

# **Integration of small area estimation and mapping techniques**

Tool for Regional Studies

Hans C.J. Vrolijk  
Wietse Dol  
Tom Kuhlman

Project code 30157

Juni 2005

Report 8.05.01

LEI, The Hague

The Agricultural Economics Research Institute (LEI) is active in a wide array of research which can be classified into various domains. This report reflects research within the following domain:

- Statutory and service tasks
- Business development and competitive position
- Natural resources and the environment
- Land and economics
- Chains
- Policy
- Institutions, people and perceptions
- Models and data

Integration of small area estimation and mapping techniques; Tool for Regional Studies  
Vrolijk, H.C.J., W. Dol and T. Kuhlman  
The Hague, LEI, 2005  
Report 8.05.01; ISBN 90-8615-010-1; Price € 20.- (including 6% VAT)  
60 pp., fig., tab., app.

Research projects are increasingly aimed at specific regions. A tool for statistics for regional studies was developed to combine all available information from the agricultural census and the Farm Accountancy Data Network. In this report a description is given of the development of methods and tools to display small area estimates in maps.

Orders:

Phone: 31.70.3358330

Fax: 31.70.3615624

E-mail: [publicatie.lei@wur.nl](mailto:publicatie.lei@wur.nl)

Information:

Phone: 31.70.3358330

Fax: 31.70.3615624

E-mail: [informatie.lei@wur.nl](mailto:informatie.lei@wur.nl)

© LEI, 2005

Reproduction of contents, either whole or in part:

- permitted with due reference to the source
- not permitted



The General Conditions of the Agricultural Research Department apply to all our research commissions. These are registered with the Cental Gelderland Chamber of Commerce in Arnhem.



# Contents

	Page
<b>Preface</b>	7
<b>Summary</b>	9
<b>1. Introduction and problem statement</b>	11
1.1 Introduction	11
1.2 Problem statement	11
1.3 Structure of report	12
<b>2. Alternative ways of mapping agricultural data</b>	13
2.1 Introduction	13
2.2 Mapping by region	13
2.3 Mapping by farm	14
2.4 Spatial analysis: constructing special regions	16
2.5 Geostatistical analysis	21
<b>3. STARS: statistics for regional studies</b>	23
3.1 Data imputation versus direct estimations	24
3.2 Estimating regional results of dairy farmers	25
<b>4. Tool for regional studies</b>	30
4.1 Introduction	30
4.2 GIS in STARS	30
4.3 Defining a map in the GIS viewer	32
4.4 Showing a map in the GIS viewer	34
4.5 Showing results of imputation procedure	38
4.6 Inspecting data	42
4.7 Defining set of available MAPS	44
<b>Literature</b>	45
<b>Appendix</b>	
1. STARS: Statistics for Regional Studies	47



## Preface

Research projects are increasingly aimed at specific regions. A tool for statistics for regional studies was developed to combine available information from the agricultural census and the Farm Accountancy Data Network. Combining information from different sources increases the reliability of estimates of small areas. This tool was developed by Wietse Dol and Hans Vrolijk. In the current project the tool was extended with the possibility to display regional results in maps. The authors would like to thank Marcel Betgen, Tim Verwaart and Foppe Bouma for their contributions in this project.



Prof. Dr. L.C. Zachariasse  
Director General LEI B.V.



## Summary

In recent years methods have become available for combining data from the agricultural census and data from the Farm Accountancy Data Network. Using both data sources enables the researcher to make more reliable estimates for small areas. Estimating results for regions easily leads to a demand to show results in geographical maps. Geographical information systems provide this opportunity. Combining the statistical methods for small area estimation and geographical information systems creates the following powerful options: (a) to show aggregated FADN data in maps; (b) to estimate the variables which are measured in the FADN for all farms in the agricultural census; and (c) to display all variables from the agricultural census and the FADN on each possible aggregation level - from county level to individual farm level. This report gives a description of a LEI project aimed at the development and use of tools to combine data from different data sources and to display results in geographical maps. These tools enable researchers to easily create maps with relevant data.



# 1. Introduction and problem statement

## 1.1 Introduction

In recent years methods have become available for combining data from the agricultural census and data from the Farm Accountancy Data Network. Using both data sources enables the researcher to make more reliable estimates for small areas (Vrolijk et al., 2002). Estimating results for regions easily leads to a demand to show results in geographical maps. Geographical information systems provide this opportunity. Combining the statistical methods for small area estimation and geographical information systems creates the following powerful options: (a) to show aggregated FADN data in maps; (b) to estimate the variables which are measured in the FADN for all farms in the agricultural census; and (c) to display all variables from the agricultural census and the FADN on each possible aggregation level - from county level till individual farm level. The latter creates the opportunity to make analyses based on spatial correlations, for example the use of pesticides in bird protection areas and the optimal location of market or distribution places.

For privacy reasons it is of course not allowed to display values of individual farm, but it is possible to process the data in such a way that detailed maps can be created without revealing sensitive information.

## 1.2 Problem statement

This report gives a description of a LEI project aimed at the development and use of tools to combine data from different data sources and to display results in geographical maps. The goal of these tools is to enable researchers to easily create maps with relevant data.

To achieve this goal, a software application has to be developed to display maps based on FADN information (for example for 14 agricultural regions) and based on data from the agricultural census (for example at the level of municipalities, 31 manure regions or 66 agricultural regions). These types of maps can be directly created based on data from the available databases. At a more detailed level, maps can be made by using STARS. Instead of actual data maps are based on estimated data. The design of the tool is displayed in figure 1.1.

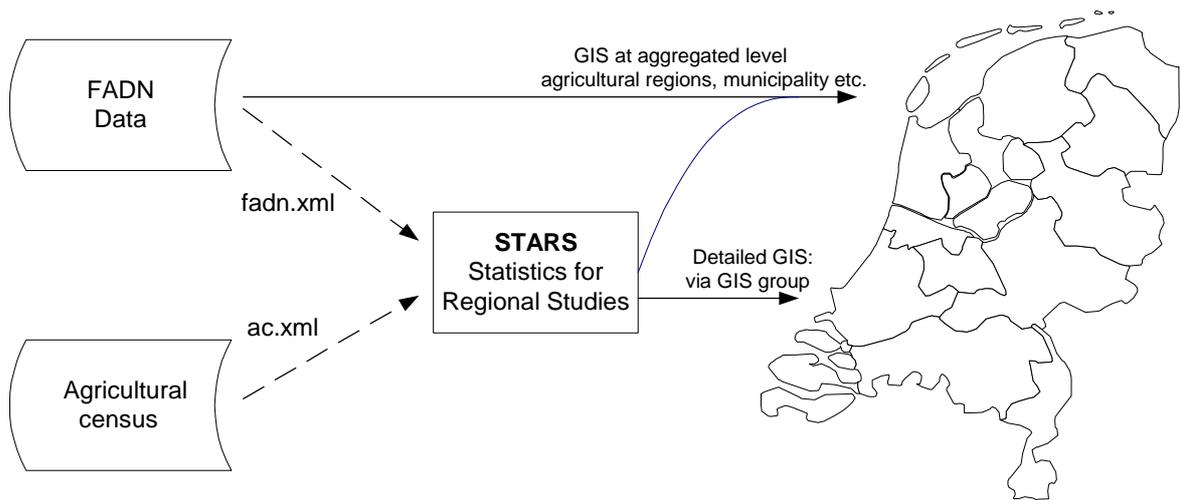


Figure 1.1 Elements of a tool for regional studies

### 1.3 Structure of report

In chapter 2 an introduction will be given in the possibilities for mapping of agricultural data. Data from the agricultural census and from the agricultural business survey can be visualised at different spatial levels and with a variety of techniques. Chapter 2 will discuss and evaluate some of these techniques. Chapter 3 gives a short introduction in small area estimation techniques as implemented in the tool Statistics for Regional Studies (STARS). Chapter 4 will give a description of the newly developed tool for regional studies.

## 2. Alternative ways of mapping agricultural data

### 2.1 Introduction

Data from the agricultural census and from the agricultural business survey can be visualised at different spatial levels and with a variety of techniques. The purpose of this chapter is to present some feasible alternatives from the GIS toolbox and discuss their relative merits.

### 2.1 Mapping by region

The simplest way to map any data is to tabulate them by region and link the resulting table to a map of the regions concerned. In the Dutch agricultural census, for example, the municipality and the agricultural region<sup>1</sup> in which the farm is situated are recorded on the census form, so the data can be easily aggregated to either of those two levels. A table with data on regions can be linked with a digital map of those regions, and any of the columns in the table can then be visualised on the map - the number of farms per region, the area cultivated, or the number of cattle. Also, for each of these themes a separate map can be shown based on the same digital data.

Apart from data actually available per region, with the aid of the STARS technique it is also possible to *estimate* them on the basis of data from the sample survey of farms (BIN). Since selected variables for each individual farm can be estimated, these estimates can also be aggregated to the aforementioned regional levels.

The simplicity of this procedure makes it suitable for automation. A standard regional level can be set (or a few levels from which the user can select); one procedure in the program generates a table for one particular variable; the program can be made to generate a default classification for that variable<sup>2</sup>, a standard color scheme can be used to visualise the classes per region; and hey presto, there is your map. Figure 2.1 shows an example.

---

<sup>1</sup> The Netherlands has been divided into 66 agricultural regions which exhibit a certain homogeneity in production conditions. The boundaries of these regions are made to coincide with municipal boundaries, and the regions can be aggregated into the 12 provinces into which the country is divided administratively. These 66 regions can also be grouped into 14 larger agricultural zones.

<sup>2</sup> The classification itself can, of course, not be uniform: a classification for the number of farms will not satisfy for the number of euros earned per year. However, GIS applications can scan a table and propose a suitable classification based on the frequency distribution of the variable measured and using a standard technique such as Jenks natural breaks, equal intervals or quantiles.

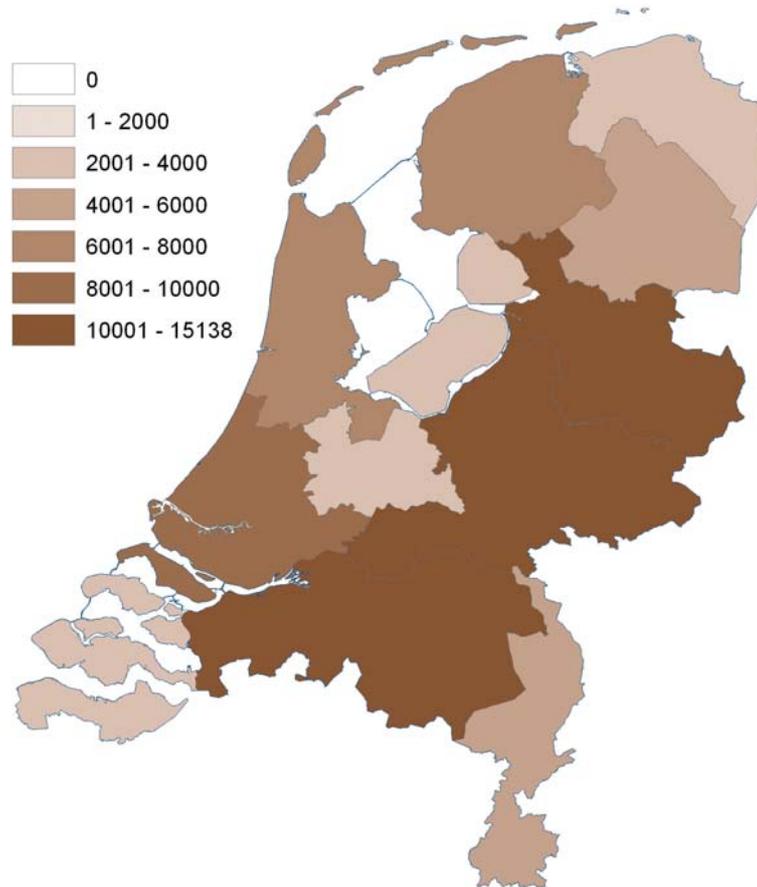
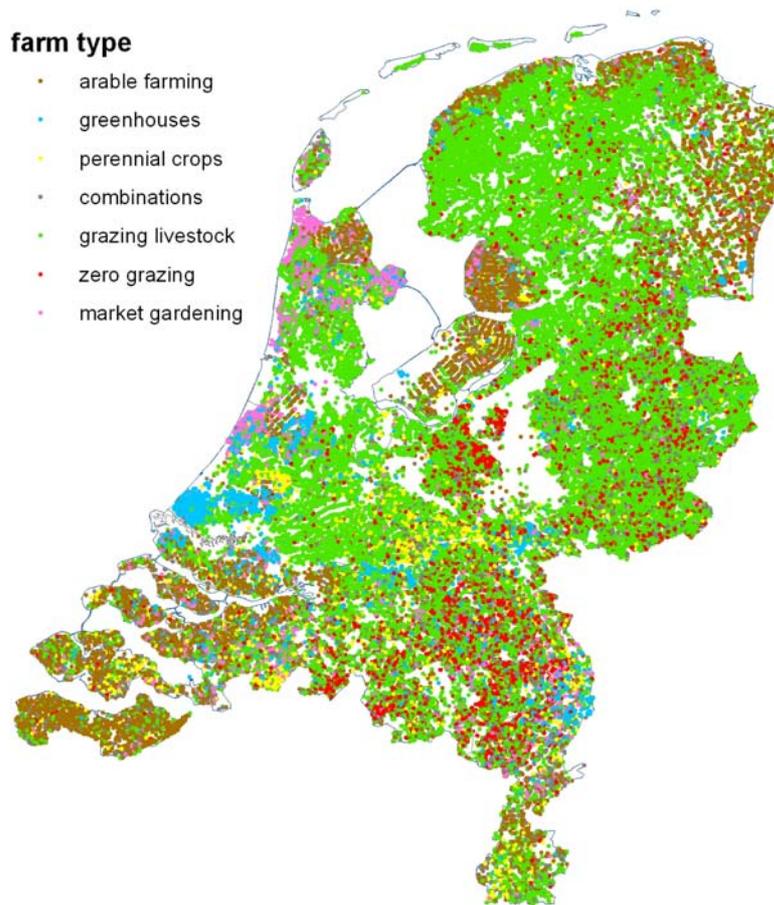


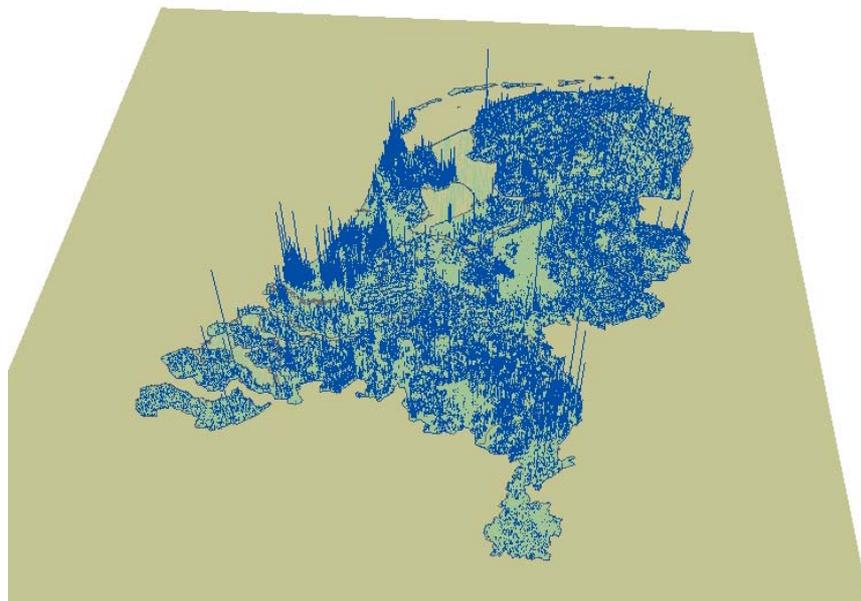
Figure 2.1 Number of farms by province  
 Source: Statistics Netherlands - Agricultural Census 2003.

### 2.3 Mapping by farm

While easy to make, maps by region suffer from a lack of detail and consequently look rather bland. Furthermore, from a scientific point of view they suffer from the disadvantage that the regions do not follow the actual spatial distribution of the phenomenon concerned. In figure 1.1, for instance, a province may show a larger number of farms than another, yet some areas within that province may be entirely without farms. If we know exactly where the farms are located (which in the case of the Dutch agricultural census we do), we could also map our variables by farm, for instance by giving the quantities different colour codes on a two-dimensional map (figure 2.2) or by making three-dimensional histograms (figure 2.3). On both maps, each individual farm is represented.



*Figure 2.2 Farms by type*  
 Source: Statistics Netherlands - Agricultural Census 2003.



*Figure 2.3 Production capacity per farm*

These pictures show something of the power of GIS tools. However, while the level of detail can be attractive, these methods also have their drawbacks:

- individual variations make it difficult to see regional patterns (the trees obscuring the view of the forest);
- there are legal restrictions against publicizing data on individual businesses (for this reason, the data reflected in figure 2.2 have been altered); and
- it is difficult to visualise data based on geographical points - as both figures 2.1 and 2.2 illustrate.

To overcome these problems, yet utilise the benefits that detailed spatial data can offer, there is a third possibility: the various techniques from the toolbox of spatial analysis.

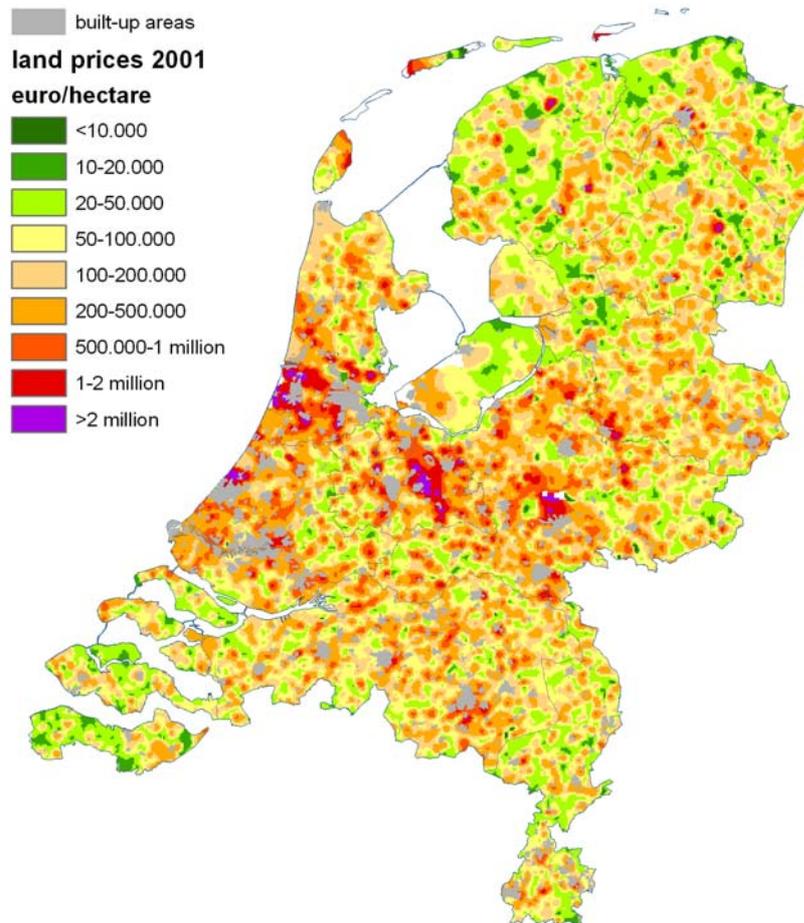
## **2.4 Spatial analysis: constructing special regions**

Spatial analysis in GIS essentially involves converting a digital map of points, lines or areas into a pattern of grid cells within a raster (McCoy and Johnston 2001). Mathematical operations can then be carried out on the values in these cells, and in this way new maps can be constructed by combining several digital maps or by simulating spatial processes.

One of the techniques of spatial analysis is interpolation. This is essentially a statistical technique, in which measurement points (e.g. farm locations) are regarded as sampling locations for a certain quantity. For instance, if we take the price of land per hectare as the quantity to be measured, we can estimate the price in different areas on the basis of actual sales in a particular year. Only some plots were sold, and these can be regarded as a sample for the value of all plots in that year. For each cell in the raster, the distance to various sample locations is calculated, and the values measured at the sample points are compared. The values for each cell are now estimated with one of a variety of mathematical formulas - the choice depending on what you assume about the nature of spatial correlation in your variable. We do not have spatial data on crop yields, so we use land prices as an example of a quantity that can be handled in this way. The result is shown in figure 2.4 (taken from Luijt et al., 2003).

A pattern emerges, with the highest values near the larger cities and the horticultural areas and relatively low values in the north and the southwest. The individual data are no longer visible: you cannot see the price of any particular piece of land.

The technique of interpolation is suitable only for those variables which can be thought of as representing sample measurements for a quantity that is present everywhere - such as the value of land or the depth of the groundwater table. Many of the variables we study in agriculture, however, are of a different nature. Farms and chickens, for instance, only exist at certain locations; with the agriculture census we have data for all these locations. In other words, we have a population, not a sample. Yet our data do not cover the whole country.



*Figure 2.4* agricultural land values, 2001  
 Source: Rural Area Service - INFOGROMA database.

For mapping such data we use the concept of density: the number of units per square kilometre or per hectare. Density can also be handled by spatial analysis, and it works very similarly to interpolation. A continuous surface is created from point measurements. Production capacity, used for figure 2.3, is a good example of a quantity that can be so measured. It is a way of adding crop areas and livestock units for comparing farms by size, and is expressed in so-called Dutch size units, abbreviated as NGE.

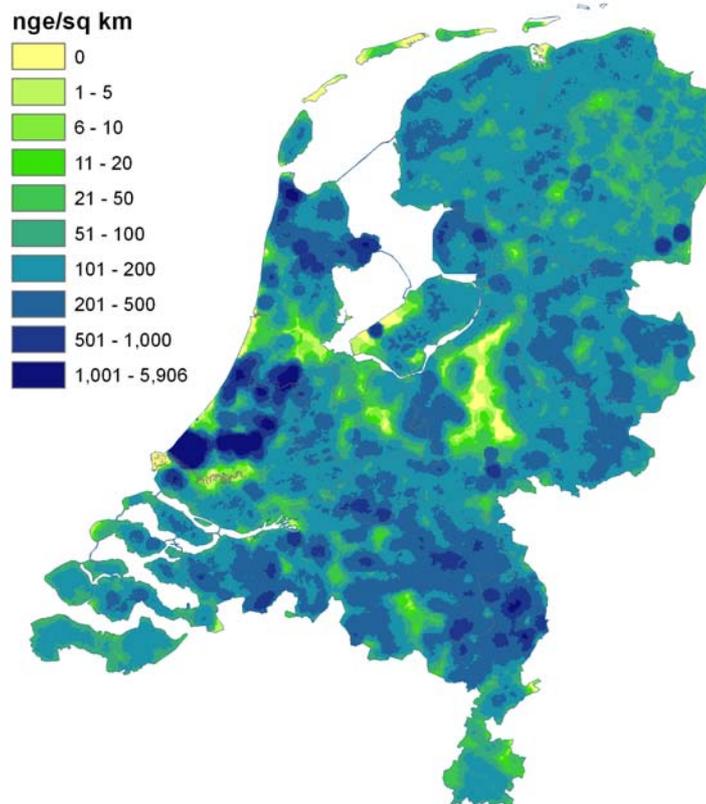


Figure 2.5 Agricultural production capacity, 2003  
 Source: Statistics Netherlands - Agricultural Census.

This kind of analysis can be manipulated in several ways: the formula can be altered; the radius within which other points are searched from each measurement point in order to establish an average can be set; and the size of the grid cells can be modified. For figure 2.5, we used a so-called simple density calculation (adding the values at the points that fall within the search area and dividing by the size of that area), a search radius of 3 km (based on the assumption that, in the Netherlands at least, most farm land is within 3 km from the main farm building),<sup>1</sup> and cells of 500x500m.

In order to illustrate the range of manipulation, figure 2.6 shows a density map of the same variable, but now with density calculated on a kernel basis, in which points near the centre of the search area are weighed more heavily; this produces a smoother distribution of values. Furthermore, the search radius has been extended to 10 km, making the pattern broader. In a way, we have now constructed our own regions - not with arbitrary boundaries but based on the actual spatial distribution of the phenomenon we are interested in.

---

<sup>1</sup> The locations we have from the agriculture census are those of the main farm buildings.

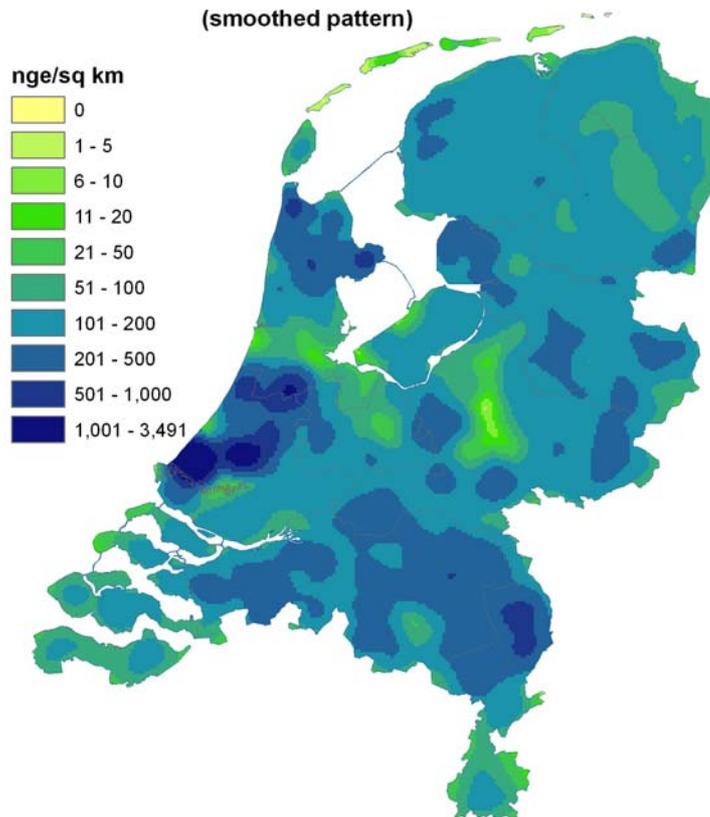


Figure 2.6 Agricultural production capacity, 2003  
Source: Statistics Netherlands - Agricultural Census 2003.

Great care is needed in choosing the right technique for a particular question, configuring it adequately, and interpreting it correctly. As an example of two very different ways to look at an issue, take figure 2.7 and figure 2.8. These examples are based on a dataset of dairy farmers in the Agriculture Census of 1999; with the aid of the STARS program, their incomes are estimated based on similar farms in the Farm Business Survey sample set. The data on which the two maps are based are exactly the same, but a different technique of spatial analysis has been chosen for each.

Orange indicates low incomes, green high ones, and yellow in between. Figure 2.6 is the result of an interpolation (in this case, using the inverse distance weighted method), whereas figure 2.7 measures density, i.e. the income earned per square km. Whereas the Veluwe forest area in the central-eastern part of the country shows low values on both maps, for most other areas the results arrived at are quite different on each map. For instance, the area west of centre between Amsterdam, Utrecht and Rotterdam (the so-called Green Heart of Holland) shows some of the highest scores on figure 2.8, on figure 2.7 the incomes are low to average. Similar differences can be found for other areas on the two maps. The reason for this is that figure 2.7 shows the income per farm - or more precisely, it estimates what a farm would earn at a particular location if there were a farm there - while the scores in figure 2.8 partly depend on the number of farms within any given area.

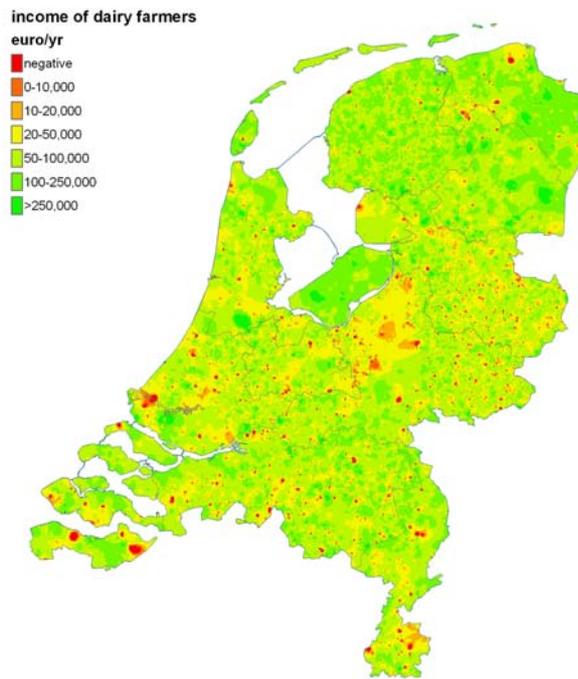


Figure 2.7 *Income of dairy farmers, 1999*  
 Source: LEI- Farm Business Survey 1999 / Central Statistical Office - Agriculture Census 1999.

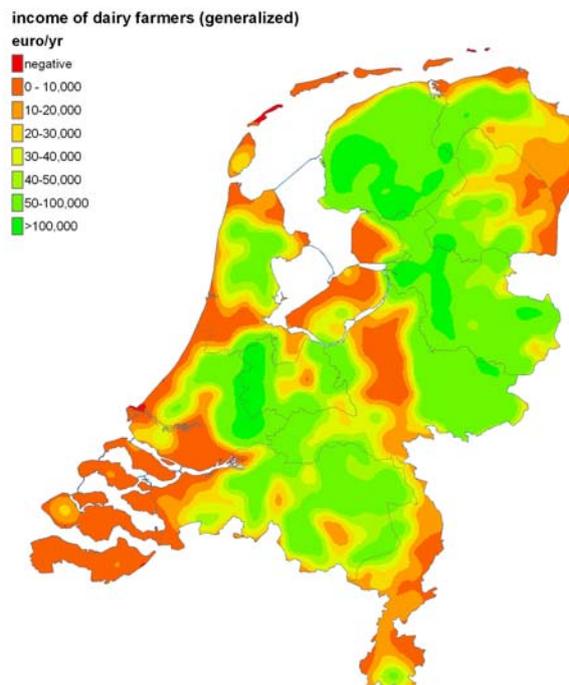


Figure 2.8 *Income of dairy farmers, 1999*  
 Source: LEI - Farm Business Survey 1999 / Central Statistical Office - Agriculture Census 1999.

The areas that have a strong green colour on figure 2.7 are not necessarily those where the earnings are good, but may be those where most dairy cows are.<sup>1</sup> Looking at both maps can also teach us something: there are areas where much dairy income is earned, but per farm the income is not high - and vice versa, for instance in the reclaimed area of Flevoland, or in the northeastern corner of the country. We must bear in mind, by the way, that data such as these are estimated from a limited sample on the basis of some matching variables.

## 2.5 Geostatistical analysis

Density is just a way of finding a pattern for the spatial distribution of a known population. When we are dealing with a sample, as in the method of spatial interpolation described above, we are making a prediction of the values in each cell. The techniques as such do not provide us with any information as to the degree of confidence with which we can make that prediction. Statistical techniques must be brought to bear.

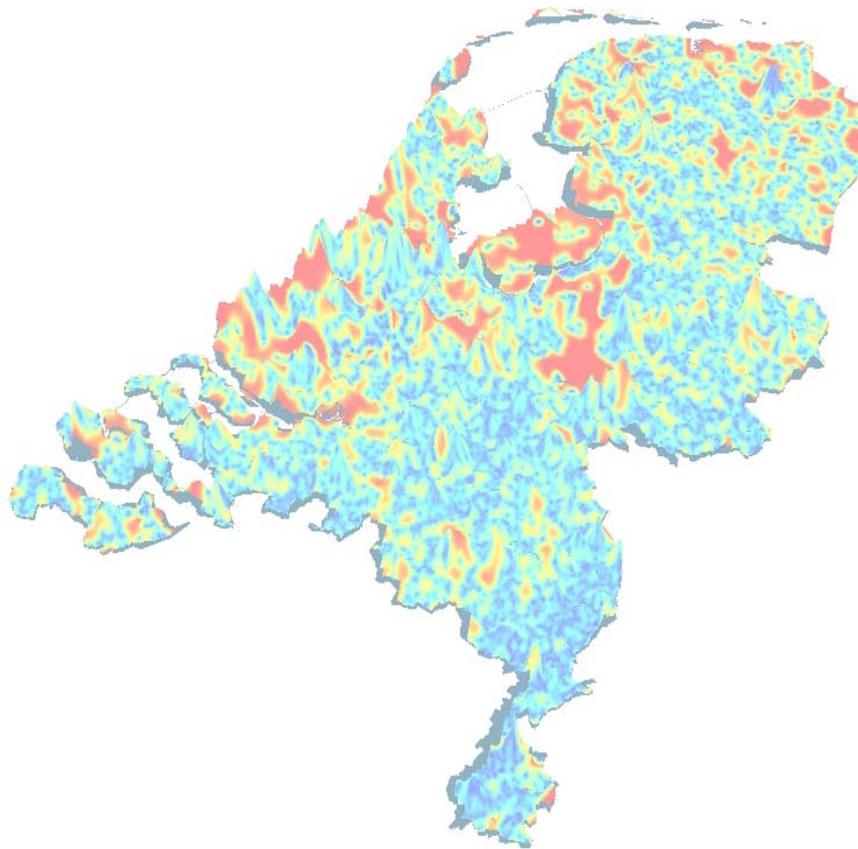
One such technique has been developed by the Centre for World Food Studies (SOW-VU) in Amsterdam: the *Mollifier* (e.g. Keyzer & Sonneveld. 1997). The Mollifier is a statistical method for calculating estimated values of one or more dependent variables on the basis of two or more independent variables. This method can also be applied in GIS, where the x- and y-coordinates determining location in space are the independent variables; the dependent variable is then measured at the point determined by x and y as the height (the z-value, in GIS jargon), which can be visualised in a three-dimensional image as we did in figure 2.3. So far it is basically the same as the kernel density function in ESRI's ArcGIS Spatial Analyst, but with a tool added to quantify the uncertainty. However, the uncertainty here is assumed to be more in the location of the observation rather than in the value observed, which makes it less useful to most of our research.<sup>2</sup>

An alternative is ESRI's Geostatistical Analyst. This is a software package which integrates geostatistical methods with GIS techniques and a GIS interface (Johnston et al. 2001). This provides us with a range of tools to diagnose the probability of our predictions. Figure 2.9 illustrates these possibilities, on the basis of the land values map shown in figure 2.4 (agricultural plots sold in 2001). The heights in this figure represent the estimated land values, the colours the reliability on the basis of standard error: blue for low error, red for high.

---

<sup>1</sup> But not only: it is a combination of income data and number of farms. The number of cows could be more easily presented directly from the Agriculture Census.

<sup>2</sup> Marc Hoogerwerf (Alterra, Wageningen University), personal communication, June 2004.



*Figure 2.9 Estimated agricultural land values and their reliability*

As mentioned in the previous section, geostatistical analysis is useful when dealing with samples, and estimating values for locations other than those in the sample. Land values such as discussed above are a good example. Some variables from the Farm Business Sample Survey (BIN), as estimated for other farms through the STARS program, can also be analysed with these techniques. In agricultural economics there may be relatively few topics where geostatistical analysis can be used, but in those cases it is a powerful tool.

### 3. STARS: statistics for regional studies<sup>1</sup>

Surveys are widely applied to provide information about important population characteristics. The datasets of surveys are mainly used to generate statistics for the whole population. Based on the observations and a set of weights an estimate can be made for the population. Given the availability of these survey datasets, it is interesting to re-use this information to make estimations for regions or specific groups. The original sample was often not designed to make this kind of estimations. The number of sample elements belonging to a region or group can be limited. This results in estimates with a low reliability.

In agriculture, data from the Farm Accountancy Data Network (FADN) are often used to estimate population characteristics. The use of FADN data in regional studies is often problematic due to the low number of observations. Several methods have been developed to use additional information to increase the reliability of estimates (Dol, 1991; Baker et al. 1994; Vrolijk and Wedel, 1996; Gelman et al. 1998; Vrolijk et al. 2002). Additional information that can be used is for example the agricultural census. The agricultural census gives a complete list of the population of farms. The amount of information in this census is however limited. In this paper we will describe an option to make use of this additional information from the census to make more reliable estimates in regional studies. The procedure has been implemented in the software tool Stars.

In a specific research project attention focuses on farms of a certain region, farms that belong to a certain type or a combination of both. We will call this group the population of interest or population in short. In the imputation procedure, for each farm in the population, a farm in the FADN sample is selected which resembles the farm as closely as possible. The researcher selects the variables, which are used to decide whether a farm resembles a sample farm. These variables are called the imputation variables. The imputation variables should be known for all farms in the sample and the population. Based on these variables the distance is calculated. Different methods are available to establish this distance. The sample farm with the smallest distance is regarded as the farm that resembles the population farm as closely as possible. For each farm in the population, 5 or 10 most similar farms are selected from the sample. These best fits are recorded together with the distance measures.

Based on these best fits, estimates can be made for a set of goal variables, which are known in the sample, but unknown for all population farms. In making estimations for the population of interest a choice can be made between simple and multiple imputations. Vrolijk et al. (2002) describe that simple imputation has the disadvantage that the variance of the estimator is underestimated. The estimated (e.g. imputed) value is treated as the real value, although there is a degree of uncertainty about this value. To overcome this problem

---

<sup>1</sup> This chapter is based on Vrolijk (2004).

multiple imputation can be used. In this option, the user can define how many of the best-fit farms will be used to make estimates about the population.

### 3.1 Data imputation versus direct estimations

The approach is illustrated in figure 3.1 and figure 3.2. Figure 3.1 describes the traditional approach (see for example Cochran, 1977). The census describes the whole population (N units). Based on the population a stratified sample is drawn. Given the number of farms in the population and the sample, weighting factors per sample farm are calculated. A weighted average of the sample observations gives a good estimation of the population.

Figure 3.2 describes the data imputation approach. The same sample as in figure 3.1 is the starting point. To make estimates of the population of interest (e.g. specific region), sample farms are matched to population farms based on the imputation variable. The sample farm that is most similar to a population farm is used to impute goal variables. The basic assumption is that if the farm is similar on the imputation characteristics, then it is likely that the farm is also similar on the goal variables. To assure that this is a valid assumption, the imputation variables have to be selected in a careful way.

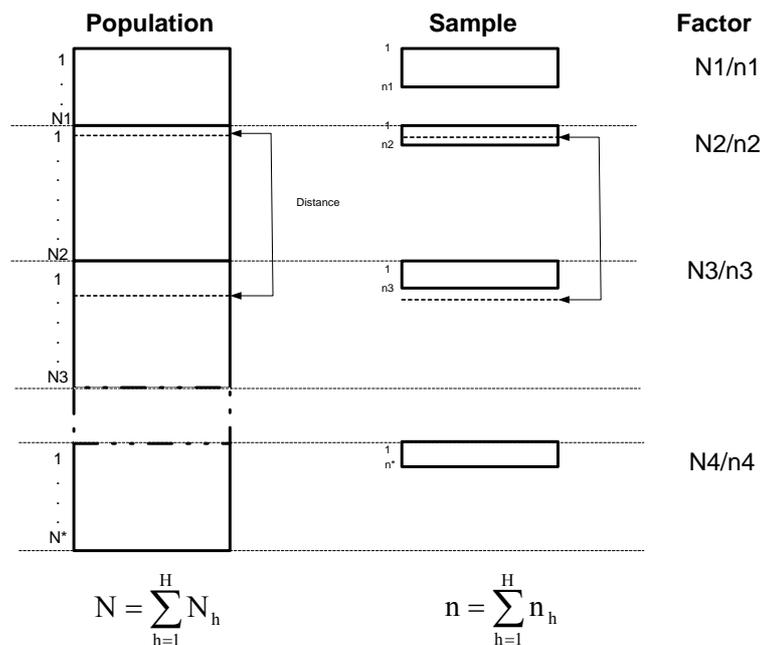


Figure 3.1 Direct estimation using weight of sample units

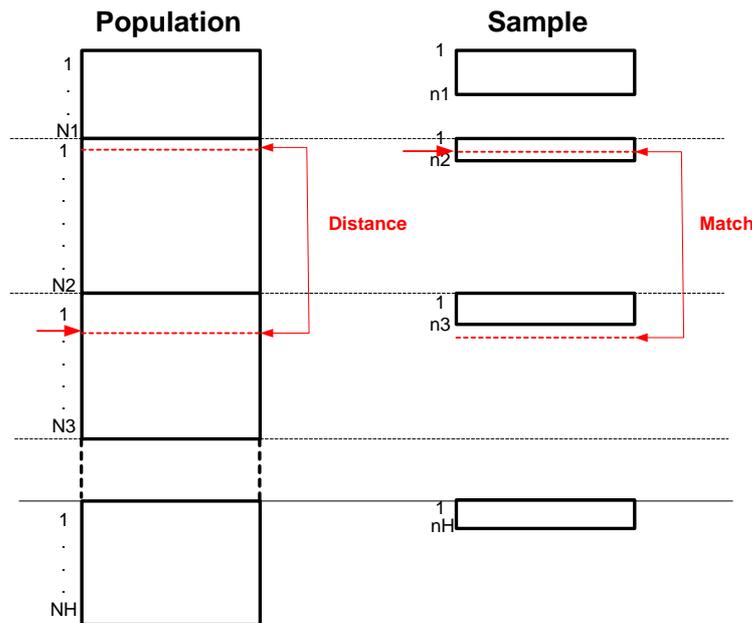


Figure 3.2 Data imputation

### 3.2 Estimating regional results of dairy farmers

In this example we explore the opportunities to make estimations for dairy farms in a municipality in the northern part of the Netherlands (black area in figure 3.3). In this example an estimate is made for the variables: total revenues, total costs, net farm result, labour income entrepreneur and number of entrepreneurs (these are the goal variables). Based on the number of observations in the FADN, it is difficult to make direct estimations. However, this municipality is part of a larger grassland area with similar production circumstances. This area, 'Noordelijk Weidegebied' (Northern Grassland Area), is one of the agricultural areas of the Netherlands (see grey area figure 3.3). 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. In the FADN, 70 dairy farms from this region are included in the sample.

In the estimation procedure a number of imputation variables is used (the choice of the variables will be explained in the next section):

- age;
- hectares grass;
- hectares fodder crop;
- number of dairy cows;
- economic size.

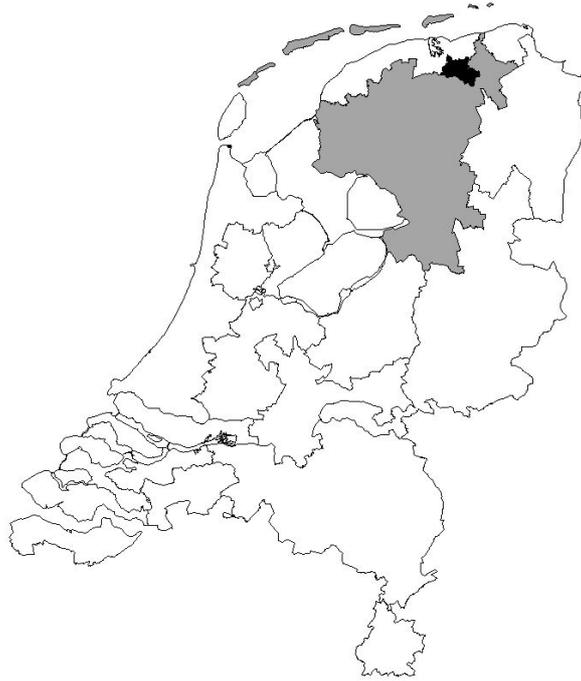


Figure 3.3 Municipality of interest (black) in Northern Grassland Area (grey)

In table 3.1 the results of the imputation process are described. In this example a single imputation is applied. For each farm in the population in the municipality the most similar farm in the FADN sample in the Northern Grass Area is selected. The similarity is based on the five imputation variables as described above (to take into account the different units of measurement the variables are standardised before calculating the distance). Subsequently the average of the imputed values for all farms in the municipality are calculated, assuming that the values of the most similar farms in the Northern Grass Area provide a good approximation of the value of that specific farm.

Table 3.1 Results of imputation process (single imputation)

	Mean	Standard error
Revenues	415,020	15,028
Costs	506,479	15,103
Net farm result	-80,069	4,581
Labour income per entrepreneur	58,066	5,010
Number of entrepreneurs	1.47	0.05

Single imputation has the disadvantage of underestimating the variance. The imputed values for a specific farm are considered as the true values, although there is a certain uncertainty about these values. In table 3.2 the results are displayed for a multiple imputation process. The three most similar farms are used to make an estimation for the municipality. In this multiple imputation process 100 independent replications are applied. In each replication one of the three nearest neighbors is randomly selected. The values of that neighbor are used to impute the values and make estimations for the region. Comparing tables 3.1 and 3.2 shows that the estimations of the means are not very different. It also shows that the variance of the estimator increases due to the multiple imputation process. This increase is caused by the addition of between replication variance. The columns Min and Max show that the estimation of the average total revenues varies between 405 and 431 thousand. This variance is added to the variance as a consequence of differences between farms within a replication (within variance). The variance increases by 10% for the different goal variables.

Table 3.2 Results of imputation process (multiple imputation)

	Mean	Standard error	Min	Max
Revenues	417,203	16,723	405,002	431,081
Costs	505,405	16,354	492,738	521,129
Net farm result	-76,984	5,502	-85,138	-69,606
Labour income per entrepreneur	63,899	6,459	56,126	75,055
Number of entrepreneurs	1.49	0.05	1.4	1.6

Until now, the quality of the imputation process is not explicitly considered. In the remaining of this section, a validation procedure is described. The quality can be judged by using the same approach for imputing values in the sample (which are known) under the restriction that the farm itself cannot be used to impute values. In this way the values of a sample farm are estimated by imputing values from one or more other sample farms that are very similar. Subsequently a statistical test can be conducted to check whether significant differences exist between the real values and the imputed values.

Table 3.3 Potential imputation variables

Age	Percentage other grazing livestock
Hectare	Percentage breeding pigs
Hectare grass	Percentage fattening pigs
Hectare fodder crops	Percentage poultry
Dairy cows	Percentage fodder crops
Dairy cows per hectare	Percentage grains
Total added value	Percentage tuberous plants
Added value pigs	Percentage other arable farming
Percentage dairy cows	Percentage horticulture open air

Table 3.3 lists all the variables that could be used as imputation variables. The inclusion of variables as imputation variables is only useful when there is some kind of logical relationship between this variable and the goal variables. Unlike regression analysis no assumption has to be made about the shape of the relationship. In table 3.4 a naïve approach has been applied in which all potential imputation variables have been used. This table shows that the values estimated by the imputation procedure are close to the real values. No significant differences can be shown by looking at the averages and the standard errors.

*Table 3.4 Comparison of real and estimated values*

	Real value	Estimated value	Standard error
Revenues	476,902	493,360	32,869
Costs	569,488	573,109	33,472
Net farm result	-79,303	-66,473	9,536
Labour income per entrepreneur	67,817	80,157	11,858
Number of entrepreneurs	1.53	1.49	0.09

An important question is whether all imputation variables are relevant in the imputation process. A balance has to be found between the correctness of the model and the simplicity of the model. In table 3.5 an extreme variant is applied in which the distance is only based on the age of the farmer and the hectares of grassland. This table shows large and significant differences between the estimated and real values. Based on this analysis the conclusion can be drawn that data imputation based on only these two variables result in a low quality.

*Table 3.5 Imputation based on age and hectares of grassland*

	Real value	Estimated value	Standard error
Revenues	476,902	355,033	21,028
Costs	569,488	459,701	14,797
Net farm result	-79,303	-91,233	9,601
Labour income per entrepreneur	67,817	12,530	10,507
Number of entrepreneurs	1.53	1	0

In table 3.6 the results for an imputation procedure based on five imputation variables is described. This table shows that the results are equally good or even better compared to an imputation procedure based on all imputation variables.

Table 3.6 *Imputation based on age, ha grass, ha fodder crops, number of dairy cows and economic size*

	Real value	Estimated value	Standard error
Revenues	476,902	470,917	34,330
Costs	569,488	560,114	33,836
Net farm result	-79,303	-76,492	9,182
Labour income per entrepreneur	67,817	68,500	11,297
Number of entrepreneurs	1.53	1.53	0.09

This approach provides the advantage that the basic assumption of the imputation process can be tested. Besides theoretical reasons, a quantitative analysis can provide support for the choice of the imputation variables.

## 4. Tool for regional studies

### 4.1 Introduction

STARS has been developed to support regional studies. Several methods for small area estimation are the basis for STARS. Given the regional characteristic of studies supported by STARS, the obvious demand to display results in GIS maps arose, and hence STARS has been extended with a component to create and show these maps. In this report the main functionality of this extension is described. For a detailed description of the functionality of STARS see appendix 1.

### 4.2 GIS in STARS

Besides creating a GIS viewer, STARS has been adapted to support the GIS functionality. Figure 4.1 shows the GIS button to access the GIS functionality. This button is available in the sample data window, in the population data window and in the window displaying the imputation results. Pressing the GIS button will open a window (figure 4.2) in which information required to define the map can be selected.

TYPE	REGION	COUNCIL	AGE	HA	HA_GRAS	HA_FEED	COWS	COWPERHA	ECSIZE	ECPI
4110	7	480	57	38.97	35.14	38.97	58	1.49	276.8	
4120	5	228	35	14.25	10.15	14.25	32	2.25	190.1	
4110	13	1685	44	10.55	5.25	6.05	70	11.57	218.23	
4110	5	230	56	55	52.3	55	63	1.15	344.73	
4120	2	1987	38	71.8	30	50.08	60	1.2	374.79	
4120	7	550	34	21.5	21.5	21.5	20	0.93	137.10	
4110	12	758	37	35.24	26.25	30.7	74	2.41	304.41	
4110	13	846	58	44.3	15.65	31.15	99	3.18	397.54	
4110	12	758	61	28.46	14.16	28.46	55	1.93	230.76	
4110	4	159	56	31.26	25.71	31.26	110	3.52	401.27	
4110	3	98	22	31.5	25.5	31.5	69	2.19	286.05	
4110	10	310	35	45.91	38.51	45.91	72	1.57	329.03	
4110	7	638	52	36.29	36.29	36.29	55	1.52	245.26	
4110	4	199	66	170.5	132.5	160.3	366	2.28	1439.4	
4110	1	70	33	39.1	37	39.1	62	1.59	268.3	
4110	3	51	30	60.94	60.94	60.94	98	1.61	440.94	
4110	3	56	38	67.65	65	67.65	98	1.45	447.8	
4110	13	1658	32	28.27	18.52	27.07	57	2.11	275.33	
4120	4	109	29	37.61	13.19	26.01	31	1.19	242.14	
4110	2	136	41	57.8	45.3	57.8	89	1.54	433.88	
4110	7	432	31	50.52	50.52	50.52	68	1.35	338.1	
4110	3	85	64	100.25	52.8	68.25	114	1.67	599.62	
4120	2	1697	54	57.32	50.48	57.32	81	1.41	523.63	

Figure 4.1 GIS button in STARS data window

Figure 4.2 shows the information, which is crucial for the interface between STARS and the GIS viewer. Three elements should be defined.

- Region indicator     The name of the variable in the dataset which gives the number of the region.
  
- Map                     Name of the map to display the information (in this case the municipalities (gemeenten) in the Netherlands). Appendix A describes how the set of available maps can be extended.
  
- Variable                Which variable or which variables should be displayed in a map. For each selected variable, an aggregation type should be selected. The available types are: Count; Average; Maximum; Minimum; Sum, Variance and Standard Deviation.

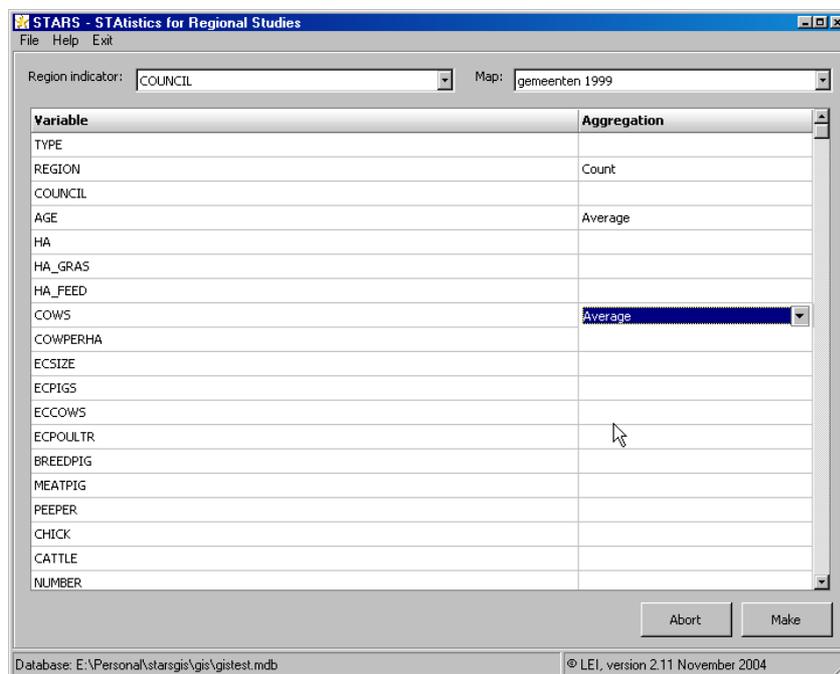


Figure 4.2 Defining the interface between STARS and GIS

After pressing the 'make' button, the appropriate information is forwarded to the GIS viewer.

### 4.3 Defining a map in the GIS viewer

The GIS viewer shows the map that is selected in figure 4.2. In the example the municipalities in the Netherlands are shown. Only the shape of the region is displayed with the default background colour.

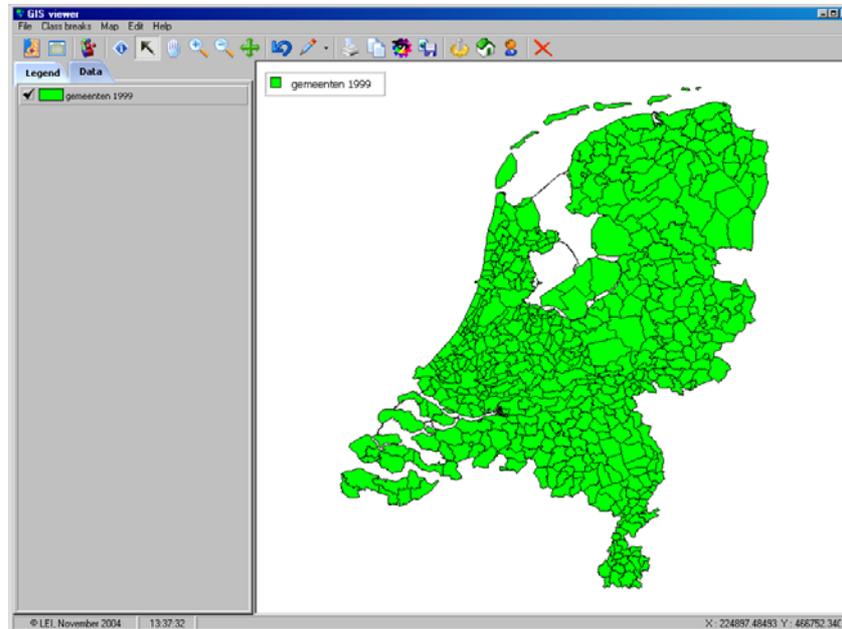


Figure 4.3 Opening window of the GIS viewer

Subsequently the content of the map should be defined based on the available information. After pressing the  button, the dialog box as displayed in figure 4.4 is shown.

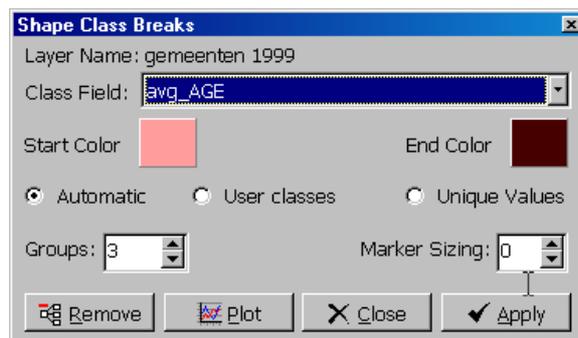


Figure 4.4 Dialog box to define map

The elements of the dialog box are:

**Class field** The name of the variable to be displayed in the map has to be selected from a list of available variables.

**Start and End Colour** The regions with lowest value will get the start colour in the map, the regions with highest value the end colour. The colours can be selected and changed according to own preferences.

*Defining classes*

**Automatic** Automatic only requires the definition of the number of groups (H). In case of automatic group definition, the set of elements are ordered according to the class field variable and divided in H groups of equal size.

**User classes** User classes gives a high degree of flexibility to define classes (see figure 4.5). The user-defined classes also allow the user to exactly determine the thresholds of the classes. This is a little bit more labour intensive but provides the highest degree of control.

**Unique values** Each unique value is represented by a different group.

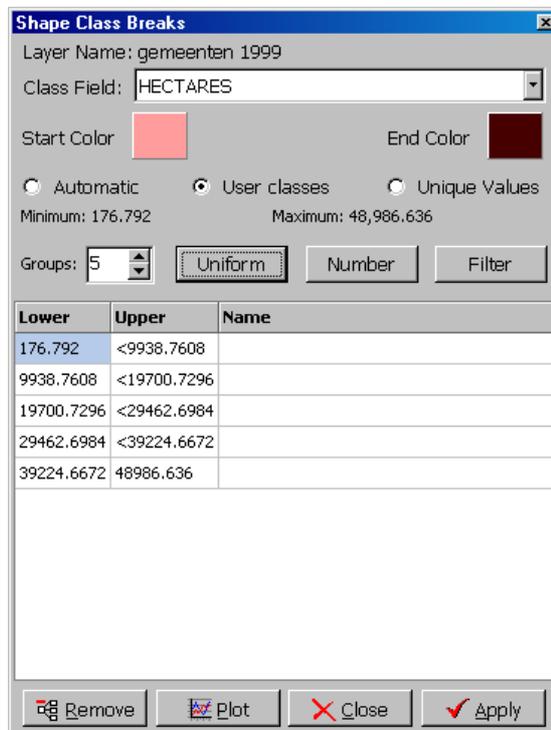


Figure 4.5 Defining User classes

User classes provide the option to manually manipulate the groups. The groups can be defined from scratch, or initial groups can be automatically generated and subsequently changed according to the preferences of the user. To generate these initial groups automatically, two options are available:

- Uniform** The first way is to define the number of groups/classes and subsequently automatically generate the groups. In case Uniform is selected, the range between the lowest and highest value is divided into equal ranges. Each range becomes one group. With skewed distributions this has the disadvantage that most of the observations will be in one of the groups.
- Number** The option Number is similar to the Automatic generation of groups, with the important difference that the group definitions can be changed. The set of elements are ordered according to the class field variable and divided in N groups of equal size.

The automatically generated initial group definitions can be changed. Defining or changing groups is difficult without a vague understanding of the values of a variable. The plot option assists in exploring the range of values that occur in the dataset. The dispersion of values as shown in the plot (see for example figure 4.6) assists the user to define the groups.

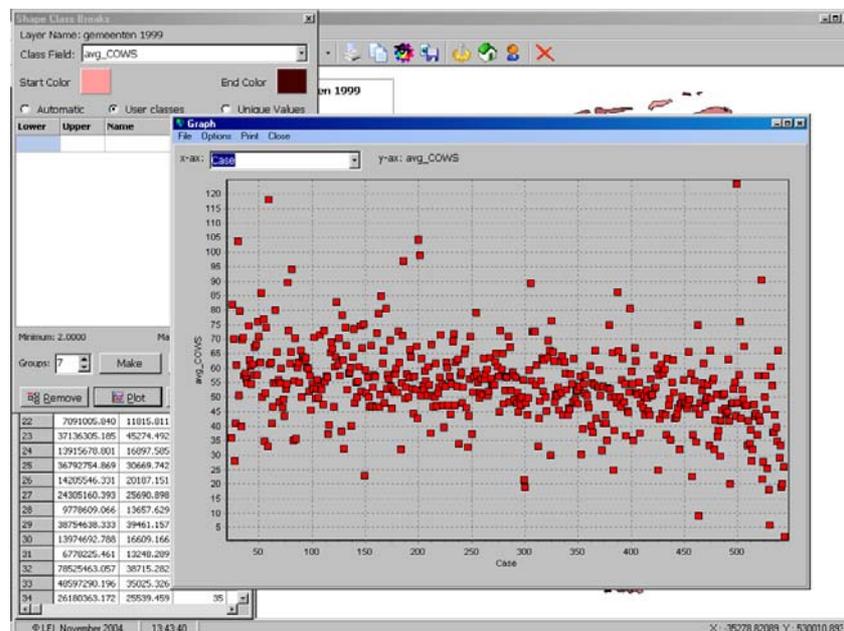


Figure 4.6 Plot to inspect distribution of average number of cows on dairy farms

#### 4.4 Showing a map in the GIS viewer

When the variable to be displayed in the map has been selected and the groups have been defined the map can be displayed by pressing the  button. Figure 4.7 shows an example of a map. In this map the average age of dairy farmers in ea 2ch municipality is displayed. The darker the colour, the higher the average age of the farmer.

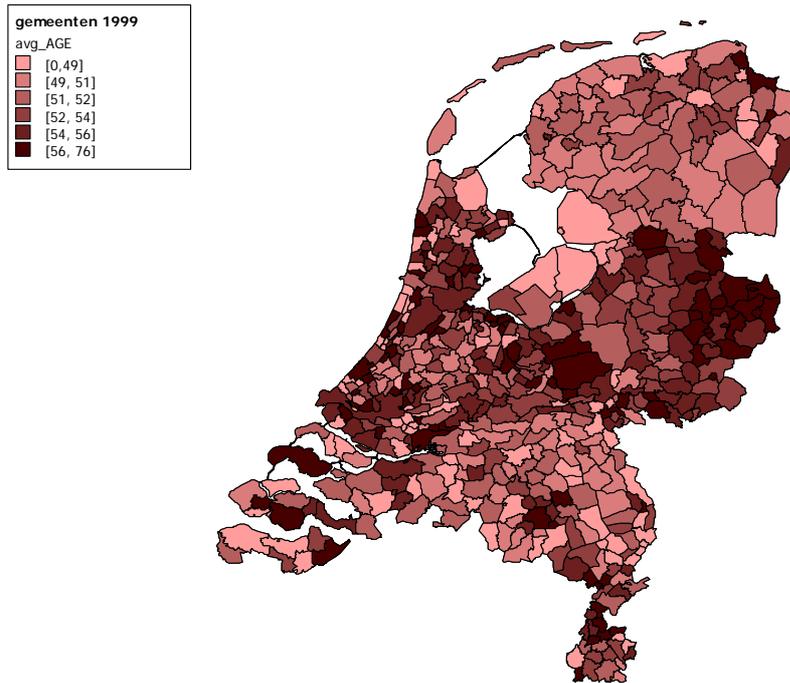


Figure 4.7 Average age of dairy farmers

For research purposes it's often interesting to inspect additional information. Pressing the  button in the button bar will display additional information for each region in the map by selecting the shape representing that region. Figure 4.8 shows all available information for the municipality Zuidhorn. The number of specialised dairy farms is 160 (count\_REGION), the average age of dairy farmers is just above 51 (avg\_AGE) and the average number of cows on specialised dairy farms is 62 (avg\_COWS).

Variable	Value
AREA	128226096.55
PERIMETER	73776.703
NLGE_R99_	27
NLGE_R99_I	26
NLGEM99NR	56
NLGEM99NM	Zuidhorn
HECTARES	12822.610
count_REGION	160
avg_AGE	51.73125
avg_COWS	62.1625

Figure 4.8 Shape information

Additional functionality can be assessed by pressing the right mouse button (see figure 4.9).

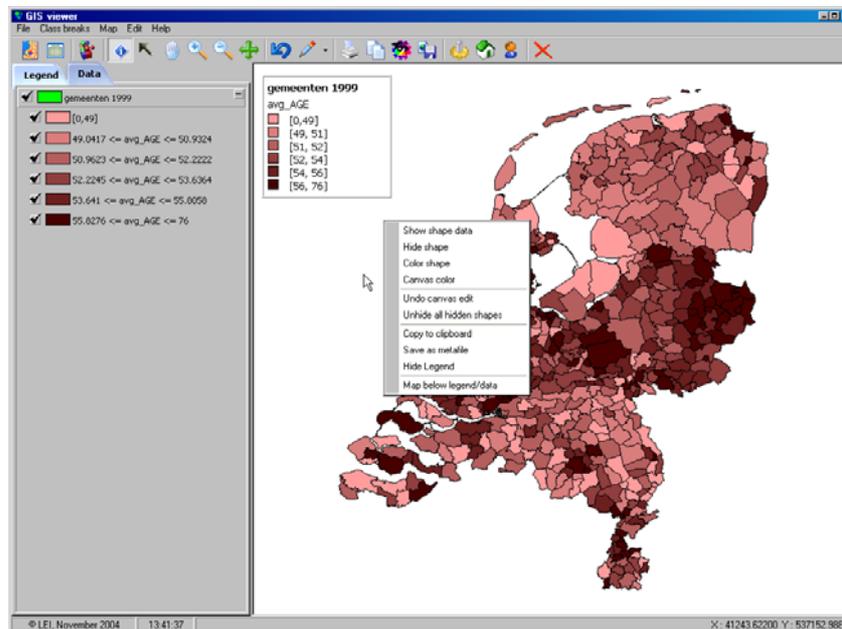


Figure 4.9 Right click to get additional functionality

Available options after right clicking in the map area are:

- |                   |   |
|-------------------|---|
| Show shape data   | Displays additional information of the selected region as displayed in figure 4.8.  |
| Hide shape        | Hides selected shapes. This option provides the opportunity to hide individual regions or parts of the map.   |
| Colour shape      | Gives the opportunity to change the colour of one or more selected regions.   |
| Canvas colour     | To change the colour of the background canvas   |
| Copy to clipboard | Provides the opportunity to copy the map, including the legend, to the clipboard. This option makes it very easy to include a map in a report. The same functionality can be assessed by pressing the  button in the button bar. |

Save as file Provides the opportunity to save the map including the legend into a file. The same functionality can be assessed by pressing the  button in the button bar. The map can be saved as a bitmap, jpg or as a metafile.

Hide legend Hides the legend.

Map below legend/data Changes the position of the map

Available options after right clicking in the legend area are:

Delete selected map Deletes the selected map

Delete all maps Deletes all maps that are open

Default map colour Changes the default map colour

Edit name Changes the name of the selected layer. This name also affects the information displayed in the legend.

Map below legend/data Changes the position of the map

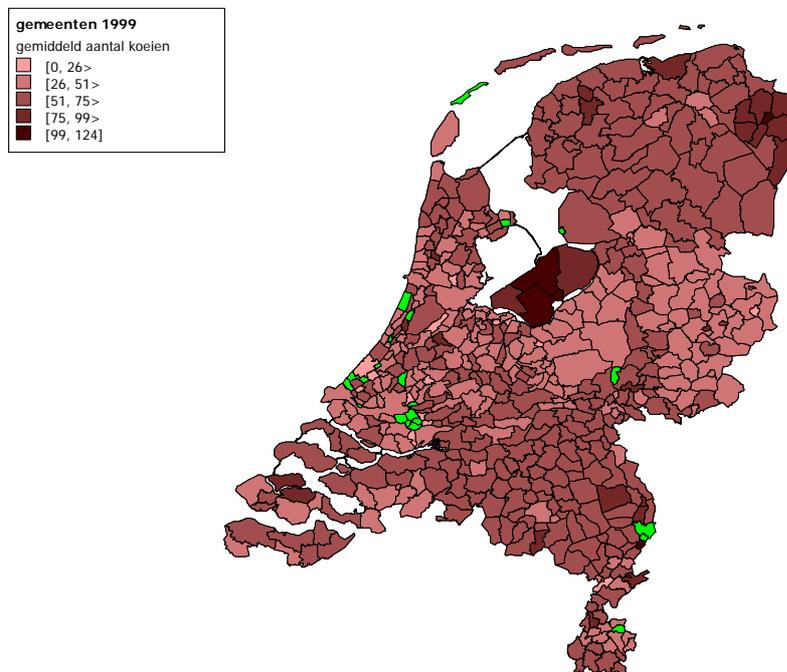


Figure 4.10 Example of average number of cows on dairy farms

## 4.5 Showing results of imputation procedure

In the foregoing we have used GIS to directly display data from data sources such as the agricultural census or the FADN sample. In this section we integrate the functionality of STARS to estimate values for regions and the functionality of the GIS viewer to display regional results in maps.

Figure 4.11 shows the STARS program in which the imputation procedure is displayed (see STARS manual for further details about the imputation procedure). For displaying the results in the GIS viewer it is essential to select a regional variable as grouping variable. In this case, council (municipality) is selected. This means that results are calculated for individual municipalities.

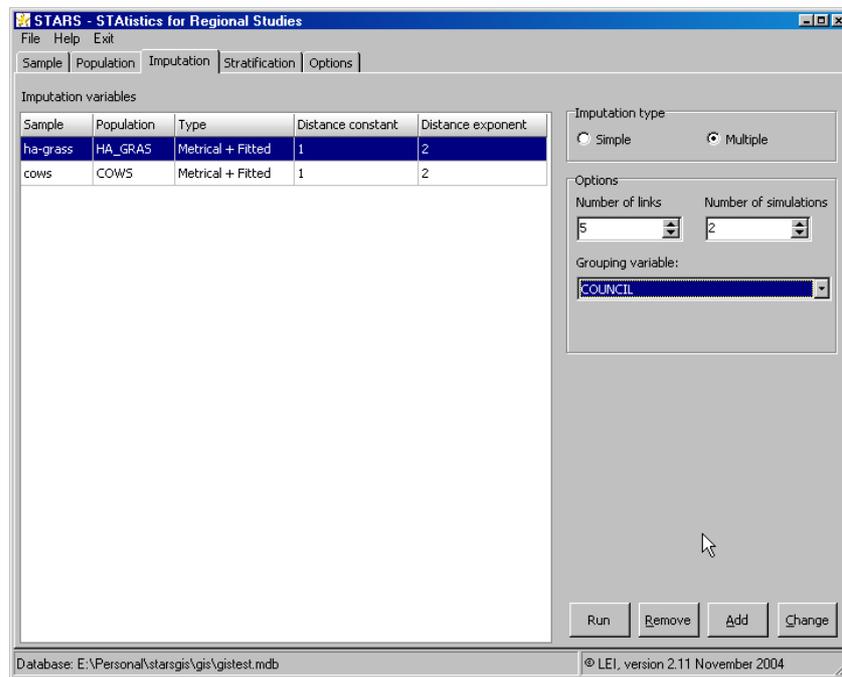


Figure 4.11 Defining imputation procedure and regional level for results

After running the imputation procedure, results are displayed as in figure 4.12. These are the aggregated results for the whole population. Pressing the 'group' button in figure 4.12 will display the group results as displayed in figure 4.13.

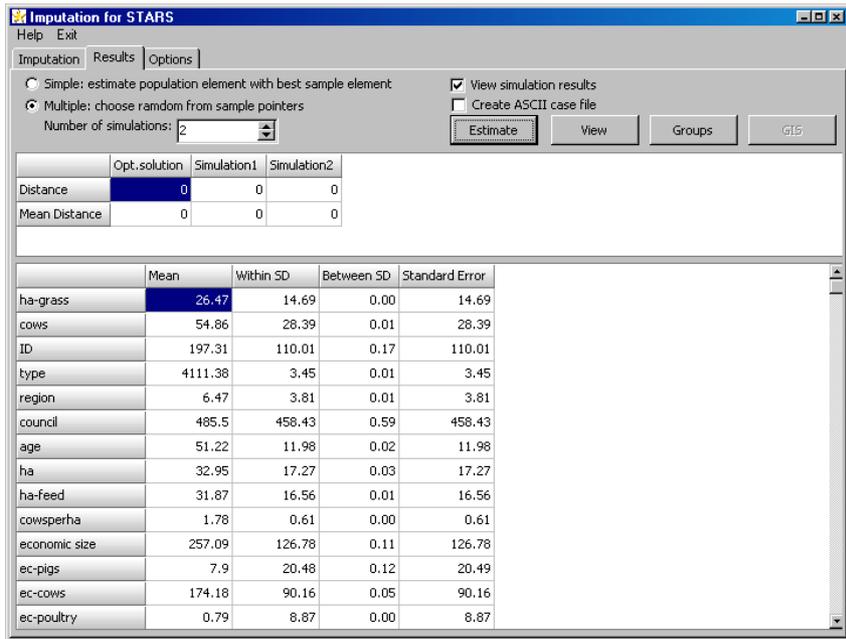


Figure 4.12 Results of imputation procedure

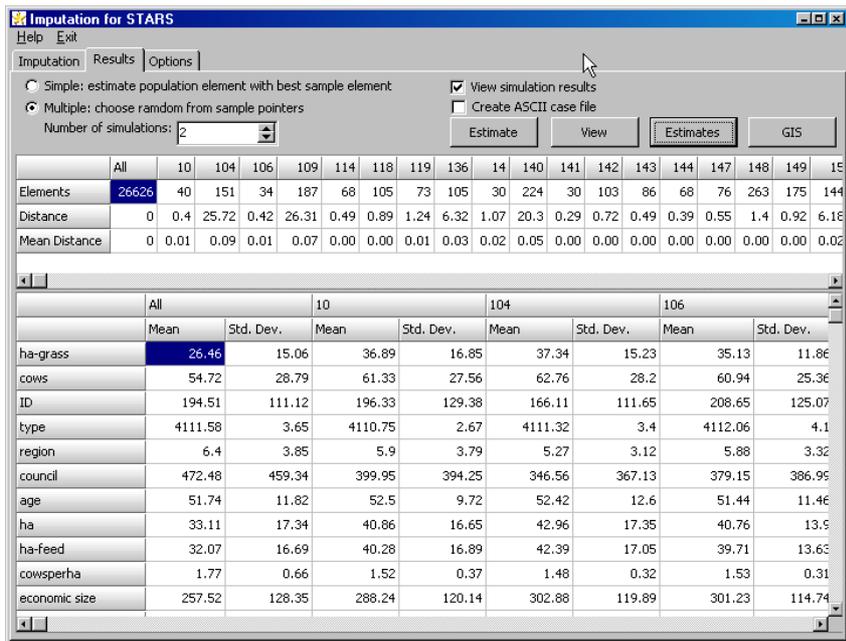


Figure 4.13 Showing results on regional level

After displaying the regional results the GIS button is enabled and can be selected. Before opening the GIS viewer the appropriate information for constructing the map has to be provided.

- Map Name of the map to display the information (in this case the municipalities (gemeenten) in the netherlands). Appendix A describes how the set of available maps can be extended.
- Variable Names of variables to be available in the GIS viewer. One or more variables can be selected.

Subsequently the make button can be selected to forward all necessary information to the GIS viewer.

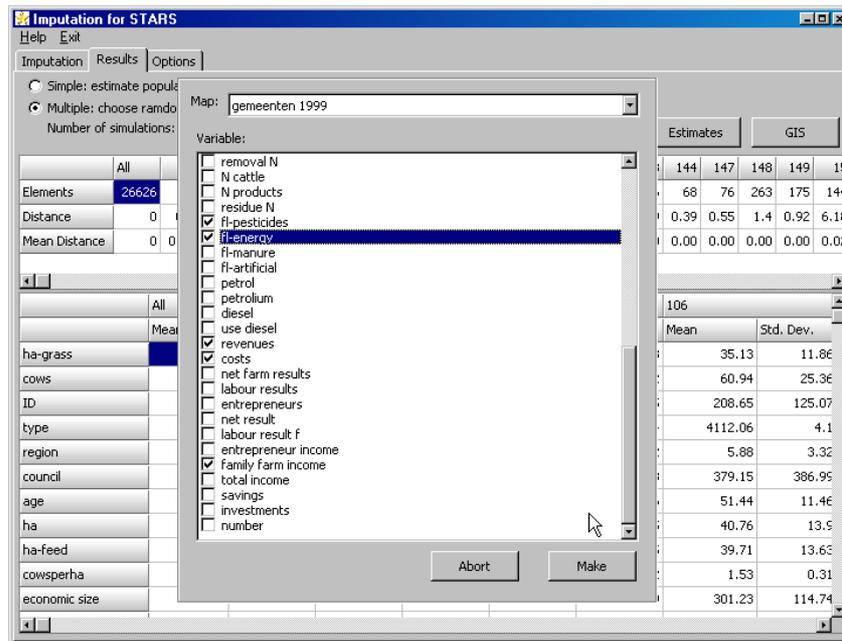


Figure 4.14 Selecting imputation results to forward to GIS viewer

In the GIS viewer the variable has to be selected which has to be displayed in the map. The mean and the standard deviation of all variables selected in figure 4.14 are available for displaying in a map (see figure 4.15).

The example in figure 4.16 is based on the imputation variables 'hectare grass' and 'number of cows' (as shown in figure 4.11). These imputation variables are used to estimate the total revenues of the dairy farm. (This is strong simplification used to illustrate the functionality. A realistic research project should consider other imputation variables to estimate the revenues.)

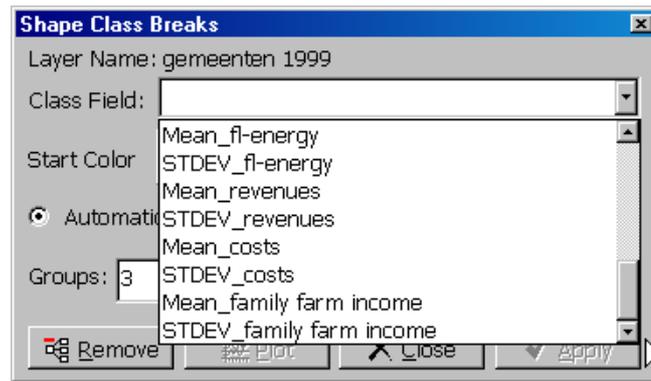


Figure 4.15 Selecting variable to display in the map

The estimated average revenues of specialised dairy farms are shown in the following figure, in which seven classes are distinguished (the classes are defined automatically).

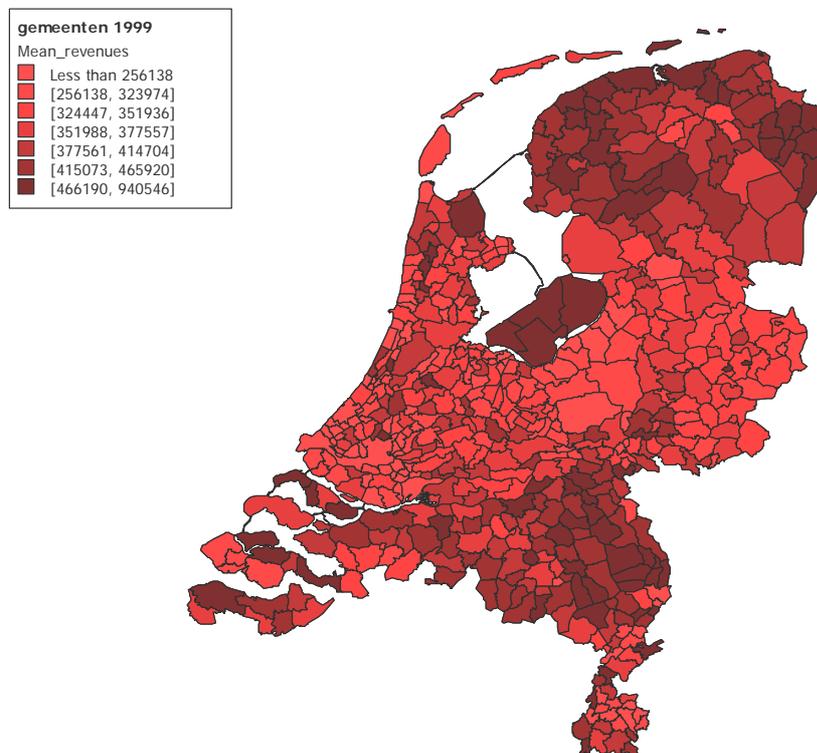


Figure 4.16 Average revenues per municipality based on imputed

The advantage of this approach can be illustrated by comparing figure 4.17 with figure 4.16. Figure 4.17 shows the estimates based on direct observations in the sample.

Given the limited number of observations (less than 300) it is obvious that sample units do not cover all municipalities. No observations are available in the green areas (the default colour of the map). In figure 4.16 estimations are made for the regions with no observations.

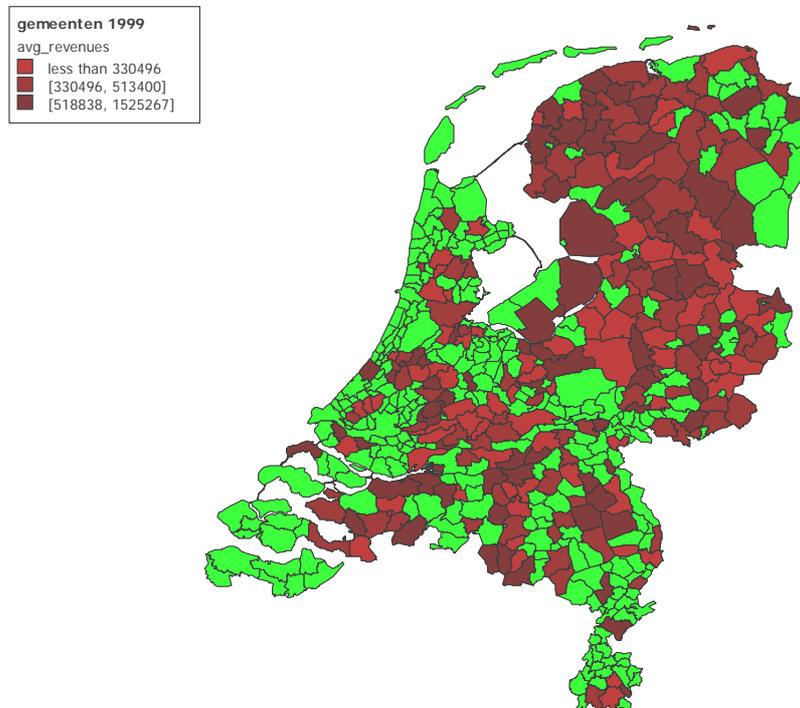


Figure 4.17 Average revenues per municipality (based on direct observations)

#### 4.6 Inspecting data

All information required to show the map in the GIS viewer is transferred to the GIS program. Besides displaying the information in the set and inspecting the data from individual regions, all data can be displayed in the data window (see figure 4.18).

Additional options are available again through right clicking in the data area. These options are:

- |                   |   |
|-------------------|---|
| Blink shape       | After selecting a data field in the dataset, blink shape offers the option to show the matching region in the map.    |
| Highlight shape   | After selecting a data field in the dataset, highlight shape offers the option to give the data field another colour. |
| Zoom to selection | Offers the option to zoom into the map on the region matching the selected data field.                                |

- Add new data Will add a new column in the dataset. Values can be entered from the keyboard or by copy and paste.
- Map below legend/data Changes the position of the map.
- Rename variable Renaming variable will also affect the way the variable is displayed in the legend.
- Sort Shows the data in sorted order.

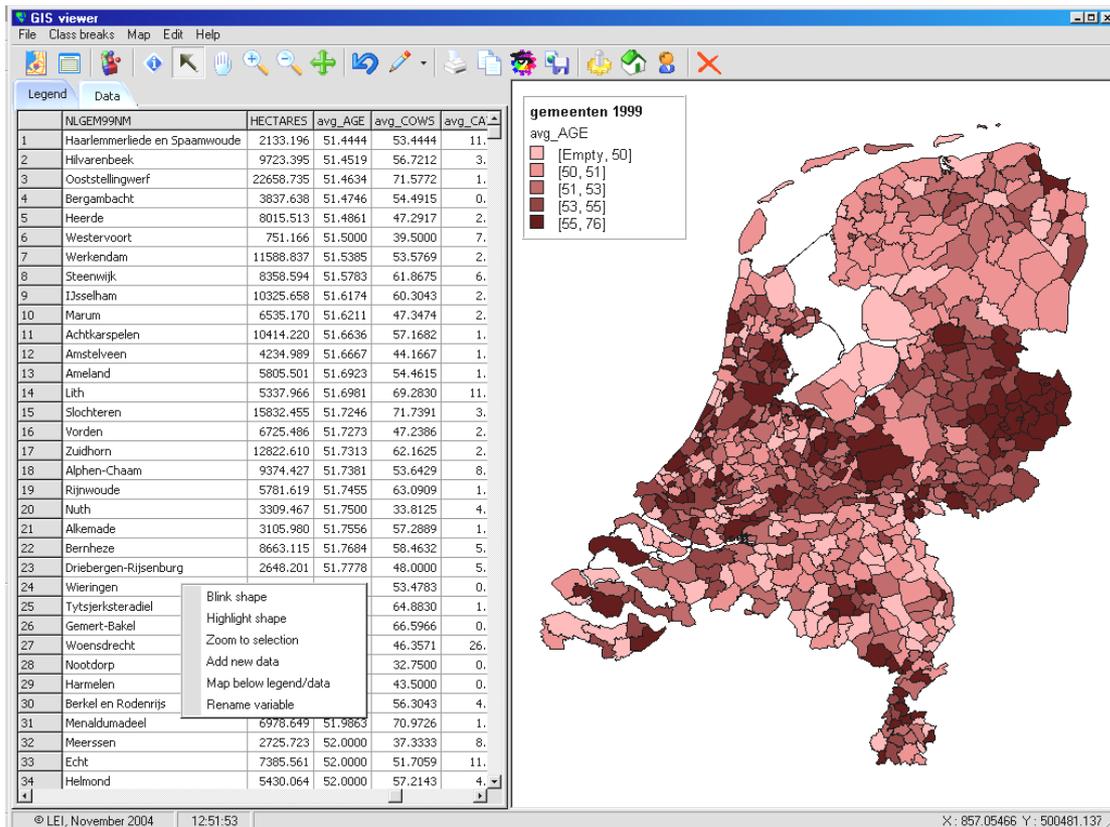


Figure 4.18 Inspecting data

A subset of data can be shown by selecting a layer from the legend (see figure 4.19). Selecting the '51.4444 <= avg\_AGE <= 53.0909' will show only those regions in the data window which belong to this group.

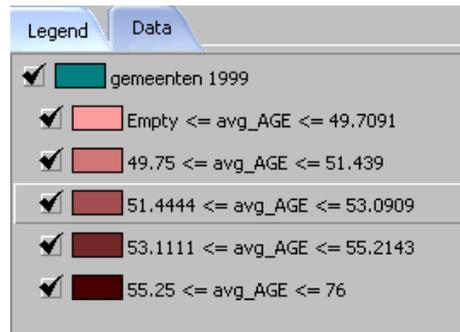


Figure 4.19 Selecting a specific layer/group

#### 4.7 Defining set of available MAPS

Clicking on the  button opens the database of available maps as displayed in figure 4.20. Clicking on a map on the left side of the window will display the matching map at the right side. With the 'add to DB' button a new map can be added to the list.

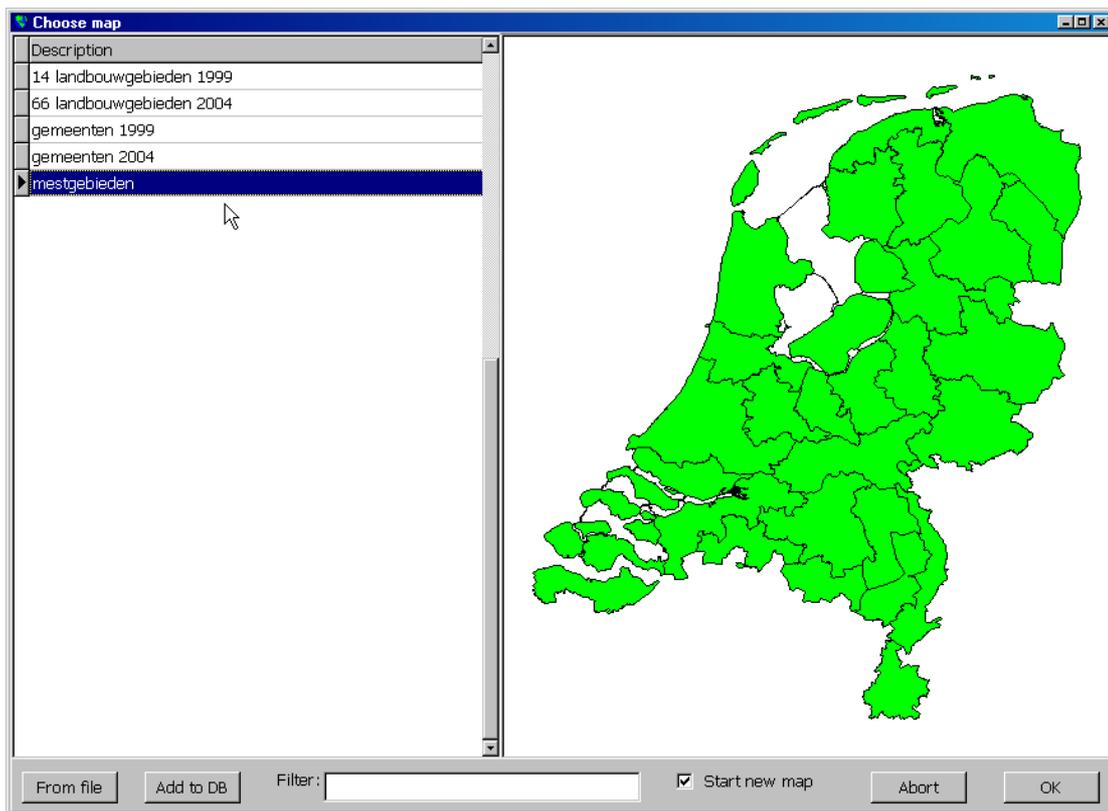


Figure 4.20 Description of available maps

## Literature

Baker, Ken, Paul Harris and John O'Brien, *Data Fusion: An Appraisal and Experimental Evaluation*. Journal of The Market Research Society, 31 (2),152-212, 1994.

Cochran, W.G.C., *Sampling Techniques*. Wiley, New York, 1977.

Dol, W., *Small area estimation, a synthesis between sampling theory and econometrics*. Wolters Noordhoff, Groningen, 1991.

Ford, Barry, *An Overview of Hot-Deck Procedures*. In: Incomplete Data In Sample Surveys, Academic Press, Volume II, Part 2, 185-207, 1983.

Johnston, K., J.M. Verhoef, K. Krivoruchko and N. Lucas, *Using ArcGIS Geostatistical Analyst*. Redlands, Cal.: Environmental Systems Research Institute, 2001.

Keyzer, M.A. and B.G.J.S. Sonneveld, *Using the mollifier method to characterise datasets and models: the case of the universal soil loss equation*. ITC Journal 3/4, pp. 263-270, 1997.

Luijt, J., T. Kuhlman and J. Pilkes, *Agrarische grondprijzen onder stedelijke druk (Agrarian land prices under urban pressure). Stedelijke optiewaarde en agrarische gebruikswaarde afhankelijk van ligging*. The Hague: LEI, Werkdocument 2003/15, 2003.

McCoy, J., and K. Johnston, *Using ArcGIS Spatial Analyst*. Redlands, Cal.: Environmental Systems Research Institute, 2001.

Vrolijk, H.C.J. and M. Wedel, *Een Datafusie-procedure voor het maken van kruistabellen*. In: Jaarboek van de Vereniging voor Marktonderzoek en Informatiemanagement, Uitgeverij de Vrieseborch, Haarlem, pp. 95-106, 1996.

Vrolijk, H.C.J., W. Dol and G. Cotteleer, *Het schatten van kenmerken van kleine deelgebieden*. Rapport 8.02.05, LEI, Den Haag, 2002.

Vrolijk, H.C.J., *STARS: statistics for regional studies, Proceedings of Pacioli 11; New roads for farm accounting and FADN*. Report 8.04.01, LEI, The Hague, 2004.



# Appendix 1 Stars: Statistics for Regional Studies

*Hans Vrolijk, Wietse Dol, Foppe Bouma*

## *Introduction*

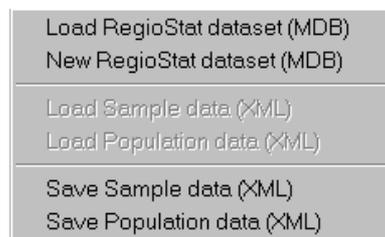
In this appendix, a description is given how to apply Stars in a research project. The following topics are discussed in this chapter: working with data files and inspecting data files; defining the imputation procedure; and finally displaying and analysing the imputation results and the estimates of the goal variables.

## *Working with data files*

A stars project is stored in a Microsoft Access database. In this database the data describing the sample and the population are stored in separate tables. The imputation procedures defined by the user are also stored in the same database so that imputation procedures can be re-used at subsequent occasions.

## *Load and save*

An existing project can be opened by choosing the menu option 'File-Load Regiostat dataset' (see figure B1.1). Using a normal windows dialog box the file can be selected and opened. A new project can be defined by selecting 'New RegioStat dataset'.

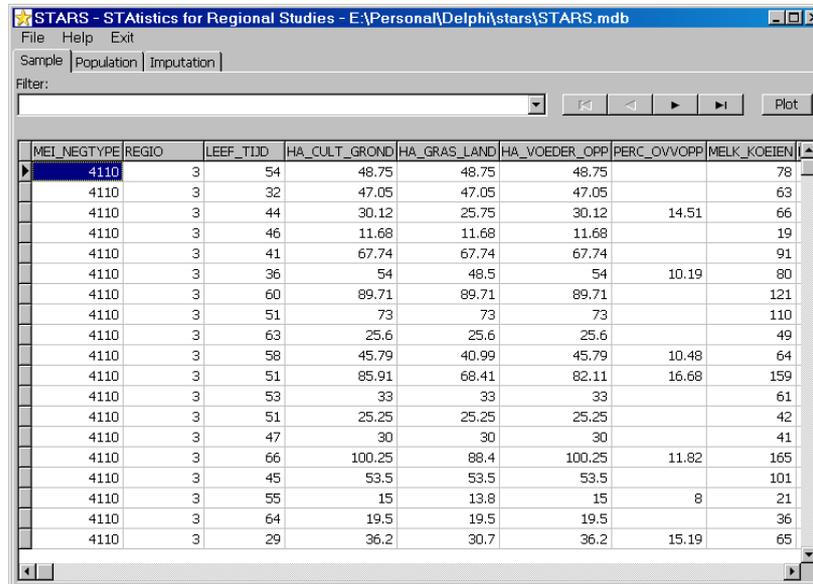


*Figure B1.1 File menu*

User defined changes are immediately stored in the database. Therefore a separate action to save a Regiostat project is not required.

## Viewing data

After loading a Regiostat project the data is displayed in two spreadsheet like forms. The columns display the different variables and the rows represent the cases. In figure B1.2 the sample data are displayed.



MEI_NEGTYPE	REGIO	LEEF_TIJD	HA_CULT_GROND	HA_GRAS_LAND	HA_VOEDER_OPP	PERC_OVVOPP	MELK_KOEIEN
4110	3	54	48.75	48.75	48.75		78
4110	3	32	47.05	47.05	47.05		63
4110	3	44	30.12	25.75	30.12	14.51	66
4110	3	46	11.68	11.68	11.68		19
4110	3	41	67.74	67.74	67.74		91
4110	3	36	54	48.5	54	10.19	80
4110	3	60	89.71	89.71	89.71		121
4110	3	51	73	73	73		110
4110	3	63	25.6	25.6	25.6		49
4110	3	58	45.79	40.99	45.79	10.48	64
4110	3	51	85.91	68.41	82.11	16.68	159
4110	3	53	33	33	33		61
4110	3	51	25.25	25.25	25.25		42
4110	3	47	30	30	30		41
4110	3	66	100.25	88.4	100.25	11.82	165
4110	3	45	53.5	53.5	53.5		101
4110	3	55	15	13.8	15	8	21
4110	3	64	19.5	19.5	19.5		36
4110	3	29	36.2	30.7	36.2	15.19	65

Figure B1.2 Data view of sample

## Plot function: Exploring data

Before using data it is always important that a researcher has an understanding of the data he or she is working with. To support this phase of getting a grasp of the data, an option to plot the data is offered. After selecting the plot button a small window is displayed in which variables can be selected that will be displayed on the axis (see figure B1.3). Besides the user defined variables the system defined variable case number can be selected.

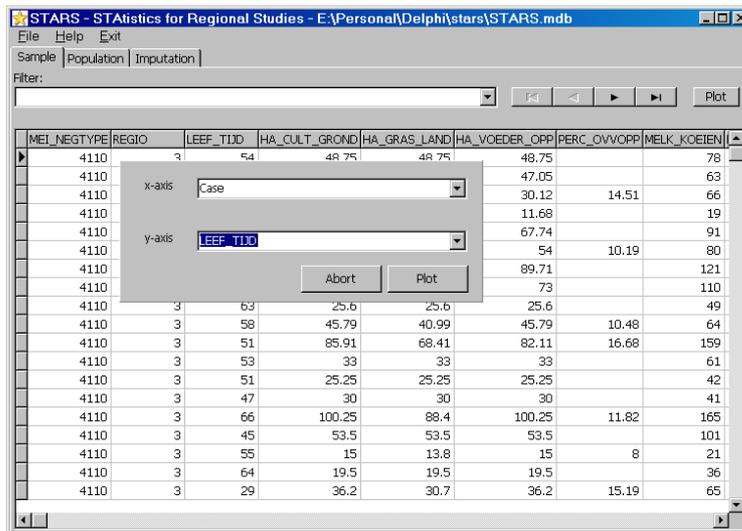


Figure B1.3 Defining a plot

The plot option supports two useful activities:

1. Analysing the distribution of values of one variable.  
By choosing the case number as the X-axis and the user defined variable (for example age) on the Y-axis the distribution can be displayed (see figure B1.4). The researcher can inspect the plot to see the distribution. Comparing the sample plot with the population plot gives an indication whether imputation is feasible. Outliers in the population with no similar farms in the sample might cause problems in the imputation procedure because no resembling farms are available for such outliers.

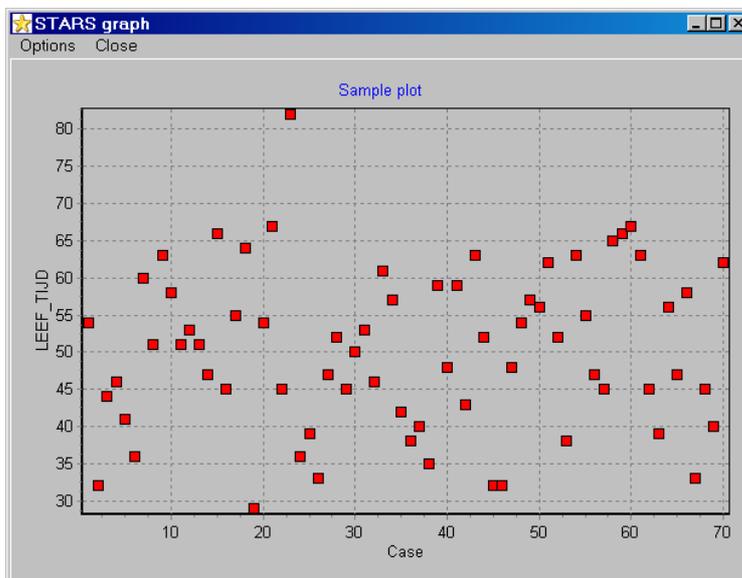


Figure B1.4 Exploring the distribution of a variable

2. Analysing the correlation between variables.  
By choosing two user defined variables the correlation between these variables can be displayed. Analysing the correlation between an auxiliary variable and a goal variable can be useful in selecting the variables which will be used in the imputation procedure. The following example shows the relation between the number of the cows and the total revenues (see figure B1.5). Given the high correlation between both variables, the number of cows could be an important imputation variable for estimating the total revenues.

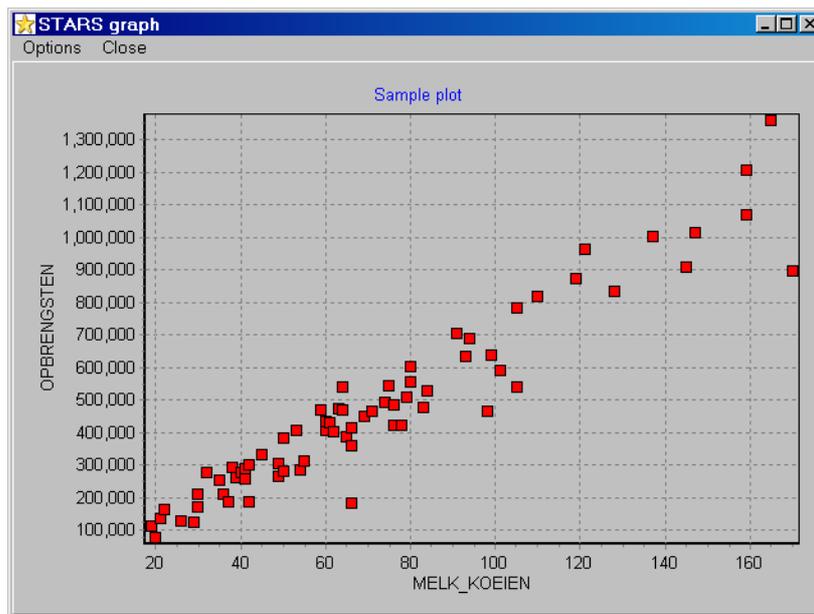


Figure B1.5 Exploring the correlation between variables

### Defining filters

Filters can be defined in data display windows. In the example in figure B1.6 only farms with an owner of more than 50 years old are selected. Subsequent user actions will be performed on the selected cases. For example, a plot will only display the selected cases and an imputation procedure will only take these cases into account.

The syntax of the filters is equal to the Microsoft Access database filters. In filters, conditions can consist of combinations of variables by using AND and OR constructions, for example 'age > 50 AND region = 10' to select farmers older than 50 years in a specific region.

Filters can be applied on both the sample and on the population data. Defining a filter on the sample implies that less farms are available in finding best fitting farms for population farms. Defining a filter on the population implies that results are generated for a smaller region or a smaller group of farms.

STARS - Statistics for Regional Studies - E:\Personal\Delphi\stars\STARS.mdb

File Help Exit

Sample Population Imputation

Filter: leef\_tijd > 50

MEL_NEGTYPE	REGIO	LEEF_TIID	HA_CULT_GROND	HA_GRAS_LAND	HA_VOEDER_OPP	PERC_OVVOPP	MELK_KOEIEN
4110	3	54	48.75	48.75	48.75		78
4110	3	60	89.71	89.71	89.71		121
4110	3	51	73	73	73		110
4110	3	63	25.6	25.6	25.6		49
4110	3	58	45.79	40.99	45.79	10.48	64
4110	3	51	85.91	68.41	82.11	16.68	159
4110	3	53	33	33	33		61
4110	3	51	25.25	25.25	25.25		42
4110	3	66	100.25	88.4	100.25	11.82	165
4110	3	55	15	13.8	15		8
4110	3	64	19.5	19.5	19.5		36
4110	3	54	100	86	100	14	159
4110	3	67	49.5	49.5	49.5		75
4110	3	82	35.74	35.74	35.74		49
4110	3	52	54.99	48.3	54.99	12.17	66
4110	3	53	25	25	25		37
4110	3	61	58.41	58.41	58.41		119
4110	3	57	67	57.4	67	14.33	137
4110	3	59	57.5	57.5	57.5		99

Figure B1.6 Defining a filter on sample or population data

Defining the imputation method

When the sample and population data are loaded the user can explore the data with the plot options. Subsequently, the user can define the imputation procedure to be applied. The imputation procedure can be defined in the screen displayed in figure B1.7. The information on the screen consists of the variables used in the imputation procedure and the general characteristics of the imputation procedure.

STARS - Statistics for Regional Studies - E:\Personal\Delphi\stars\STARS.mdb

File Help Exit

Sample Population Imputation

Imputation variables

Sample	Population	Type	Distance constant	Distance
LEEF_TIID	LEEF_TIID	Metrical + Fitted	1	2
SBE	SBE	Metrical + Fitted	1	2
HA_GRAS_LAND	HA_VOEDER_OPP	Metrical + Fitted	1	2
MELK_KOEIEN	MELK_KOEIEN	Metrical + Fitted	1	2
HA_CULT_GROND	HA_CULT_GROND	Metrical + Fitted	1	2

Imputation type

Simple  Multiple

Options

Number of links: 5

Number of simulations: 5

Grouping variable: -none-

Normalization using whole population

Run Remove Add Change

Figure B1.7 Defining the imputation procedure

On the left side of this screen the imputation variables are displayed. In the separate columns the following information is given for each imputation variable:

Sample	The variable name in the sample
Population	The variable name in the population
Type	Type of fit
Distance constant	Distance constant of the variable
Distance exponent	Distance exponent of the variable

These variables will be explained in more detail in the section Defining Imputation variables.

In the right hand side of the window (figure B1.7) the general characteristics of the imputation procedure can be defined.

#### *Defining imputation type*

Single imputation	For each farm in the population the best fitting farm in the sample is selected. Best fitting is defined based on the imputation variables.
Multiple imputation	For each farm in the population not only the best fitting farm in the sample is selected, but n best fitting farms are selected. N can be defined in the imputation options.

#### *Defining options for multiple imputation*

The following options are only available when 'multiple imputation' is selected.

Number of links	Number of links defines how many best fitting farms are selected for each population farm.
Number of simulations	Number of simulations defines how many simulations are run to make an estimation of the goal variables. In a simulation for each population farm a farm is randomly selected from the list of best fitting farms.

#### *Defining imputation variables*

After selecting Add (or Change) from the window displayed in figure B1.7 the right half side of the screen changes (figure B1.8). In this part imputation variables can be defined and added to the list.

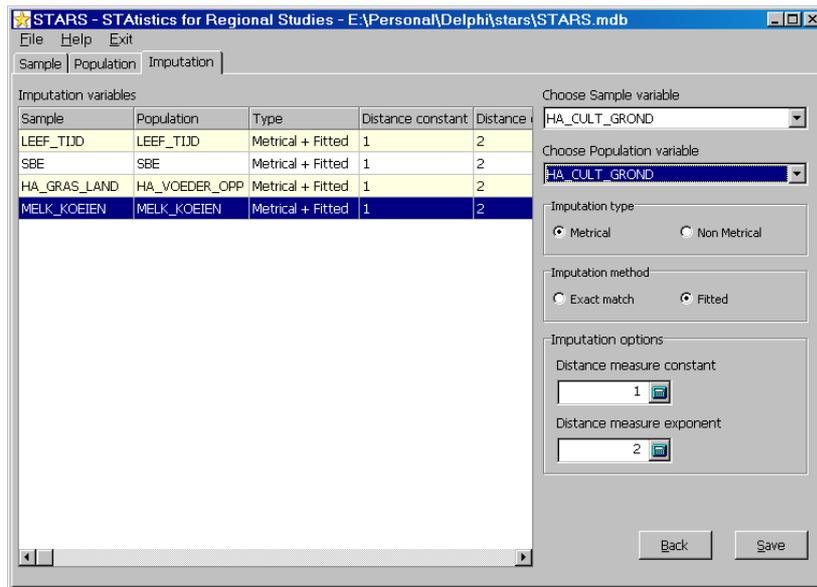


Figure B1.8 Defining imputation variables

To define an imputation variable the following steps have to be taken.

*Selection of variables*

- Population variable      Name of the variable in the population
- Sample variable      Name of the same variable in the sample. The name of the population variable and the sample variable can be the same, but that is not necessarily true. The content of the variable should however be the same.

*Type of variable*

- Metrical      A metrical variable implies a variable on an interval or ratio scale.
- Non metrical      A non metrical variable implies a variable on a nominal or ordinal scale.

*Type of match*

- Exact match      In case of exact match a population farm can only be matched to a sample farm when the values are exactly the same. This matching type is therefore mainly useful when the number of different values is limited. This will often be the case for non-metrical variables.

Fitted In case of a non-metric variable the type of match is exact. Only for metric variable it makes sense to define a distance and to minimise this distance.

### *Distance measures*

Distance measure constant Distance measure constant  $C_i$  gives a weight to the dissimilarity on variable  $i$ .

Distance measure exponent Distance measure exponent  $EXP_i$  determines whether a linear increase in difference between the sample and population farm on a variable result in a proportional increase in distance or in more or less proportional increase.

These two values are the parameters of the equation to determine the distance between a sample farm and a population farm. The distance is calculated as:

**Error! Objects cannot be created from editing field codes.**

in which:

$D_{j,k}$	Distance between sample unit $j$ and population unit $k$
$\alpha_i$	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 $i$
$\beta_i$	Exponent of variable $i$
$j,k$	Unit identifier
$i$	Variable identifier

### *Grouping variable*

A grouping variable can be defined to display the results for different groups separately. For example, in a certain project the researcher tries to estimate variables for a province. Within this province the researcher might also be interested in the means in different municipalities. In this case the population is defined as all farm in the province and municipality can be defined as the grouping variable in order to display results by municipality. The grouping variable does not have any impact on the imputation process itself, it is only used to present the results.

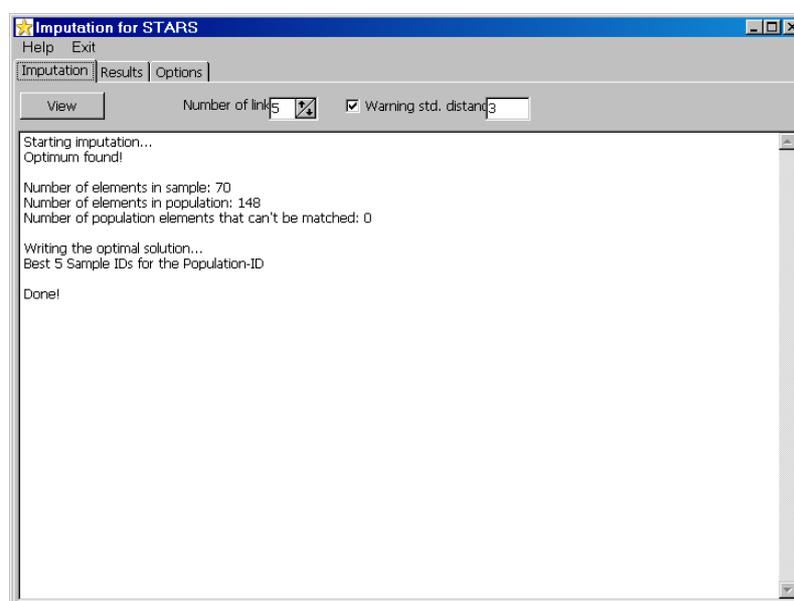
### *Normalisation using whole population*

In case a filter has been defined on the population not all population elements are considered in the imputation process. This also means that the normalisation step is performed on only the selected cases. In some instances it might be more useful to perform the normalisation step on the whole population. For example, when the researcher wants to

look at the results for a special group. In that case the results should be still consistent with the analysis on the population level.

### *Displaying and analysing results*

After selecting the button to run the imputation procedure some details about the imputation procedure are displayed (see figure B1.9). The details listed are the number of sample farms included in the procedure, the number of population farm for which best matching farms were searched, and the number of population farms that could not be matched. Non-matching farms should be a clear warning to the researcher to analyse the imputation process. Especially imputation variables that have to be matched exactly can cause non-matches.



*Figure B1.9 Details of imputation procedure*

Besides these characteristics of the imputation procedure, the researcher can look into the details of the matches. By clicking the view button (figure B1.9) the researcher can choose one of the following options:

- |             |  |
|-------------|--|
| ID's        | For each population farm the ID's of the sample farms are displayed which most closely match the population farm (see figure B1.10).     |
| Distances   | For each population farm the distances to the best fitting farms are displayed.  |
| Non-matches | The ID's of the population farms can be displayed for which no fit could be realised. This option is only relevant if non-matches occur. |

	Link1	Link2	Link3	Link4	Link5
1	55	4	17	47	69
2	16	5	27	26	6
3	32	12	44	55	35
4	44	32	10	55	12
5	44	32	12	55	35
6	32	44	35	42	12
7	21	28	1	10	66
8	60	43	23	61	9
9	64	50	31	13	14
10	61	66	10	41	1
11	8	68	27	34	5
12	10	61	44	66	55
13	38	14	37	35	22
14	6	16	63	29	53
15	53	42	25	63	29
16	46	36	19	35	25
17	55	12	43	10	50
18	19	46	2	36	25
19	29	6	42	53	25
20	13	31	14	64	50
21	7	70	8	54	34

Figure B1.10 Id's of most similar sample farms

View results

Single or multiple imputation

- Single imputation For each farm in the population the best fitting farm in the sample is selected. Best fitting is defined based on the imputation variables.
- Multiple imputation For each farm in the population not only the best fitting farm in the sample is selected, but n best fitting farms are selected. N can be defined in the imputation options.

	Opt. solution	Simulation1	Simulation2	Simulation3	Simulation4	Simulation5
Distance	55.46	119.02	127.47	106.83	126.81	119.27
Mean Distance	0.07	0.16	0.17	0.14	0.17	0.16

	Mean	Within SD	Between SI	Standard Error
LEEF_TIJD	51.18	10.02	0.22	10.02
SBE	303.7	125.49	3.95	125.57
HA_GRAS_LAND	37.84	15.69	0.39	15.7
MELK_KOEIEN	63.36	27.12	0.96	27.14
HA_CULT_GROUND	39.97	16.53	0.29	16.53
MEI_NEGTYPE	4110	0	0	0
REGIO	3	0	0	0
HA_VOEDER_OPP	39.95	16.51	0.3	16.52
PERC_OVVOPP	4.99	7.58	0.47	7.6
MELK_KOEIEN_PERHA	1.59	0.23	0.02	0.24
SBE_VARKENS	0.32	1.97	0.13	1.97
PERC_MELKVEE	70.82	3.69	0.19	3.7
PERC_OVERIG_WEIDE_V	0.94	1.71	0.12	1.71
PERC_TELLEN	0.12	0.08	0.05	0.08

Figure B1.11 Results of imputation

### *Checkboxes*

- View simulation results    Selecting the view simulation results check box enables to view the results of separate simulations in a multiple imputation process.
- Display totals              Default the mean values after imputation are displayed. By selecting The Display totals check box the totals can be displayed instead of the means.

### *Buttons*

Selecting one of the following buttons results in output screens. The 3 buttons are shortly described and subsequently a more extensive interpretation of the output screens is given.

- Estimate                      Do conduct the estimation process with a single or multiple imputation as selected by the user.
- View                            Clicking the view button displays the results of the individual simulation runs.
- Groups                         Clicking the groups button displays the mean results for separate groups. This button is only available when a grouping variable is selected in figure B1.7. If totals are required the display totals check box should be selected.

### *Interpretation of output screens*

After selecting the estimate button, the upper half of the output screen gives the degree of fit in the different simulations (see figure B1.11). The characteristics given are:

- Distance                      Summated distance of all matches.
- Mean Distance                Average distance. This value is equal to the summated distance divided by the number of imputation variables and the number of population farms.

The lower half of the output screen gives the estimations for the goal variables. The results are displayed for the imputation variables and the goal variables. The data provided depends on whether single or multiple imputation is selected.

In case of single imputation, 2 statistics are given for each variable:

- Mean                            The mean in the population calculated based on the imputed values.
- Within SD                      Standard deviation of the imputed values for each farm in the population.

For imputation variables the next 2 statistics are also given:

- Mean Population      The mean of the true values in the population. Imputation variables are known for all cases in the population. This enables a comparison between the real values and the imputed values. This comparison can be used in a verification of the imputation procedure.
- Within SD Population      Standard deviation of the true values in the population.

In case of multiple imputation, 4 statistics are given for each variable:

- Mean      The mean in the population calculated based on the imputed values. The reported values is the average of the separate simulations.
- Within SD      Standard deviation of the imputed values for each farm in the population. The reported values is the average of the separate simulations.
- Between SD      Standard deviation of the series of means. This value gives an indication of the stability of the estimations of the mean.
- Standard Error      Standard error of the mean.

*View*

Selecting the view button displays the results of the individual simulation runs (see figure B1.12). The reported statistics are:

- Mean      The mean in the population calculated based on the imputed values in a specific simulation.
- Std. Dev.      Standard deviation of the imputed values for each farm in the population in a specific simulation.

		LEEF_TIDC	SBE	HA_GRAS_LANI	MELK_KOEIEI	HA_CULT_GRON	MEL_NEGTYP	REGIO	HA_VOEDER_OF	PERC_O
Opt	Mean	51.48	303.1	38.86	62.49	40.56	4110	3	40.56	
	St.Dev.	635.66	3966.19	509.93	822.64	530.65	49829.95	36.37	530.61	
1	Mean	51.36	303.1	38.28	63.34	40.25	4110	3	40.25	
	St.Dev.	634.4	3973.97	504.22	834.17	528	49829.95	36.37	528	
2	Mean	51.18	302.12	38.08	63.22	39.87	4110	3	39.85	
	St.Dev.	631.74	3940.33	500.24	828.01	522.24	49829.95	36.37	522.15	1
3	Mean	51.41	301.98	37.69	63.03	39.88	4110	3	39.87	
	St.Dev.	635.02	3954.5	492.94	831.32	523.66	49829.95	36.37	523.6	1
4	Mean	50.86	300.71	37.27	62.3	39.59	4110	3	39.54	
	St.Dev.	629.11	3960.7	489.81	825.28	519.38	49829.95	36.37	518.15	1
5	Mean	51.09	310.6	37.89	64.92	40.27	4110	3	40.26	
	St.Dev.	631.04	4090.21	496.11	859.28	528.77	49829.95	36.37	528.71	1

Figure B1.12 Results per simulation

## Groups

Clicking the groups button displays the mean results for separate groups (see figure B1.13). In the upper half of the screen the following information is displayed:

Elements	Number of element per group. In the heading the number of the group is displayed. Each value in the group is considered as a separate group.
Distance	Summated distance of all farms.
Mean Distance	Average distance. This value is equal to the summated distance divided by the number of imputation variables and the number of population farms.

	All	1666	1672	1673	1676	171	226	228	384	397	453	483	484	492	493	495	496
Elements	2850	179	254	1	1	1	1	1	1	1	2	1	34	67	233	208	1
Distance	11754.01	3189.35	380.55	0.48	2.13	1.41	0	0.07	0.16	0	0.02	0.00	37.47	583.07	1458.4	244.36	3.0
Mean Distance	0.37	1.62	0.32	0.04	0.19	0.13	0	0.01	0.01	0	0	0	0.1	0.93	0.57	0.11	0.0

	All	1666		1672		1673		1676		171		
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
opp	7.99	14.42	11.29	20.43	10.47	17.87	-	-	-	-	-	-
nge	129.14	133.64	182.35	169	74.66	62.12	-	-	-	-	-	-
ha_hak_mei	0.81	4.61	3.73	10.54	1.64	5.82	-	-	-	-	-	-
ha_groent	0.21	1.36	0.85	2.7	0.6	2.31	-	-	-	-	-	-
ha_boomkwekerij_r	0.28	0.69	0.03	0.23	0.81	1.01	-	-	-	-	-	-
koe_mei	7.39	19.78	0.6	6.63	7.76	22.56	-	-	-	-	-	-
fokz_mei	0.95	11.71	1.79	16.86	0	0	-	-	-	-	-	-
vleesva_mei	5.66	50.47	9.99	80.08	3.39	47.39	-	-	-	-	-	-
vleesk_mei	0	0	0	0	0	0	-	-	-	-	-	-
legh_mei	27.65	584.71	13.69	183.12	0	0	-	-	-	-	-	-
vleesy_mei	0.68	4.55	0.43	3.54	0.45	3.71	-	-	-	-	-	-

Figure B1.13 Result by group

In the lower half of the screen the estimates of the mean and standard deviation for the total population are displayed in the first two columns, in a similar way as in figure B1.11. In the next columns the mean and standard deviation are displayed for each separate group. If the number of elements in a group is less than 2, no results are displayed.

## Validation of results

As described in section 2 the basic assumption underlying data imputation is that if the farm is similar on the imputation characteristics, then it is likely that the farm is also similar on the goal variables. This makes the selection of the imputation variables an essential step. Besides theoretical ideas about the dependency between imputation

variables and goal variables, the plot option provides a helpful tool in exploring the distribution of variables and identifying relations between variables.

Stars provides a number of statistics to help in judging the validity of results.

**Mean Distance** Average distance. This value is equal to the summated distance divided by the number of imputation variables and the number of population farms.

**Identifying 'bad' best matches** In figure B1.9 a checkbox is available to highlight matches where the mean distance is more than a predefined number of standard units.

**Imputed mean vs. population mean** In figure B1.11 the results of the imputation process are displayed. If single imputation is selected, then the estimates for the imputed values are displayed together with the population averages for the imputation variables. The mean of the imputed values together with the standard deviation can be used to test whether the value is significantly different from the population average.

**Stability of results** The between standard deviation is an indicator for the stability of the results. The means of 95% of the simulations are between the mean and plus or minus two times the standard deviation.