the nearest farm of all other farms for ten different successions of farms.

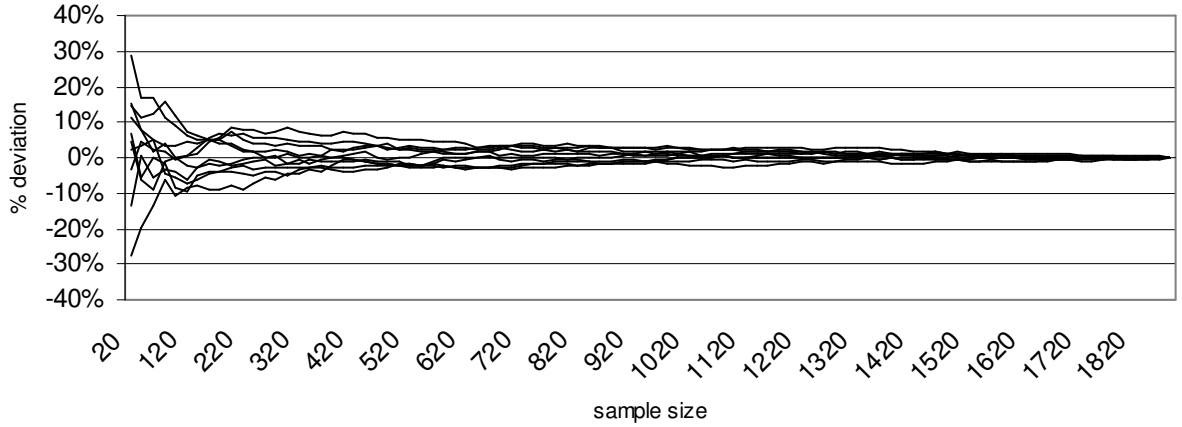


Figure 7: The deviation of the average distance to all other farms for ten different successions of farms.

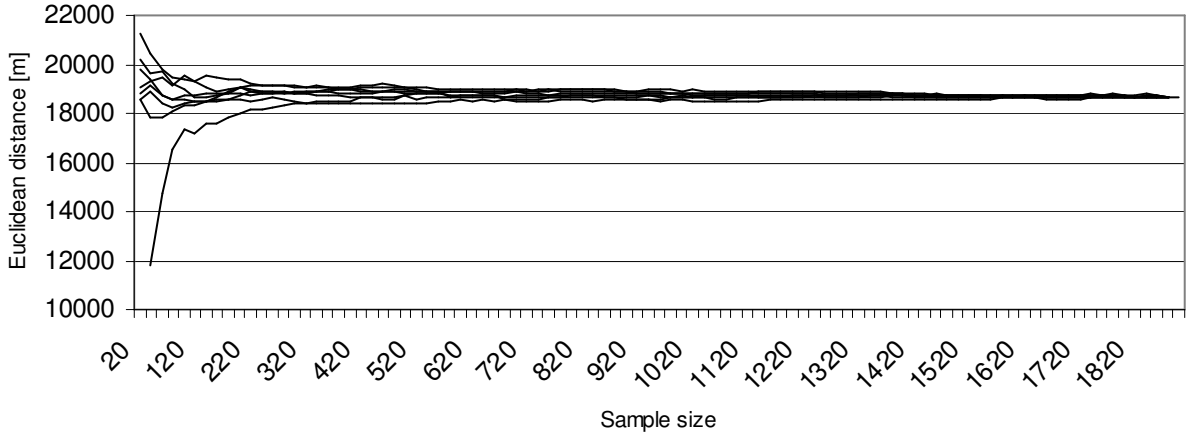
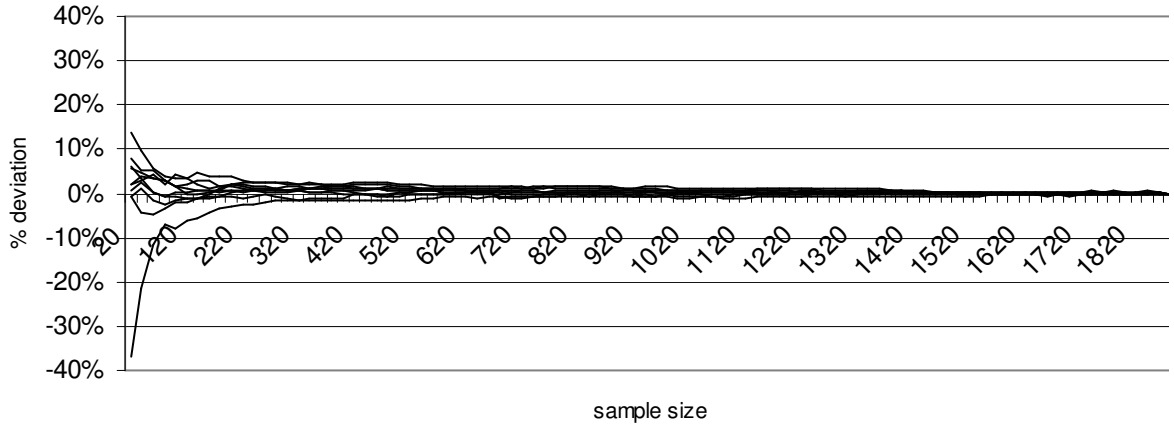


Figure 8. The relative deviation of the average distance to all other farms for ten different successions of farms.



As one can see from the above plots the different ordering of farms will fluctuate around the values for the complete region with a spread that diminish with the larger sample size. This spread have we used to pass on the reliability of a number of the different categories (presented in table 1) and results are shown in appendix 1. For the category "All Farms" as well as the five field size categories are the relative difference between the maximum and minimum values presented for the sample size 20, 100, 400 and 1000. This is done as a function of the average value for all the 11 categories used in this study.

From the shown values in appendix 1. one can see that the size of the fluctuations to a far larger degree depend on the farms chosen as the point of reference than the different categories, that the rest of the farms are categorized under. Because the differences between the tables are much larger than the deviations between the categories. Once more this is due to the number of the individual farms that fulfils a given type description. This is apparent when the values of the most common groups are compared with the less common groups, such as the 24,97% spread for the sample size 20 for 21-50 ha farms against the category "All farms" where as for the group >200 ha is the value 466,76% for the same. At the same time can one also see that some groupings such as the group 0-20 ha farms and 21-50 ha farms produce better results than the "All farms" group. This demonstrates that some of the sub-groupings one can make of the FADN sample can actually reveal better insights to the spatial distribution of the farms in a region than using only averaged considerations.

Discussion

Neither the presented method nor the more commonly used methods based on indirect statistics and remotely sensed data will ever be able to recreate a 100% accurate location of the farms in a region as long as "representative" farms from the FADN sample are used. The challenge is to find the most reliable method. Each methodology has its strengths and weaknesses that differ from location specific settings. The actual procedure of using the data holds another set of challenges. Guiding the location of farms by average values will of course produce false locations. The question is however whether it reduces mistakes to a larger degree than a random location procedure would produce. A question we hope to investigate in the near future.

Conclusion

The ability to find values through supplementing field studies to help the location of farms for FADN-based spatial models has been demonstrated. Both the variability of an individual farms spatial relationship as well as the average values of farm categories found in the FADN sample has been investigated with regard to variation in sample size and composition.

Acknowledgement

The research for this paper has received funding from the European Commission's 6th Framework Programme (MEA-Scope, STReP No. 501516). This publication reflects only the views of the author. The Community is not liable for any use that may be made of the information therein.
We are grateful for the data supplied by Chris Kjeldsen, DIAS.

References

Beers, G. et al. (2001): Pacioli 8 –Innovation in the FADN, Report 8.01.02 Agricultural Economics Resarch Institute (LEI), The Hague

Elbersen, B et al. (2006): Protocols for spatial allocation of farm types,

SEAMLESS No.010036 Deliverable number: PD4.7.1, Wageningen, NL

Fais,A & Nino,P.(2004): Mapping the Spatial Distribution of Plant Diseases, 24th Annual ESRI International User Conference Proceedings: pap1820. San Diego, California USA.

Fais,A et al. (2005): Microeconomic and GEO-Physical data integration for Agri-environmental analysis, georeferencing FADN data: A case study in Italy, Paper prepared for the XIth seminar of the EAAE "The Future of Rural Europe in the Global Agri-Food System", Copenhagen, Denmark, 24-27 August, 2005

Meier, B (2004): The role of cash flow indicators in understanding farm households, Paper prepared for OECD "Workshop on Information Needs for the Analysis of Farm Household Income Issues", Paris, France 29-30 April 2004

Meier, B (2005):"Organic" Sampling and Weighting in Farm Accountancy Data Networks –A Discussion Note on Standard Gross Margins and Calibration, from "Towards a European Framework for Organic Market Information" Proceedings of the Second EISfOM European Seminar, Brussels, November 10-11, 2005

Appendix 1:

All farms

Summary:	20	100	400	1000
All FARMS	50.57%	12.53%	3.38%	2.25%
0-20ha farms	50.34%	12.72%	3.54%	2.30%
21-50ha farms	50.46%	12.07%	3.98%	2.38%
51-100ha farms	51.07%	12.03%	3.20%	2.04%
101-200ha farms	51.03%	14.23%	2.44%	1.90%
more than 200ha farm	50.56%	14.08%	4.01%	2.53%
plant_production farm	50.69%	13.83%	3.00%	2.17%
1-50 animal unities	50.46%	12.06%	3.70%	2.30%
more than 50 animal u	50.47%	11.57%	3.78%	2.28%
pock	50.56%	12.42%	3.93%	2.41%
dairy	50.49%	11.58%	3.78%	2.23%

0-20ha farms

Summary:	20	100	400	1000
All FARMS	33.32%	7.55%	4.18%	2.21%
0-20ha farms	34.33%	7.66%	4.16%	2.21%
21-50ha farms	34.94%	6.80%	4.44%	2.22%
51-100ha farms	29.12%	7.08%	4.60%	2.18%
101-200ha farms	29.34%	11.01%	5.03%	2.22%
more than 200ha farm	40.49%	9.82%	5.16%	2.69%
plant_production farm	32.40%	9.93%	4.39%	2.23%
1-50 animal unities	34.04%	6.98%	4.27%	2.20%
more than 50 animal u	34.31%	6.32%	4.51%	2.21%
pock	36.01%	7.56%	4.56%	2.32%
dairy	33.42%	6.01%	4.52%	2.15%

21-50ha farms

Summary:	20	100	400	1000
All FARMS	24.97%	10.89%	7.06%	2.44%
0-20ha farms	26.72%	11.25%	6.59%	2.32%
21-50ha farms	25.13%	10.39%	6.66%	2.10%
51-100ha farms	20.97%	10.98%	8.77%	2.87%
101-200ha farms	28.92%	14.62%	7.79%	3.35%
more than 200ha farm	35.98%	14.67%	7.57%	2.56%
plant_production farm	28.74%	13.36%	6.84%	2.69%
1-50 animal unities	23.56%	10.10%	7.06%	2.28%
more than 50 animal u	21.47%	9.47%	7.90%	2.32%
pock	26.96%	11.05%	6.75%	2.19%
dairy	20.51%	9.37%	8.12%	2.38%

51-100ha farms

Summary:	20	100	400	1000
All FARMS	304.90%	76.24%	62.59%	18.08%
0-20ha farms	305.09%	76.24%	62.65%	18.04%
21-50ha farms	299.00%	76.07%	62.82%	18.01%
51-100ha farms	305.41%	76.22%	62.35%	18.23%
101-200ha farms	322.47%	77.18%	61.93%	18.28%
more than 200ha farm	313.54%	76.87%	62.88%	18.28%
plant_production farm	314.12%	76.83%	62.27%	18.09%
1-50 animal unities	301.14%	75.95%	62.69%	18.05%
more than 50 animal u	297.67%	75.85%	62.87%	18.15%
pock	301.58%	76.05%	62.90%	18.07%
dairy	297.17%	75.69%	62.72%	18.13%

101-200ha farms

Summary:	20	100	400	1000
All FARMS	132.39%	87.32%	97.66%	45.47%
0-20ha farms	132.21%	86.80%	97.75%	45.21%
21-50ha farms	133.25%	87.21%	97.77%	45.55%
51-100ha farms	132.51%	88.68%	97.09%	46.39%
101-200ha farms	129.74%	87.42%	97.80%	45.12%
more than 200ha farm	134.77%	86.60%	99.28%	45.11%
plant_production farm	129.64%	86.60%	97.57%	44.54%
1-50 animal unities	133.27%	87.41%	97.64%	45.67%
more than 50 animal u	134.95%	88.28%	97.89%	46.58%
pock	133.08%	86.84%	98.09%	45.56%
dairy	135.11%	88.61%	97.54%	46.58%

more than 200ha farms

Summary:	20	100	400	1000
All FARMS	466.76%	281.75%	91.46%	50.22%
0-20ha farms	468.18%	283.22%	91.94%	50.19%
21-50ha farms	482.32%	282.26%	92.78%	49.96%
51-100ha farms	459.67%	276.07%	88.91%	50.16%
101-200ha farms	419.13%	282.50%	88.99%	51.50%
more than 200ha farm	486.57%	283.89%	94.94%	50.56%
plant_production farm	439.52%	284.31%	90.62%	50.88%
1-50 animal unities	477.63%	281.32%	92.09%	50.05%
more than 50 animal u	488.82%	278.57%	91.62%	49.51%
pock	481.55%	282.93%	92.82%	49.82%
dairy	486.15%	277.33%	91.33%	49.75%