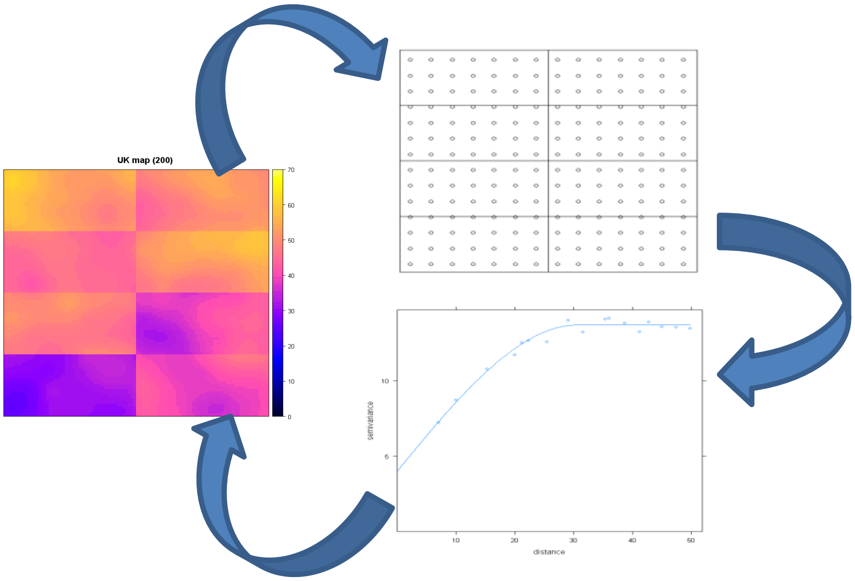
Centre for Geo-Information

Thesis Report GIRS-2011

Interpolation and Infence methods for the Learning Map

Enhancing the learning experience of students with a real-time sampling interpretation system

Menno van der Veen



9-6-2011

WAGENINGEN UNIVERSITY
WAGENINGEN UR

Interpolation and Infence methods for the Learning Map

Enhancing the learning experience of students with a real-time sampling interpretation system

Menno van der Veen

Registration number 88 07 25 856 010

Supervisor:

Dr.Ir. Sytze de Bruin

A thesis submitted in partial fulfilment of the degree of Master of Science
at Wageningen University and Research Centre,
The Netherlands.

9-6-2011
Wageningen, The Netherlands

Thesis code number: GRS-80436
Thesis Report: GIRS-2011-11
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing

Foreword

This thesis was written as a part of a SURFnet funded project called the Learning Map. It intends to present the most suitable techniques for real-time intrepretation of sampled spatial data from an educational point of view.

Contents

Summary 1

1. Introduction 2

2. Methods 5

2.1 Learning Map scenarios 5

2.2 Sampling methods 8

2.2.1 Sampling theory 8

2.2.2 Selected sampling methods 11

2.3 Inference and interpolation 15

2.3.1 Mean value estimation 15

2.3.2 Point predictions for mapping purposes 17

2.3.3 Assesment Criteria 24

2.4 Matching scenarios and statistical methods 26

2.4.1 Scenario 1: Spatial mean of study area 26

2.4.2 Scenario 2: Chloropleth map 26

2.4.1 Scenario 3: Continuous map 27

2.4.2 Scenario 4: Continuous map using spatially exhaustive secondary information 27

2.5 Application tools 29

3. Implementation and assessment 30

3.1 Test field 30

3.1.1 Test field construction 30

3.1.2 Scenario 1: Spatial mean of study area 34

3.1.3 Scenario 2: Chloropleth map 40

3.1.4 Scenario 3: Continuous map 41

3.1.5 Scenario 4: Continuous map using spatially exhaustive secondary information 48

3.2 Hoeksche Waard case 60

4. Discussion 65

5. Conclusions 69

6. Recommendations 71

7. References 72

List of figures

Figure 1: Flow chart showing goals and the presence/absence of secondary information that define a scenario. 7
Figure 2 : The expectation of relative suitability (%) for design-based and model-based approaches as a function of spatial resolution averaged over a large number of cases (De Gruijter, et al., 2006). 10
Figure 3: Simplified decision tree for selecting a sampling method 11
Figure 4: Example of a 95% confidence interval around the mean $\overline{x}$ 15
Figure 5: Example of an experimental variogram with a spherical model fitted through the points. Made with the R extension Gstat (Pebesma, 2004: section). 20
Figure 6: Descion tree for sampling strategy; choices are based on goals, restrictions and secondary information 28
Figure 7: Gaussian Random field of 100x100 pixels 31
Figure 8: Layout of polygons and their values 31
Figure 9: Trend along the y-axis with values ranging from 0-25. 32
Figure 10: Complete Test Field; summation of the random field, ... 32
Figure 11: True mean values of strata containing all three layers; random field + polygons +trend, a histogram of values in the Testfield, Semivariogram at all distances and at distances up to 50 with fitted function: nugget 2.23, sill 53.52 and range 48. 33
Figure 12: Example of a Simple Random Sampling pattern 34
Figure 13: Compact Geographical Stratification into 35
Figure 14: Example of Centred Grid Sampling 36
Figure 15: Experimental variogram including 37
Figure 16: Example of a Stratified Random Sampling 38
Figure 17: Mean and 95% confidence intervals for strategies of Scenario 1. 39
Figure 18: Estimated mean of the strata, ... 40
Figure 19: Point prediction map and Kriging variance map, using Centred Grid Sampling with a sample size of 49 and a spherical model with nugget 3.65, psill 50.59 and range 52.53. 43
Figure 20: Point prediction map and Kriging variance map, using Simple Random Sampling with a sample size of 50 and a spherical model with nugget 0, sill 53.60 and range 44.10. 44
Figure 21: Point prediction map and Kriging variance map, using triangular Centred Grid Sampling with a sample size of 56 and a spherical model with nugget 2.88, psill 59.48 and range 59.02. 45
Figure 22: Point prediction map and Kriging variance map, using squared Centred Grid Sampling with a sample size of 100 and a spherical model with nugget 0, psill 56.61 and range 47.44. 46
Figure 23: Point prediction map and Kriging variance map, using squared Centred Grid Sampling with a sample size of 196 and a spherical model with nugget 2.61, psill 54.28 and range 52.76. 47
Figure 24: Point prediction and Kriging variance map with UK, using squared Centred Grid Sampling with a sample size of 49 and a spherical model with nugget 2.17, psill 26.49 and range 28.0. 49
Figure 25 Point prediction map and Kriging variance map with UK, using squared Centred Grid Sampling with a sample size of 100 and a spherical model with nugget 0 , psill 33.60 and range 29.44. 50
Figure 26: Point prediction map and Kriging variance map with UK, using squared Centred Grid Sampling with a sample size of 196 and a spherical model with nugget 1.95 , psill 30.17 and range 29.75. 51
Figure 27: Point prediction map and Kriging variance map with UK (strata +trend), using squared Centred Grid Sampling with a sample size of 49 and a spherical model with nugget 1.6 , psill 10.08 and range 28.76. 52
Figure 28: Point prediction map and Kriging variance map with UK (strata +trend), using squared Centred Grid Sampling with a sample size of 100 and a spherical model with nugget 1.28 , psill 14.13 and range 27.10. 53
Figure 29: Point prediction map and Kriging variance map with UK (strata +trend), using squared Centred Grid Sampling with a sample size of 196 and a spherical model with nugget 4.0 , psill 9.69 and range 30.72. 54
Figure 30 Left; New Test Field without strata. This new field is a summation of figures 6 and 8. It has a global mean of 30.63. Right; semivariogram of the New Test Field. 55

Figure 31: Regression Kriging prediction and variance maps using square Centred Grid Sampling with a sample size of 49. The variogram is modeled using a Spherical function with; nugget 1.23, psill 18.84 and range 31.39. The trend is linear with, intercept at y-axis 16.22 and a slope of 0.28.................................. 57
Figure 32: Regression Kriging prediction and variance maps using square Centred Grid Sampling with a sample size of 100. The variogram is modeled using a Spherical function with; nugget 0.98, psill 18.41 and range 33.14. The trend is linear with, intercept at y-axis 15.78 and a slope of 0.29.................................. 58
Figure 33: Regression Kriging prediction and variance maps using square Centred Grid Sampling with a sample size of 196. The variogram is modeled using a Spherical function with; nugget 3.28, psill 16.78 and range 43.32. The trend is linear with, intercept at y-axis 15.89 and a slope of 0.29.................................. 59
Figure 34: Agricultural “Field K” displaying the boundaries of management zones and all 25 sampling locations. (Heijting, et al., 2010).. 60
Figure 35: Simple Random Sampling pattern for different sample sizes 2,4,6,8,10,15, 20 and 25. 62
Figure 36: Stratified Random Sampling pattern for different sample sizes 2,3,4 and 5 per stratum. 63
Figure 37 Mean and 95% confidence intervals for Simple Random Sampling and Strafied Random Sampling with different sample sizes from case 1. SRS2-SRS20 represent Simple Random Sampling with sample sizes; 2,4,6,8,10,15, 20 and 25. STR10-STR25 represent Stratified Random Sampling with sample sizes; 10,15,20 and 25... 64

List of tables

Table 1: Inference results from Simple Random Sampling 34
Table 2: Inference results from Compact Geographical Stratification, CGS1-CGS3 represent different strata sizes. 35
Table 3: Prediction results from Block Kriging 37
Table 4: Inference results from Stratified Random Sampling 38
Table 5: Performance criteria for methods of scenario 1. Number in between brackets indicates the sample size.. 39
Table 6: Inference results for strata separately using Stratified Random Sampling with a sample size of 6/stratum. The real mean per stratum is given as reference. 40
Table 7: Mean, maximum variance and performance criteria of OK point predictions for 5 different sampling situations: row 1 matches fig. 18, row 2 matches fig. 19, row 3 Matches fig.20, row 4 matches fig.21 and row 5 matches fig. 22. 42
Table 8: Mean, maximum variance and performance criteria for UK (only trend) with 3 different sampling situations: sample size 49 matches Figure 24 , 100 matches Figure 25 and 196. Matches Figure 26......... 48
Table 9: Mean, maximum variance and performance criteria for UK (trend and strata) with 3 different sampling situations: sample size 49 matches Figure 27Figure 24 , 100 matches Figure 28 and 196. Matches Figure 26. 48
Table 10: Regression Kriging results from the Test Field for squared Centred Grid Sampling with sample sizes; 49,100 and 196. 55
Table 11: Mean, maximum variance and performance criteria for Regression Kriging (only trend) on the New Test Field without strata using the y-coordinates to construct a linear model. Sample size 49 corresponds to figure 27, 100 to figure 28 and 196 to figure 29. 56
Table 12: Mean, maximum variance and performance criteria for Regression Kriging (trend and strata) on the (orignal) Test Field using the y-coordinates and the strata to construct a linear model...................... 56
Table 13: Inference results from Simple Random Sampling on the Hoeksche Waard case 62
Table 14: Inference results from Stratified Random Sampling 63

Summary

The objective of this thesis was to find suitable mapping or inference techniques for the Learning Map. From the spatial interpolation exercise in 2004 it was concluded that there are two main approaches in selecting methods for automatic mapping; (1) the use of simple and robust methods applicable to multiple situations, or (2) the use of some kind of decision framework guiding the user towards the implementation of a method more adapted to a specific type situation (EUR, 2005).
As the second approach leaves considerable room for choices by the user and thereby for learning, it was thought to be most useful to the Learning Map.

The first requirement in this approach of building a decision framework was that there had to be a set of circumstances which are of influence on mapping and sampling to base the selection of methods on.

Section 2.1 deals with this question by the formulation four hypothetical "scenarios" which users of the Learning Map may encounter.
It was deliberately chosen to formulate these scenarios quiet broadly, only specifying the goal, possible secondary information and limiting factors for a survey. Thus giving a general direction for the search rather than a detailed case description.
This was done because although the decision framework should lead to a more adapted method, it should also still be simple enough to be used by relatively inexperienced students.

Section 2.2 gives a theoretical description of the relationship between sampling and the way the sampled results can be used in deriving information about a certain variable within an area. Also sampling methods are selected which are to be used in the Learning Map. In section 2.3 inference and interpolation methods are selected which can be used in combination with the sampling methods. In section 2.4 the sampling strategies (sampling method + inference or interpolation method) are matched to the scenarios. Figure 6 summarizes this process and is the decision framework mentioned in the introduction.

The selected strategies are tested on a constructed dataset of which all characteristics are known in section 3 from which results conclusions are made about the educational added value of the proposed methodology. The main conclusion is that the Learning Map in the proposed form does answer to most of the objectives that were stated in the introduction but that significant work is still needed, both in technical as in conceptual terms, in order to come to a working concept version.

1. Introduction

Background

Real time mapping applications are a relatively new development in GIS made possible by the increasing availability of continuously measured spatial data. The added value of real time mapping however has been recognized ever since the development of Geostatistical techniques in the 1960s (Brenning and Dubois, 2008).
This added value can be attributed to the fact that real time mapping allows users to make decisions very quickly based upon the generated results. A particularly interesting and much discussed application of real time mapping is the response to and the management of natural disasters and environmental pollution such as seismic activity, oil spills and toxic waste pollution (Groat, 2004; De Jesus, et al., 2008).
A recent reminder of the fact that we live in a densely populated country where accidents with potentially hazardous pollution do occur, is the fire at a chemical production company at Moerdijk on the 5th of January 2011. Although in this case the impacts remained fairly limited in the sense that no life threatening concentrations of chemicals were released into the air or water, it does indicate the need for tools which enable quick risk assessment.
It is these kinds of situations, combined with the fact that densely inhabited areas are never far away in The Netherlands that make real time mapping applications for emergency management so interesting. One example of such a mapping system for risk management in the Netherlands is Automap (Hiemstra, et al., 2007).
This automated mapping system is designed to use measurements from the National Radioactivity Monitoring network (NRM) to predict the distribution of radioactive pollution throughout the country. Development of similar systems could lead to quicker and more efficient response to environmental emergencies.

Because of the importance of the developments concerning automated mapping within the Geo-information field, it is thought that MGI students should be at least familiar with the concept of automatic mapping. This, together with the need to keep up with growing use of (mobile) technology for educational purposes (MOBIlearn, 2005), has led to the development of a new educational project called The Learning Map (LM).
This project is an initiative of the Wageningen UR centre for Geo-information and it is intended as a new educational tool using mobile technology which can be used by first year MSc students Geo-information science in the course Remote Sensing and GIS integration.

The Learning Map consists of three main components:

1. Data are collected by field measurement and is gathered on a mobile device with a wireless internet connection.
2. This data are then sent to a server where they are analysed and used to create a map of the area or give information about the spatial mean.
3. The map or information of the study area is displayed on the same mobile device used to gather the data.

It is the continually repeating of these three steps that results in a near real time map of the area; The Learning Map.

The users of the learning map are expected to be relatively inexperienced in all of the three elements mentioned above. The educational value of the Learning Map can be maximized by actively involving students in all three steps; data collection, mapping/inference and interpreting results. Also important is showing the added value of real time (web) applications in spatial and environmental issues.

So by using the Learning Map, students (LM objectives);

1.) Are introduced to different types of data collection (sampling) and learn about the consequences this has.
2.) Learn about different ways of using sampled data to generate a result that matches the goal of their survey.
3.) Are able to use data collected by others as to show the advantage of a real time measurement system.
4.) Interpret their own results and relate this to the choices made in sampling and/or inference/interpolation methodology and compare this to the results of other students.

Objective and research questions

The main focus of this thesis is on the 2nd point in the process of the Learning Map as described above, the objective is: *to find suitable mapping (interpolation) or inference techniques that can be used to interpret the sampled data.*

Many mapping and inference techniques exist, all relying on different assumptions about the spatial variation of the target variable in an area and thereby generating different results. It is this diversity and the fact that many methods also require the setting of one or more parameters that makes it difficult to choose a mapping method before the data are collected to base this choice on, as with automated mapping (Brenning and Dubois, 2008).

The difficulty of automated mapping was clearly defined by the spatial interpolation comparison exercise of 2004 (EUR, 2005). In this exercise participants were asked to make automated predictions from two datasets, the results show that not all participants were able to generate a meaningful result and that there is quite some difference in the generated results.

To overcome the difficulty of selecting an appropriate method there are essentially two options (EUR, 2005);

1.) To use a very basic but robust model, enabling users to generate quick results without prior knowledge but with relatively high uncertainty about the accuracy of the predictions. Or;
2.) To use some sort decision tree containing selection criteria. This may lead to better results but does require some form of extra knowledge or the making of assumptions.

In this thesis the second option is explored offering students with a variety of options for spatial surveys. To do this a number of scenarios are developed which function as a theoretical framework, stating the goal and some other characteristics of the survey that are to be considered in the search for suitable interpolation/inference methods. These scenarios form the foundation on which the choice for a certain methodology is based.

The comparison made above with completely automated mapping methods may seem to imply that the goal for the Learning Map is also to create a completely automated system for which only input needs to be generated by sampling.
However applying such a “black box” approach where input automatically leads to output does not lead to an optimal learning experience. Whereas “real” automated mapping systems such as Automap (Hiemstra, et al., 2007) really focus on working without human interference, the Learning Map is, or should be the opposite. Users have to be included in the process to some degree in order to learn about the decisions involved in spatial surveys.

This educational purpose is why the focus is mainly on finding methods which are easy to apply and understand by students.. Selecting such methods means that as much time as possible can be put into sampling and interpreting data rather than on theory.

The research questions of this thesis are:

1. *Which scenarios are relevant for the learning map project given the project background?*
   Considering the possible applications and the educational purpose; what might the user be interested in? What are interesting secondary circumstances to consider? What limitations concerning mapping/inference should the users learn about?
2. *What are criteria on which an interpolation or inference method can be judged?*
3. *Which interpolation or inference methods are most suitable given the different scenarios?*
4. *Can these interpolation methods be implemented for automatic interpolation or inference within a statistical computing environment?*
5. *How do these methods perform judged on the formulated criteria?*

Structure

The structure of the report follows the sequence of the research questions. First, in section 2.1 the concept of the scenarios is explained and worked out in more detail. Section 2.2 deviates from the sequence of the research questions and from the initial scope the research as well. It deals with data acquisition, which is essential to include because data acquisition (sampling) cannot be seen separately from any inference or interpolation method (De Gruijter, et al., 2006).
Section 2.3 lists possible inference and interpolation methods plus several criteria with which they can be assessed.
The sampling methods and interpolation or inference methods are matched to the scenarios according to their suitability for each scenario in section 2.4. In section 3 these strategies, sampling method + inference/interpolation methods, are put to the test in the statistical computing environment R (R Development Core Team, 2010). Finally, in section 4 the results are discussed and conclusions are made.

2. Methods

2.1 *Learning Map scenarios*

According to Brenning and Dubois (2008), every dataset may require a different interpolation method as the "best' method varies according to the purpose of the study and the characteristics of the dataset.

To find methods that are able to process the data and also give an output which is of use to the surveyors requires the matching of the goals of a survey, the known characteristics of the study area and the limitations in terms of time/money or accessibility of the area.

Grouping these circumstances leads to the formulation of what is termed scenarios in this thesis. A scenario is, according to the Oxford Dictionary: "A postulated sequence or development of events" or "a setting", in other words a given set of circumstances. The circumstances taken into account are:

1.) The goal of the survey. This is the most important circumstance; what is it the user wants as an output? It was decided to consider three options with regard to the spatial resolution of the output: (1) The global mean, (2) mean of sub-areas, (3) a detailed map (point predictions).
As there is an endless variety of spatial resolutions possible between the global mean and point predictions as in a continuous map, the mean of sub-areas (up to 1/8 of the total area) was chosen as an intermediate form.
These three outputs are expected to give a good impression about different possibilities for mapping/inference whilst not overwhelming students with options.

2.) The prior knowledge about the area. Is there any secondary information which can facilitate a better or more detailed estimation or prediction? Again there is a whole range of possible types of secondary information including secondary variables, trends and strata (Li and Heap, 2008). To limit the choice in methods somewhat only two different sources of secondary information are used: (1) Strata, (2) a trend in the data.

3.) The limitations in terms of money and time. The most costly phase in a survey is often sampling (De Gruijter, et al., 2006). This is why the limitation considered in in the scenarios and the search for methods is the sample size.
It may well be possible that a mapping method is very suitable for a scenario in terms of the goal and prior knowledge but because it is not affordable to acquire enough data the user is forced to adjust the goal of the survey.

The following section contains a description of four scenarios which are first of all based on the goal of the survey but also take into account the other two points; facilitation or limitation in reaching that goal.

Scenario 1: Spatial mean of a certain property for the whole study area
Two possibilities are considered where the objective is to obtain the spatial mean:

1.) The goal of the surveyor is to obtain the spatial mean with a certain degree of accuracy.
2.) The goal of the surveyor is to obtain as much information about the properties of a certain variable in the area with a very limited budget. The best guess in this case will most likely be an estimate of the spatial mean.

Within this scenario there is also room for the use secondary information like using sampling within strata to come to a more accurate estimate. These options will be explored further in sections 2.2 and 2.3.
A well-known application for spatial mean estimation is temporal monitoring of animal and plant populations within certain areas (Gibbs, et al., 1998).

Scenario 2: Spatial mean of user defined sub-strata within the study area
It is also possible that a surveyor is interested in obtaining information about specific areas within the field but that the required level of detail (i.e. spatial resolution) is not very high.
Sampling and inference can then be done from these more homogenous areas. Resulting in a choropleth map of the area, giving more detail about the area than only the spatial mean.
An example from soil science where these kind of chloropleth or mosaic maps are used is soil-landscape mapping. Here boundaries in soil types or classes are mapped as abruptly changing fields. These boundaries can for instance be detected through visual interpretation by an expert from Remote Sensing images (De Bruin, et al., 1999).

The secondary information used for stratification can be of different kinds. For instance the spatial distribution of a secondary variable with a known relation to the variable that is to be estimated. Or, like the case mentioned above; knowledge exists about where boundaries occur due to certain land uses.

Scenario 3: Continuous map of the area with no secondary information
With this scenario the goal is to produce a continuous map of the whole area using a spatial prediction technique. Taking into account the LM objectives set out in the introduction and the aim to produce a system which allows user interaction. The goal will be to let the users produce a model about the spatial variation themselves.
This approach has as a downside that a sample of sufficient size will have to be acquired which might prove difficult (section 2.3).
An application where field scale continuous maps are needed containing the spatial variation of certain soil parameters is precision agriculture. Resources like nutrients can be far better managed when specific deficit regions are known to the farmer (McBratney and Pringle, 1999).

Scenario 4: Continuous map using spatially exhaustive secondary information

Just as with Scenario 3 the intention here is to create a continuous map of the area. The difference being that in this scenario there is secondary information of some sort that can help improve the result of the interpolation. To limit the search for methods only one type of secondary information will be used as an example; a trend.

A trend is defined as a smooth change in an underlying variable; resulting in a varying mean (Webster and Oliver, 2007). It is assumed that this trend covers the entire study area (=exhaustive data).

Figure 1 gives a schematic overview of the four scenarios. Although a scenario is normally a given set of circumstances where no decisions are to be made, Figure 1 is shaped like a decision tree. This is on the one hand because the goals are part of the scenarios and there is a choice to be made there. On the other hand this figure is also meant as the basis for an interface for users in setting up their survey as it automatically leads relatively inexperienced users with a certain goal in mind to the selection of an accompanying methodology.

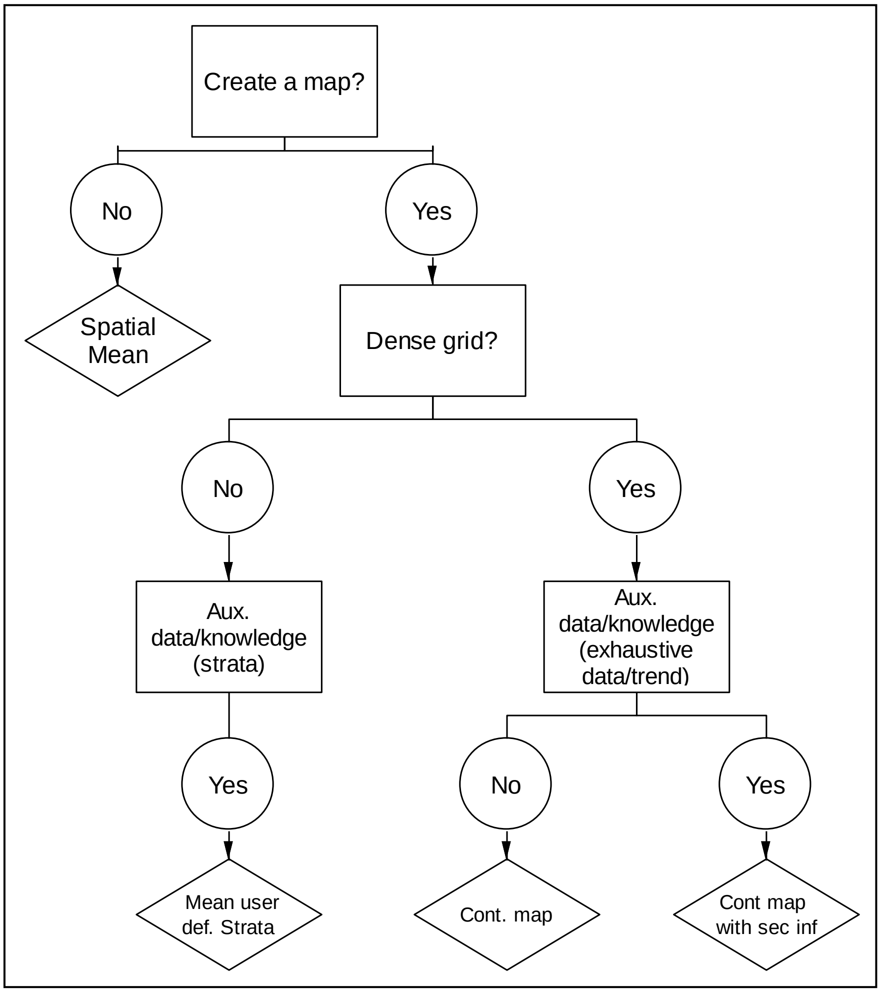


Figure 1: Flow chart showing goals and the presence/absence of secondary information that define a scenario.

2.2 *Sampling methods*

At the basis of each study involving inference or interpolation is the collection of data. The way in which the survey is carried out may have a profound influence on the outcome of any prediction or estimation. This is why this section will deal with some important characteristics of spatial sampling and the influence this has on the type of inference or interpolation. Following this theoretical section is a selection of possible sampling methods for the Learning Map.

2.2.1 Sampling theory

A fairly recent and extensive work on spatial sampling/surveying has been written by De Gruijter, et al. (2006); “Sampling for natural resource monitoring”. This book has been used as a guideline in selecting suitable sampling methods and providing some necessary theoretical background about sampling.
An important division that De Gruijter, et al. (2006) make in sampling is between design based and model based strategies. With strategies being the combination of a sampling method and an inference/interpolation method.
The difference between design and model based strategies is that with a design based approach the selection of sampling units is done randomly, thereby enabling unbiased inference. With model based methods sampling units are or not necessarily randomly selected and inference or interpolation is based on some type of model about the spatial variation in the study area.
One of the most important consequences of using a design based or a model based strategy is that the first allows for unbiased uncertainty estimation while the second does not. This is because a requirement of design based sampling methods is that they have to be p-unbiased, meaning that the average of all possible sample realisations gives the true mean. With model based strategies the uncertainty is not estimated directly from the sample but through the model. Because a model is just our best estimate of the truth repeated realisations are not likely to give the exact truth, therefore the estimate it is not p-unbiased (De Gruijter, et al., 2006).

Brus and De Gruijter (1997) extensively discussed the two approaches. Concluding from this discussion De Gruijter, et al. (2006) give the “ideal” circumstances for both:

Design based

1.) The aim of the survey is to produce an estimate of the frequency distribution of the target variable or a parameter of this distribution. For instance the mean or the standard deviation.
2.) A minimum sample size of 5-10 sampling units can be afforded in order to have an idea of the sampling error.
3.) Sampling can be done at random.
4.) An unbiased estimate is important.
5.) An objective assessment about the uncertainty of the estimate is important.

Model based

1.) The aim of the survey is to produce a prediction at specific points in the area or a prediction of the distribution over the whole area such as with mapping.
2.) A medium sample size can be afforded in order to be able to construct a model to describe the spatial dependence between points. Or:
3.) A reliable model of the variation is available. (not used in this thesis)
4.) Strong autocorrelations exist in within the area.

It can be noticed from the points mentioned above that the desired spatial resolution, i.e. the size of the separate outputs, plays an important role in defining the suitability of a strategy for a certain survey. De Gruijter, et al. (2006) call the separate parts for which an output is wanted domains.

This relation between suitability and spatial resolution (domain size) is shown graphically in Figure 2. As can be seen from this figure the suitability of neither design nor model based strategies is ever 100%. This is because besides just spatial resolution there are also other factors which are of importance. These factors raise questions like whether p-unbiasedness is strictly needed, and whether there is a model available about spatial variation (De Gruijter, et al., 2006).

So although between the ideal situations there is a range of circumstances where both strategies can be applied. Generally it can be said that the suitability of design based methods is greatest for global quantities; these are large domains like the global spatial mean and decreases for smaller domains like strata and eventually point predictions such as gridded values in a continuous map. (see Figure 2)

This division in suitability between global and local quantities relates closely to the different scenarios, where scenarios 1 and 2, the estimation of global and partial spatial means, are more about global quantities (Chapter 7, De Gruijter, et al., 2006) and scenarios 3 and 4 are about local quantities; the prediction at points. (Chapter 8, De Gruijter, et al., 2006).

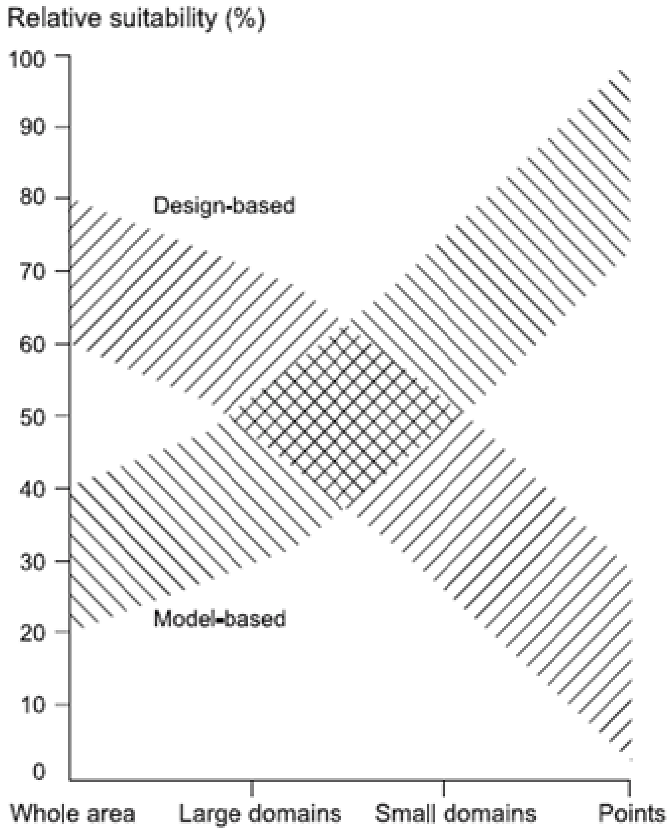


Figure 2 : The expectation of relative suitability (%) for design-based and model-based approaches as a function of spatial resolution averaged over a large number of cases (De Gruijter, et al., 2006).

2.2.2 Selected sampling methods

To aid the spatial/environmental scientist in finding a suitable method De Gruijter, et al. (2006) have created a useful tool: a decision tree for selecting design based methods. The book of de Gruijter, et al. (2006) as a whole and consequently the decision tree as well, was written as a practical guide for surveyors making it a very complete and therefore detailed instrument.

As we only consider a limited amount of cases in this thesis, i.e. the four scenarios, it makes no sense to incorporate the entire decision tree of De Gruijter, et al. (2006) in our own decision framework for the Learning Map. Figure 3 therefore represents a simplified decision tree, containing only the choices that are thought to be applicable to the Learning Map. In the sub-sections following from Figure 3, the sampling methods and the choices leading to those methods are briefly explained.

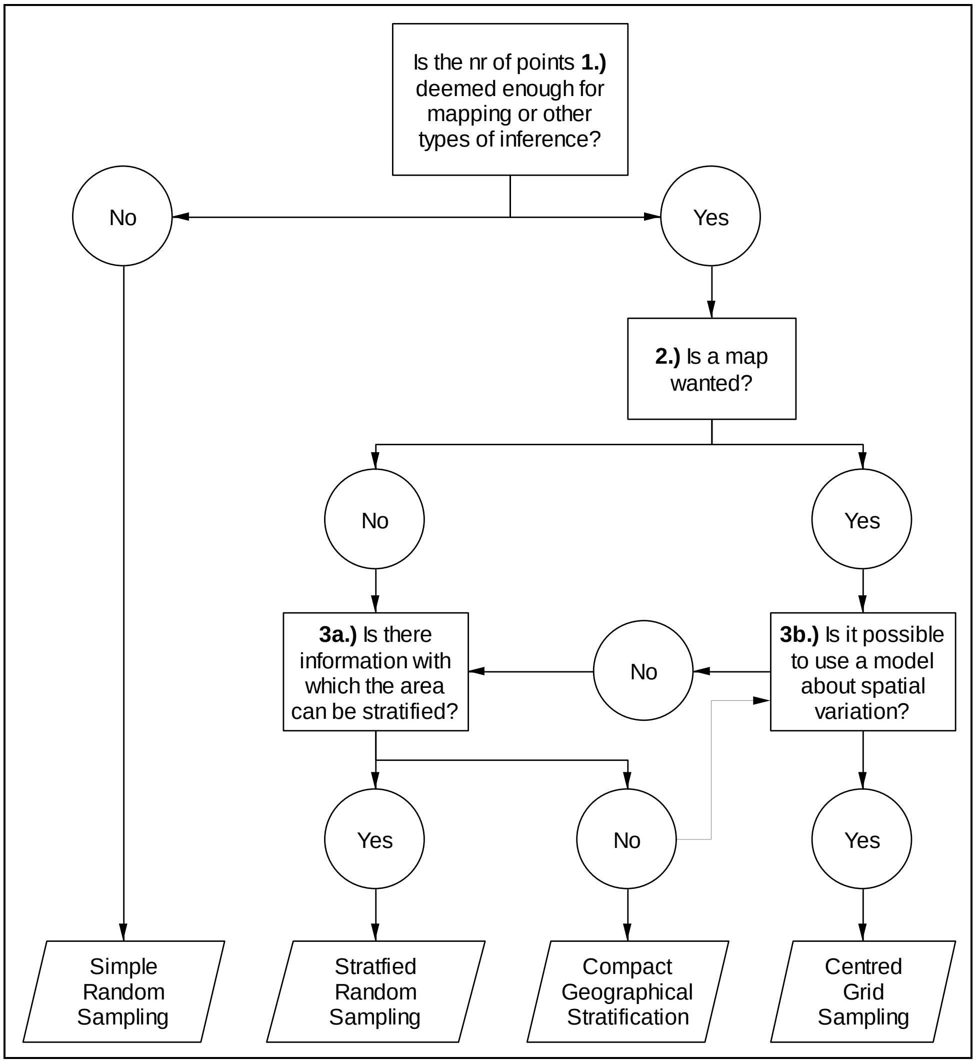

Is the nr of points **1.)** deemed enough for mapping or other types of inference?
No
Yes
2.) Is a map wanted?
No
Yes
3a.) Is there information with which the area can be stratified?
No
3b.) Is it possible to use a model about spatial variation?
Yes
No
Yes
Simple Random Sampling
Stratfied Random Sampling
Compact Geographical Stratification
Centred Grid Sampling

Figure 3: Simplified decision tree for selecting a sampling method

Simple Random Sampling

This design is most likely the simplest form of sampling; it only involves randomly selecting points from a distribution of coordinates and checking whether they actually lie within the area. The only attribute that is to be chosen in this design type is the sample size.

Although the simplicity of Simple Random Sampling is an advantage it is also very inefficient compared to other sampling methods, especially for larger sample sizes (De Gruijter, et al., 2006).

This inefficiency is twofold; (1) The spatial coverage may be poor as certain areas could be underrepresented, possibly leading to a distorted outcome. For instance in the case of inference of the spatial mean, the mean may be under or overestimated due to poor coverage. This also means that it is often the case that the variance of the mean, which is used as a measure of uncertainty in mean estimation, can be decreased for the same sample size by using a different sampling method.

The other inefficiency (2) is of practical nature as the irregular placement may lead to a more time consuming survey due to travel time.

This is why Simple Random Sampling is usually only applied when just a small sample size can be afforded. This results in answering negatively to question 1 in Figure 3, which corresponds to the second option for scenario 1 (see section 2.1).

The formulation of question one, needs some explanation as it introduces two important concepts; mapping and other types of inference. Both of these concepts will be worked out further section 2.3, the important thing here is that for these mapping or other inference types larger sample sizes are needed. For instance to sample regularly at certain distances or to cover all strata.

Regrettably no clear-cut limit in sample size can be given in Figure 3 instead of question 1. Although De Gruijter, et al. (2006) use a sample size of 30 or less as a rule of thumb in their decision tree. This choice is debatable because the needed sample size for both mapping as other types of inference really depends on the size and spatial variability of the area in question.

Stratified Random Sampling

When there is information with which the area can be divided up into clearly different parts in the sense of mean value, variance or cost of measurement, stratification could be a good way to improve the precision of a survey without increasing the total sample size. Different methods of stratification exist and globally they can be divided into two groups (De Gruijter, et al., 2006):

- Stratification based on a classified ancillary variable, for instance an existing soil or other thematic map.
- Stratification based on a quantitative ancillary variable with a known relationship to the target variable by cluster analyses.

The first method is fairly straight forward; the boundaries of the strata can be formed by the classes of an earlier map. Options are existing maps like land use, soil or elevation but also a division into (agricultural) management zones is possible.

The second method is a bit more complicated; A much used clustering method is the K-means algorithm which involves using the known ancillary variable to form clusters. The area is discretized into objects in accordance with the ancillary variable and thus forms the attribute space. The clustering procedure is an iterative process which involves the following steps:

1.) Choose k points in the attribute space which is formed by the ancillary variable. These points will form the initial group centroids (multivariate means).
2.) Assign each object to the group with the nearest centroid.
3.) Recalculate the centroids for the groups
4.) Repeat steps 2 and 3 until no objects are transferred any more (De Gruijter, et al., 2006).

Compact Geographical Stratification

One way to achieve good spatial coverage thus capturing the large scale spatial variation. Presumably leading to a lower variance without using non-random sampling techniques (which mean using a model based inference method) is by using Compact Geographical Stratification (Walvoort, et al., 2010).
This type of stratification divides the area when no ancillary variable or thematic map is available. This is done through a k-means clustering algorithm, in which the cells of a fine discrete grid are used as objects and the geographical coordinates of the midpoints of the cells are used as the classification variable (Walvoort, et al., 2010).
First an initial stratification is made based upon a number of prior sample locations or points are selected for full coverage, then the distance (in this case Mean Squared Shortest Distance) of each cell relative to centroid of these clusters is calculated thus (re)assigning or swapping the cells between the clusters. This is done through the R package spcosa (Walvoort, et al., 2010), which is described in section 2.5.

In Figure 3 it can be seen that the user arrives at this sampling type through answering no to question 2 and 3a which means the goal of the survey is not mapping and that there is no way of stratifying the area based on secondary information. Commonly the method is used in such a way that the number of strata is maximized and so also maximizing the spatial coverage. This means the number of strata is equal to the sample size divided by two, because a minimum of two samples per stratum is needed to estimate the variance. The idea behind this choice is that when it is not possible to stratify the area it is good practice to cover the area as good as possible (Walvoort, et al., 2010).

Centred Grid Sampling

Another way of obtaining good spatial coverage is through Centred Grid Sampling. Centred Grid Sampling is a fairly simple sampling method to carry out, the procedure is to place a regular grid over the area in such a way that it is expected to capture the spatial variation as good as possible. This deliberate placing makes it a sampling method which is necessarily part of a model based strategy as there is no randomness in it to base the accuracy estimation on; it is p-biased (De Gruijter, et al., 2006).

For mapping purposes this is not a problem because according to Figure 2 the suitability of model based methods for small domains as with mapping is almost 100%. This is why answering yes to question 3b in Figure 3 leads to Centred Grid Sampling.

Although model based methods are not preferred in case the required output is a large domain like the spatial mean, it is possible. This possibility is expresses by the dotted line. The path followed in this particular case would be; 1.) Yes, 2.) No, 3a.) No 3.b) Yes.

This last question (3b) from Figure 3 implies two things:

1. That there is a model about spatial variation available or a sufficient sample size (section 2.3.2) can be taken in order to construct such a model.
2. It is not necessary to obtain a p-unbiased estimate of the uncertainty like with design based methods.

Different shapes and sizes of grids are possible, much used shapes are triangular, hexagonal and square grids. Triangular grids are supposedly most efficient as they minimize the underrepresented area for a certain sample size (De Gruijter, et al., 2006). The shapes mainly influence the uncertainty where it concerns point predictions. Maximum uncertainty in the prediction of points occurs at locations farthest away from sampling locations.

Webster and Oliver (2007) argue that although triangular grids give the optimal spacing leading to lower maximum uncertainty (variance), the difference with square grids is so small that often square grids are preferred because they are easier to work with.

2.3 *Inference and interpolation*

The aim of this section is twofold; (1) to give techniques with which mean values or point prediction can be calculated for different sampling methods and (2) to give measures with which those outcomes can be compared.
In the equations the following notation will be used; $\bar{z}$ for average, $\hat{z}$ for estimate and $\tilde{z}$ for a prediction of a variable.

2.3.1 Mean value estimation

Simple Random Sampling

Obtaining the mean value of a population or area through Simple Random Sampling is done by estimation; the mean of the total area, $\hat{\mu}$,is estimated through calculating the mean of the sample, this is done according to;

$$\hat{\mu} = \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$ **Equation 1**

Where $\bar{z}$ is the sample mean and *n* is the sample size. Because the sampling method is p-unbiased, the variance of the mean of variable *z* in the area can be estimated by (De Gruijter et al., 2006):

$$\hat{\sigma}^2(\bar{z}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}(z_i - \bar{z})^2$$ **Equation 2**

Where *n* is the sample size, z_i is the value of the *i*th sampling unit and $\bar{z}$ is the sample mean.

Where $\hat{\sigma}^2$ is the estimated population variance or sampling variance and *n* is again the sample size (Webster and Oliver, 2007). A common way of representing reliability in means estimation is by using confidence intervals or limits. A confidence interval gives an upper and a lower limit around the estimated mean in which the real mean can be assumed to lie. The assumption that the mean lies within this interval is made with a certain degree of confidence usually measured in percentages or as a ratio (see Figure 4).

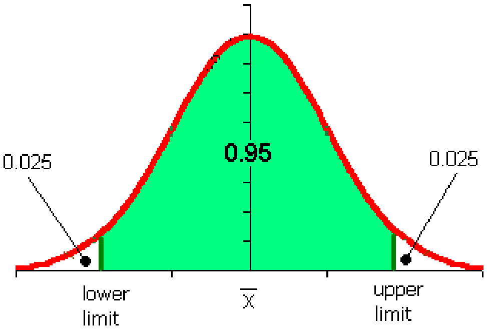


Figure 4: Example of a 95% confidence interval around the mean $\bar{x}$.

The confidence interval can be calculated by:

$$\bar{z} - ys/\sqrt{n} \text{ and } \bar{z} + ys/\sqrt{n}$$ **Equation 3**

Where y in this calculation is an indication of how far the estimated mean, $\bar{z}$, lies from the real mean μ measured in standard deviations and s is the standard deviation of the sample. The value of y can be calculated by:

$$y = \frac{z - \mu}{\sigma}$$ **Equation 4**

Of course the problem is that μ is unknown and so y cannot be calculated on basis of Equation 4. A way of deriving y is through considering a standard normal deviate, where y is a variable which is normally distributed with a mean of 0 and a variance of 1. Lists containing these values and their cumulative probabilities have been published which enable the calculation of confidence limits.

There is however one constraint; with small sample sizes, when $n < 60$, s^2 is not a good estimate of σ^2. So with small sample sizes y is replaced by the student's t which is given by (Webster and Oliver, 2007):

$$t = \frac{\bar{z} - \mu}{s / \sqrt{n}}$$ **Equation 5**

Stratified Random Sampling

Inference from this type of sampling is relatively straightforward and very similar to Simple Random Sampling. The equations used to estimate the global mean and its variance are:

$$\boldsymbol{\hat{\mu} = \sum_{h=1}^{H} \alpha_h \bar{z}_h \quad , \quad \hat{\sigma}^2(\bar{z}) = \sum_{h=1}^{H} \alpha_h^2 \, \hat{\sigma}^2(\bar{z}_h)}$$

Equation 6

Where α_h is the relative area of stratum h, $\bar{z}_h$ is the estimated mean of stratum h and $\hat{\sigma}^2(\bar{z}_h)$ is the estimated variance of the stratum mean $\bar{z}_h$:

$$\sigma^2(\bar{z}_h) = \frac{1}{n_h(n_{h-1})} \sum_{i=1}^{n_h} (z_{hi} - \bar{z}_h)^2$$ **Equation 7**

Where n_h the sample size within stratum h is, z_{hi} is the ith observed value of z in stratum h and $\bar{z}_h$ is the mean calculated from the sample in stratum h.

Ideally one would want to know beforehand how large the sample size should be for a certain level of precision. However the same problem with respect to the required sample size applies to stratified Random Sampling as with Simple Random Sampling. Prior information is needed in order calculate this.

2.3.2 Point predictions for mapping purposes

Most methods for obtaining the mean of an area are estimations. They assume the real mean to be within an interval with a specified degree of confidence. A map containing a grid of values derived from a sample containing fewer points however is a prediction. It predicts the values at unsampled locations by applying some type of spatial prediction model.

The prediction methods mentioned in this section are all model based. Meaning that it is not necessary to implement a p-unbiased (random) sampling method for uncertainty estimation as this can be derived from the model.

Because random sampling is not required, sampling can be done purposively as to cover the area as good as possible; see section 2.2.2.

As was already mentioned in the introduction many methods for spatial mapping in environmental sciences exist which can turn point measurements of a variable into smooth prediction maps for a certain area. Li and Heap (2008) give a fairly extensive overview of methods and also a description of how suitable they are for different applications.

In the spatial interpolation comparison exercise of 2004 (EUR, 2005), three different types of methods were used for automatic interpolation; Neural Networks, Splines and Geostatistical functions. From the results it can be seen that sub-types of all three methodologies have very good as well as far less results.

Because of this variation in results from different methods, it was chosen only to include different flavours of Kriging (Geostatistics) in the framework. Because it is thought that in combination with different forms of inference and sampling they already give the users of the Learning Map (1st year MGI students) sufficient options from an educational point of view. Adding to this is the fact that the users are already familiar with concept of Kriging and its implementation in the statistical computing environment R (R Development Core Team, 2010) through the course Spatial Modelling and Statistics (GRS-30306).

The fact that the basic theory behind these methods is already known to most participating students makes its implementation for the Learning Map more efficient as more time can be spent on sampling and interpreting the results rather than on theory.

Ordinary Kriging

Kriging is a generic term for a group of prediction methods which are all based on the principle that the value at an unobserved location can be predicted by considering it to be a (linear) combination of values nearby (Goovaerts, 1997).

Many types of Kriging exist, taking into account various external factors. All Kriging methods however are based on the same principles. These principles will be explained by the description of a basic form of Kriging; Ordinary Kriging (OK).

Ordinary Kriging, described by Webster and Oliver (2007) as the most robust Kriging metho. It is also the most commonly applied form of Kriging. In their comparison of spatial interpolation methods where 51 studies were reviewed, Li and Heap (2008) observed that Ordinary Kriging was used no less than 37 times.

The prediction model for Ordinary Kriging can be described by (Hengl, 2009):

$$\tilde{z}(s_0) = \mu(s_0) + \varepsilon'(s_0) \qquad \textbf{Equation 8}$$

Where $\tilde{z}(s_0)$ is the predicted value at unsampled location s_0 , μ is a global mean or trend component and $\varepsilon'(s_0)$ is the spatially correlated variation. Unlike Simple Kriging (SK) where the mean is assumed to be known and constant over the whole area, Ordinary Kriging can account for fluctuations of the mean by only considering a local neighbourhood centred around a location s_0 that is to be predicted. Within this local neighbourhood the mean (μ Equation 8) is supposed to be constant but unknown. The predictor for OK at unsampled location s_0 then becomes (Goovaerts, 1997):

$$\tilde{z}(s_0) = \sum_{i=1}^{n(s)} \lambda_i z(s_i) \text{ with: } \sum_{i=1}^{n(s)} \lambda_i = 1 \qquad \textbf{Equation 9}$$

Where $z(s_i)$ is the measured value at location s_i and where λ_i are the associated Ordinary Kriging weights, which determine how much the value at a specific location counts in the prediction for another location.

These Kriging weights can be derived by solving the Ordinary Kriging system:

$$\sum_{j=1}^{n(s)} \lambda_j \cdot \gamma(s_i - s_j) + \varphi = \gamma(s_i - s_0) \qquad \text{with: } \sum_{i=1}^{n(s)} \lambda_i(s_i) = 1 \qquad \textbf{Equation10}$$

Where $\gamma(s_i - s_j)$ is the semivariance at distance between *i* and *j*, φ is the Lagrange parameter which forces the constraint to be obeyed and $\gamma(s_i - s_0)$ is the semivariance between *i* and *0*.

Furthermore the prediction error can be quantified by means of the Ordinary Kriging Variance:

$$\sigma^2(s_0) = \sum_{i=1}^{n} \lambda_i \cdot \gamma(s_i - s_0) + \varphi \qquad \textbf{Equation 11}$$

Where the semivariance is usually smaller when there are many observations in the local neighbourhood or when there is strong spatial correlation.

The semivariance, γ , used in Equations 10 and 11 is a much used concept in Geostatistics, it is half the expected squared difference between two points (Hengl, et al., 2009):

$$\boldsymbol{\gamma(s_i, s_0) = \gamma(h) = \frac{1}{2}\, E[\left(z(s_i) - z(s_i + h)\right)^2]} \qquad \textbf{Equation 12}$$

Where the variance is supposed not to depend on the actual locations of , s_i and s_j but rather on the distance ,*h*, seperating the two using the method of moments (Kerry and Oliver, 2007). Semivariances are estimated from the sampled data by:

$$\boldsymbol{\hat{\gamma}(h) = \frac{1}{2n}\sum_{i=1}^{n}(z(s_i) - z(s_i + h))^2} \qquad \textbf{Equation 13}$$

Where *n* is the number of point pairs included in the calculation , $z(s_i)$ is the value of the variable at location s_i and $z(s_i + h)$ is the value at a location seperated by distance *h*.

An important step is the plotting of the estimated semivariances $\hat{\gamma}(h)$ from the sampled data (Equation 13) against the distance *h* in the so called experimental (semi)variogram. The points in this graph are not single values of semivariance (singe point pairs), because plotting these would result in a difficult to interpret cloud of points, but they represent semivariances averaged over a range of distances (*h*) called lags (Hengl, 2009).
An example of such an experimental variogram with a model fitted through is given in Figure 5. Three important characteristics are indicated in red in this figure; the nugget, the sill and the Range.
The nugget indicates the short distance spatial variation which includes measurement errors. The Sill is the value of the semivariance at which it stops increasing with greater distance (no spatial correlation between points at that distance) and the Range is the value of *h* at which the sill is reached.

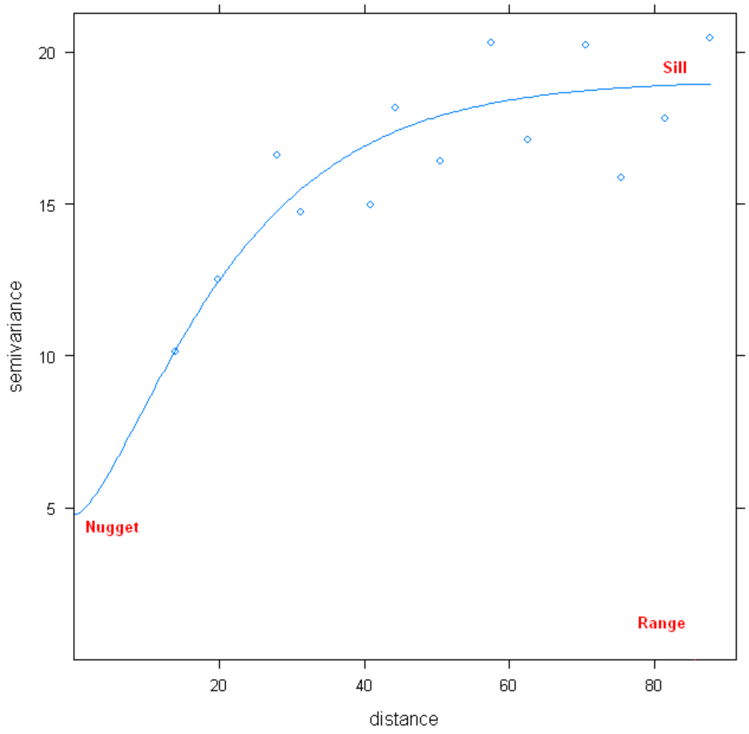


Figure 5: Example of an experimental variogram with a spherical model fitted through the points. Made with the R extension Gstat (Pebesma, 2004: section).

When the experimental variogram is plotted, a model can be fitted through the points This step is very important for the outcome of the Kriging interpolation as the model determines the relation between distance and values used for solving the OK system. Different types of functions can be fitted, but the most commonly used are Spherical, Exponential and Gaussian (Burrough and McDonnel, 1998).

Some discussion exists on the subject of the sample size needed to construct a reliable experimental variogram. Factors affecting the required sample size are the type and amount of spatial variation. For example a much larger sample size might be needed in the case of anisotropy, which means the spatial variation behaves differently in different directions.
However when there is no reason to assume such special circumstances most authors agree to a sample size of at least 50 as a rule of thumb for variogram construction. (Burrough and McDonnel, 1998; Webster and Oliver, 2007; Hengl, 2009)

Kriging with a trend

As stated in section 2.1 the secondary information used in scenario 4 to create a continuous map of the area is a trend (a form of exhaustive data). Numerous Kriging methods exist which take into account secondary information in one way or another. Among those is a group of three methods which are designed especially for use in the case of a trend; Universal Kriging (UK), Kriging with an External Drift (KED) and Regression Kriging (RK), all of which are applied fairly often (Li and Heap, 2008). According to Hengl (2009) because the techniques are fairly similar, some confusion seems to exist within the Geostatistical literature about which term corresponds to which technique.

Although some authors refer to UK as a method where only coordinates are used as secondary variable, it is essentially the same method as KED . Both methods estimate the trend coefficients and the Kriging weights in one (Kriging) process.

Regression Kriging on the other hand refers to the case where the trend and the Kriging weight estimation are separated. Kriging predictions are made from residual values (sample-trend function) and then adding the trend function to those predictions.

Universal Kriging

The basic idea behind UK is supposing that the mean (μ in Equation 8) in the local neighbourhood around location s as described for Ordinary Kriging is not constant within the local area but varies smoothly within the local neighbourhood and thus also in the entire study area.

This trend in the mean $\mu(s_0)$ is modelled as a linear combination of functions $f_k(s_0)$ (Goovaerts, 1997):

$$\boldsymbol{\mu}(\boldsymbol{s_0}) = \sum_{\boldsymbol{k=0}}^{\boldsymbol{K}} \boldsymbol{\alpha_k}(\boldsymbol{s_0})\boldsymbol{f_k}(\boldsymbol{s_0})$$

Equation 14

Where the coefficients $\alpha_k(s_0)$ are unknown coefficients and assumed constant within each local neighbourhood. The factor K in this equation indicates the number of components making up the trend.

These K+1 unknown coefficients are filtered from the linear estimator by imposing a constraint on the weight similar as with OK (Equation 9), allowing the UK prediction to be written as (Goovaerts, 1997):

$$\tilde{z}(s_0) = \sum_{i=1}^{n(s)} \lambda_i^{UK} z(s_i) \quad \text{with} \quad \sum_{i=1}^{n(s)} \lambda_i^{UK} z(s) f_k(s_i) = f_k(s) \;\; k = 0,...,K$$

Equation 15

Where, $z(s_i)$ is the measured value at location s_i and λ_i^{UK} are the Universal Kriging weights.The Kriging weights can be obtained by solving the UK system:

$$\sum_{j=1}^{n(s)} \lambda_j \cdot \gamma(s_i - s_j) + \varphi + \sum_{k=0}^{K} \varphi_k f_k(s_i) = \gamma(s_i - s_0)$$

Where:

$$\sum_{i=1}^{n(s)} \lambda_i^{UK}(s_i) f_k(s_i) = f_k(s_0) \quad k = 0,....,K$$

Equation 16

K
’

Which is essentially the same system as with OK. The semivariances at all distances can be derived by constructing an experimental variogram from the residual values (sampled – trend value) and fitting a function through it. So the only difference with OK is the addition of the trend factor $f_k(s_i)$.

A possibility of how to describe the trend is by using a linear function of the spatial coordinates. Resulting in three functions (Webster and Oliver, 2007):

$$f_0 = 1, \qquad f_1 = x, \qquad f_2 = y$$

Equation 17

In which $f_0$is the spatial variation part like with OK, $f_1$is the part of the trend working in the x direction and $f_2$is the part of the trend working in the y direction.

Regression Kriging

In contrast to Universal Kriging, RK seperates the two processes involved in the prediction; trend estimation and deriving the Kriging weights.

The process of RK is best described by the following 5 steps (Hengl, 2009):

1.) Determine a trend model from the secondary data, commonly done by using Ordinary Least Squares (OLS).
2.) Derive the residuals by calculating the difference between the sampled data and the trend model at all sample locations.
3.) Construct an experimental variogram from these residuals and estimate a function to model the spatial variation.
4.) Perform interpolation, using Ordinary Kriging, based on this residual variogram model.
5.) Add the trend model to the interpolated values.

Block Kriging

Although kriging is mostly used for point predictions it can also give predictions for larger areas by changing the support, this is called Block Kriging (BK). Block Kriging uses the sampling units to interpolate values for the rest of the area (the block, denoted as *D*).

If the averaging process is linear the formula used to calculate the predicted block mean becomes:

$$\tilde{\mu} = \frac{1}{N} \sum_{i=1}^{N} z(s) \qquad \textbf{Equation 18}$$

Where $z(s)$ are the prediction values at all locations s. Another approach is to estimate the mean directly from the sampled data through:

$$\tilde{\mu} = \sum_{i=1}^{N} \lambda_i z(s_i) \qquad \textbf{Equation 19}$$

Where λ_i is the are the block kriging weights and $z(s_i)$ is the value at the sampled location s_i. λ_i is determined such that $\tilde{\mu}$ is unbiased and the prediction variance is minimal, which can be done by solving the Ordinary Block Kriging system:

$$\sum_{j=1}^{n(s)} \lambda_j \cdot \gamma(s_i - s_j) + \varphi = \gamma(s_i - D) \quad \text{with} \sum_{i=1}^{n} \lambda_i = 1 \qquad \textbf{Equation 20}$$

Where $\gamma(s_i - s_j)$ is the semivariance at distance between *i* and *j*, φ is the lagrange parameter representing the extra uncertainty because of the unknown mean and $\gamma(s_i - D)$ is the point to block semivariance; the average semivariance between s_i and all locations within block *D*.

The Block Kriging variance is given by:

$$\tilde{\sigma}^2 = \sum_{i=1}^{n} \lambda_i \gamma(s_i - D) + \varphi - \gamma(D - D) \qquad \textbf{Equation 21}$$

Where $\gamma(D - D)$ is the average value of semivariance within area D.

2.3.3 Assesment Criteria

The accuracy of spatial prediction models is commonly assessed by looking at the difference between the predicted value and the observed value at specific locations. (Hengl, 2009)

In their review of spatial interpolation methods, Li and Heap (2008), list a number of frequently applied measures. In this research three assessment measures will be used; The mean error (ME), the Root Mean Squared Error (RMSE) and the Mean Absolute Error (MAE).

The Mean Error is often used as an indicator for the degree of bias, which is described by Hengl (2009) as; "the accuracy of estimating the central population parameters". The Mean Error is calculated as the mean difference between observed and predicted values in an estimation or prediction:

$$ME = \frac{1}{n}\sum_{i=1}^{n}(\tilde{z}_i - z_i)$$

Equation 22

Where n is the number of observations, $\tilde{z}_i$ is the predicted or estimated value of observation location *i* and z_i is the observed value at observation location *i*.

Although the ME gives a good indication of the bias, better results for accuracy measurement are obtained with the RMSE which deals with the problem of positive and negative by squaring the differences, summarizing them and then taking the square root of the result:

$$RMSE = \left[\frac{1}{n}\sum_{i=1}^{n}(\tilde{z}_i - z_i)^2\right]^{1/2}$$

Equation 23

Where the symbols are the same as with equation 22.

A problem that might arise with the use of the RMSE is that large errors may have a relatively greater influence than small errors (Li and Heap, 2008).
This is why also the MAE will be calculated; this method is essentially the same as the ME only the values with which it is calculated are absolute:

$$ME = \frac{1}{n}\sum_{i=1}^{n}|\tilde{z}_i - z_i|$$

Equation 24

In the test phase, when assessing the interpolation/inference methods, a test dataset will be used. All values of this dataset are known beforehand therefore making it possible to calculate the difference between the observed and the prediction values for all locations (section 3.1).

When using “real” datasets only the observed values at the sampling locations are known to the user. This makes it more difficult to calculate the difference in predicted and observed values. Because collecting a separate validation data can be expensive and time consuming validation is often done through cross-validation (Hengl, 2009).

Cross-validation is a technique where the original data-set is divided such that one part can be used for validation and the other for prediction. Hengl (2009) mentions two types of cross-validation:

- K-fold cross-validation: where the sample is split up into k equal parts and each time one part is left out for validation.
- Leave one out cross-validation: all sampling points but one are used in the prediction calculation and the prediction value is compared to the sampled value. This is done for each point, thus enabling the user to single out problematic values or points.

2.4 *Matching scenarios and statistical methods*

Sections 2.2 and 2.3 have concentrated on the theoretical background of sampling and inference/interpolation from the sampling results. The combination of these two resulting in several sampling strategies. This section will match these strategies with the scenarios from section 2.1. Besides a description motivating the match scenario-strategies given in the next sections, also another flow chart was constructed; Figure 6. This figure is in fact a combination of Figure 1 and Figure 3. It gives a schematic overview of all the major factors involved in choosing an interpolation or inference method for the Learning Map. It is the decision framework mentioned in the introduction.

2.4.1 **Scenario 1: Spatial mean of study area**

In section 2.1 it was stated that there are effectively two reasons as to why the spatial mean is to be estimated: (1) because no more information is needed. The spatial mean of the area gives enough information for the purpose of the survey. When taking into account the sampling method decision tree of Figure 3, which makes a distinction between stratifying and obtaining good spatial coverage. Creates three strategies for obtaining the spatial mean with this goal in mind:

- Compact Geographical Stratification
- Stratified Random Sampling
- Centred Grid Sampling + Block Kriging

The other reason why only to estimate the spatial mean is that the resources and auxiliary knowledge (about strata or a spatial variation model) are so limited that it is not expected that anything more than the spatial mean can be estimated with sufficient accuracy. This necessarily leads to:

- Simple Random Sampling

2.4.2 **Scenario 2: Chloropleth map**

As mentioned in the sampling section different methods may be used for stratification, but there are also different reasons as to why stratification should be applied. In the previous scenario the primary reason for stratification was efficiency; a better estimate or less costly assessment of the whole area is expected when dividing the area into more homogenous parts.

The goal in scenario 2 is slightly different; the aim is to create a chloropleth map of the area. Once strata have been defined Simple Random Sampling and its inference can be used within in these strata to obtain an estimate of their means, variances and confidence intervals.

Although this strategy creates a map with less detail than those of scenario 3 and 4 a benefit is that model free (p-unbiased) uncertainty estimation is possible.

2.4.1 Scenario 3: Continuous map

Recalling section 2.1 and looking at Figure 6 shows that in this scenario the aim is to create a map containing point predictions of the area. Furthermore there is no secondary information which the user can use to improve his prediction model.

In the case of insufficient knowledge to implement a more sophisticated prediction model Ordinary Kriging offers a good solution. Section 2.3.2 shows that it only uses the values at sampled locations in relation to the distances from each other to construct the model by fitting a function through the variogram.

2.4.2 Scenario 4: Continuous map using spatially exhaustive secondary information

The secondary information used in this scenario is a trend (exhaustive data) . Several types of Kriging can be used to incorperate knowledge about a trend in the predictions, some of which are quite similar (Hengl, 2009). Two of these are described in section 2.3.2; Universal Kriging and Regression Kriging.

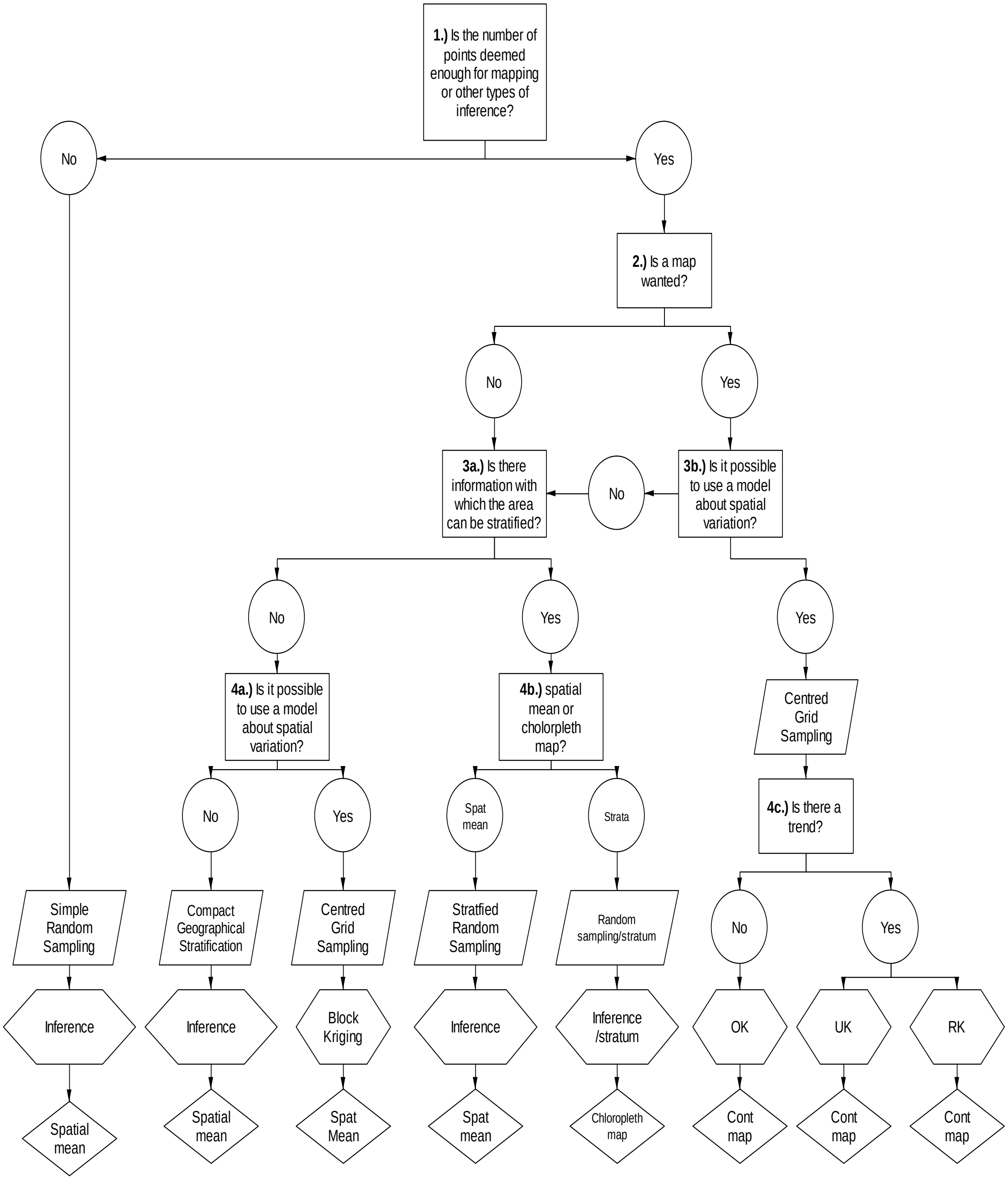

1.) Is the number of points deemed enough for mapping or other types of inference?
No
Yes
2.) Is a map wanted?
No
Yes
3a.) Is there information with which the area can be stratified?
No
3b.) Is it possible to use a model about spatial variation?
No
Yes
Yes
4a.) Is it possible to use a model about spatial variation?
4b.) spatial mean or cholorpleth map?
Centred Grid Sampling
4c.) Is there a trend?
No
Yes
Spat mean
Strata
No
Yes
Simple Random Sampling
Compact Geographical Stratification
Centred Grid Sampling
Stratfied Random Sampling
Random sampling/stratum
Inference
Inference
Block Kriging
Inference
Inference /stratum
OK
UK
RK
Spatial mean
Spatial mean
Spat Mean
Spat mean
Chloropleth map
Cont map
Cont map
Cont map

Figure 6: Descion tree for sampling strategy; choices are based on goals, restrictions and secondary information

2.5 *Application tools*

As was already briefly mentioned in the introduction, the environment which was chosen for the computations and graphical representation is R. Besides the fact that R is freely available through the internet, another major benefit is that although it is not especially designed for spatial applications, it is easily extended for specific uses by means of extension packages (R Development Core Team, 2010). These extension packages are also free and can be downloaded through The Comprehensive R Archive Network (CRAN). The following sub-sections will give a brief description of the tools (packages) which will be used to implement the different strategies.

Sp

Sp is an essential package when dealing with spatial data. It allows users to create, deal and display different types of spatial data like lines, polygons and grids (Pebesma and Bivand, 2005). A particularly useful and much used function in this thesis from this package is *spsample* which allows the user to draw random as well as regular samples from an area.

Spcosa

For Compact Geographical Stratification and sampling from those strata the R package *spcosa* for compact geographical stratification was used (Walvoort, et al., 2010).
Two types of outputs are possible from this package: equal area partitioning and unequal area partitioning. Depending on how large a sample size can be afforded two types of sampling can be done within the strata: sampling only at the centroid (only one sample per stratum) or Stratified Simple Random Sampling (at least two samples per stratum) . Of these two the latter is preferred as it makes estimation of the uncertainty possible because it is a special case of Stratified Random Sampling. The package provides special functions for the implementation of these equations.

Gstat

Most of the calculation needed for inference of the mean and its variance are provided in the basic R functionalities. This not the case for the model based strategies. A package that makes Geostatistical analyses (kriging) possible is Gstat (Pebesma, 2004). This R package allows the user to plot the experimental variogram and manually determine the nugget, sill and range as well as choose a function for the model. The package also contains a function that uses this manual estimate of the model and improves the fit by assigning weights depending on the distance *h* and the number of point pairs *n* in a certain lag. All three types of Kriging (OK,UK an RK) can be implemented by using Gstat. For OK and UK the result also provides uncertainty estimates in the form of Kriging variances besides just the prediction values.
For RK however there is no such precooked variance estimator available because only the residual values are used within the Kriging function. The variance can however be estimated by adding the Kriging variance of the residuals to the variance estimated by the trend model (Hengl, 2009).

3. Implementation and assessment

Now that strategies have been identified for the scenarios in the previous section. In this section the theories are implemented in the statistical computing environment R (R Development Core Team, 2010) and are consequently assessed according to the criteria as defined in section 2.3.3.

First, in section 3.1, a synthetic data set is created of which the values are known at every location. This dataset will be called Test Field hereafter. All strategies are demonstrated on this Test Field first because the truth is perfectly known. Which makes it possible to test the performance of strategies compared to each other and explain the differences by considering the characteristics of the Test Field.
Although the Test Field might give a good indication about the steps involved in implementing the strategies and how they work, it is also important to know if and how they work on data that are actually sampled.
This is why in section 3.2 the first two scenarios are implemented on a real case:
This case is one where different soil parameters (pH, soil moisture, Na, Mg etc.) were randomly sampled on an agricultural field also containing strata in the form of management zones defined by the farmers.

3.1 *Test field*

As stated before, the goal of this constructed reality is to demonstrate how the different strategies work and how they can be implemented. Because values at all locations are known, every type of sampling can be implemented. Equally important is the fact that with a known truth map the accuracy criteria (ME, MAE and RMSE) from section 2.3.3 can be calculated very precisely.
Accurate calculation of these criteria allows us to judge the performance of the strategies compared to one another. This way the added value of using secondary information in means estimation and mapping can be assessed and it can also be seen whether special types of sampling perform better than Simple Random Sampling.

3.1.1 Test field construction

To be able to really test the results between different strategies, the secondary information (strata and a trend, see section 2.1) must be incorporated in the Test Field. This is done by adding three processes together to create the test field: (1) A random field, (2) strata with a fixed value, and (3) a trend along the y-axis (Y-coordinate * 0.25). The following sub- sections will briefly describe how these components are made and what their characteristics are. Finally summation of the values for each pixel created by the three processes will give the characteristics of the test field.

Random Field

The first layer, the random field is generated by first making a grid of 100 x100 pixels. These pixels are given values by using a prediction model with characteristics:

$$\tilde{z} = 20 + \varepsilon'$$

Which means that the value at a certain location is defined as having a global mean of 20 and the variation around this mean is given by a spatial correlation function. This is a spherical variogram model with parameters: Sill 20, Range 40 and nugget 2. Executing this prediction model on the grid results in Figure 7.

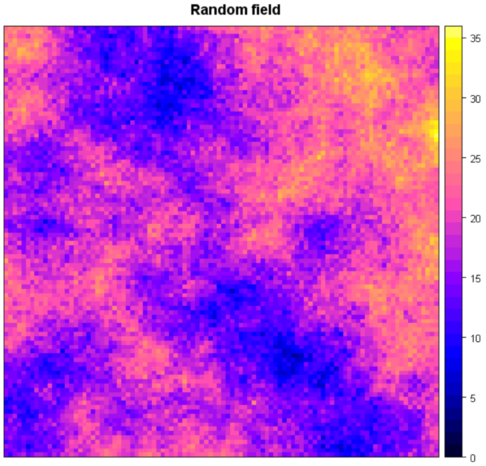


Figure 7: Gaussian Random field of 100x100 pixels with mean 18.01 and variance 20.43.

Strata

The strata are formed by creating a *SpatialPolygonsDataFrame* containing 8 polygons and assigning values to these. Figure 8 shows the layout of these polygons and their corresponding fixed values.

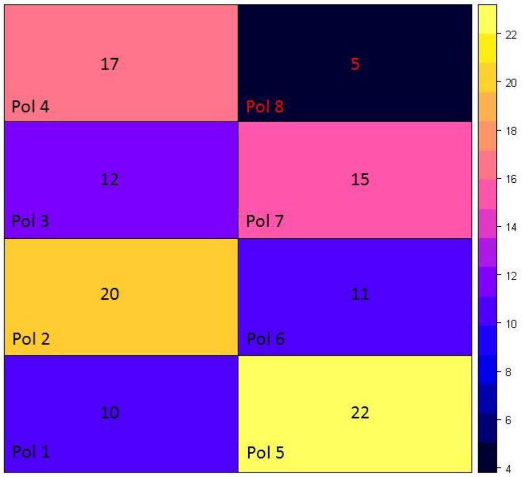


Figure 8: Layout of polygons and their values

Trend

The third layer, a trend in the y direction , contains values ranging from 0.25 to 25. The values of this layer are created by multiplying the value of the y-coordinate with 0.25 for each grid cell, resulting in a stepwise increasing trend in the y-direction (Figure 9).

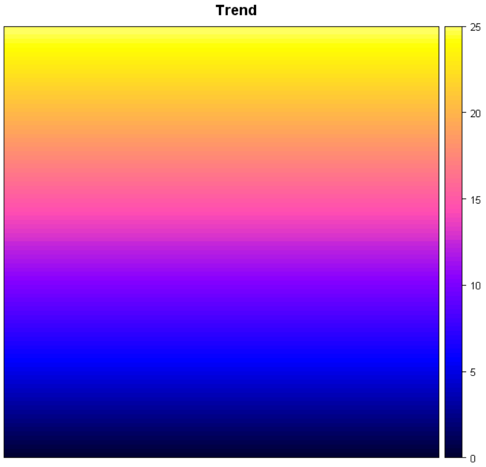


Figure 9: Trend along the y-axis with values ranging from 0-25. Created by assigning a value of 0.25*Y-coordinate to each cell.

Overlay

Simply adding the values from the three layers described above for each grid cell together resulted in the test field of Figure 10.

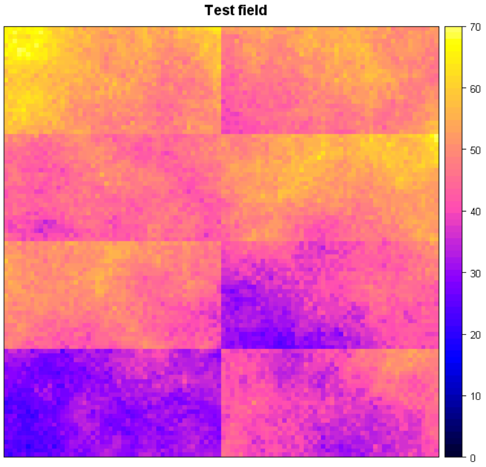


Figure 10: Complete Test Field; summation of the random field, strata and a trend in the y-direction (figures 7-9).
The test field has a global mean of 44.63.

Besides the global mean, which is 44.63, the mean of the strata (including trend, polygons and random field) is also of interest as one of the goals is to estimate them. These are shown in Figure 11 along with some other characteristics of the Test Field. The semivariograms in the bottom pictures of Figure 11 show how much the trend influences the spatial variation at longer distances. At shorter distances the spatially correlated variation still dominates.

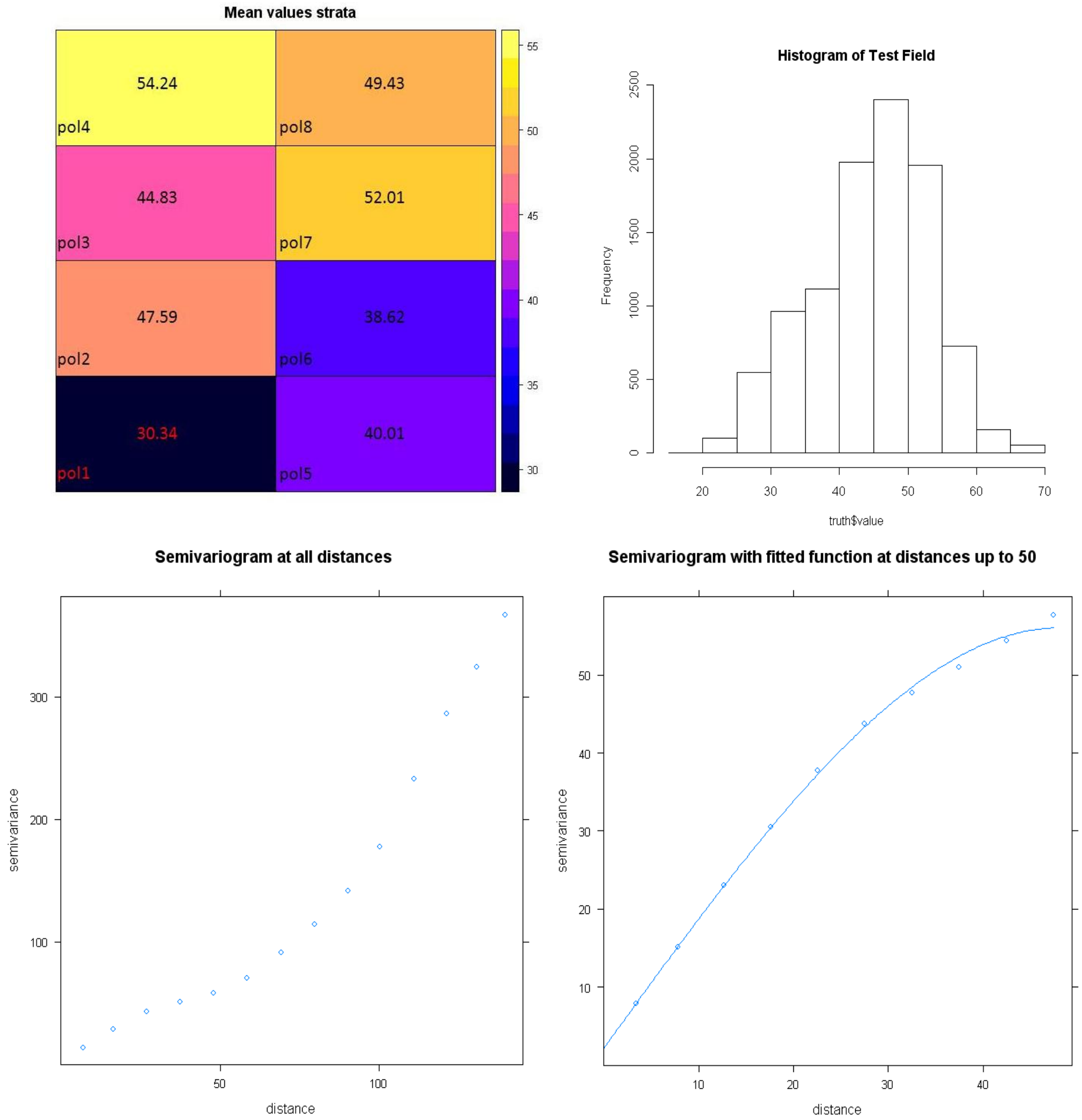


Figure 11: True mean values of strata containing all three layers; random field + polygons +trend, a histogram of values in the Testfield, Semivariogram at all distances and at distances up to 50 with fitted function: nugget 2.23, sill 53.52 and range 48.

3.1.2 Scenario 1: Spatial mean of study area

Simple Random Sampling

Implementing SRS is relatively straightforward; sampling is done with the *spsample* function from the *sp* package (see section 2.5), an example of such a random sampling pattern with a sample size of 10 is given in Figure 12. As can be seen the spatial coverage is not very good.

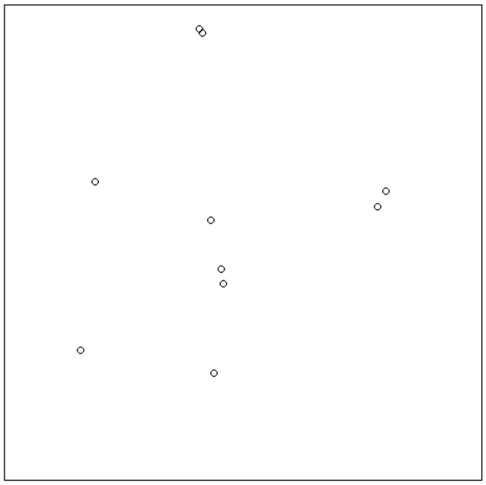


Figure 12: Example of a Simple Random Sampling pattern with a sample size of 10. Sampling points are displayed within the boundaries of the Test Field.

The estimation of the (global) mean, population variance, variance of the mean and confidence intervals are then made from the sampled values using the equations as described in section 2.3.1.

Table 1 shows the results of the aforementioned estimations for Simple Random Sampling with different sample sizes; 2,4,6,8, 10 and 50.

In the methodology section on SRS it was supposed that when more than 10 locations can be sampled, other more efficient sampling strategies should be applied. The SRS result with a sample size of 50 can be used as a comparison to other strategies.

Table 1: Inference results from Simple Random Sampling

Sample size	Mean	Variance	Var of mean	95% conf. int.		Error
				Left	right	
2	46.44	49.88	24.94	31.24	61.63	1.81
4	43.12	61.97	15.49	37.65	48.58	-1.51
6	45.23	62.35	10.39	42.01	48.45	0.6
8	41.58	153.07	19.13	38.02	45.15	-3.05
10	43.12	63	6.3	41.35	44.88	-1.51
50	45	74.74	1.49	44.66	45.34	0.37

Compact Geographical Stratification

As was mentioned in section 2.5, Compact Geographical Stratification can be implemented in R by using the *spcosa* package developed by Walvoort et al. (2008). This package enables the user to easily stratify and sample the study area with the desired sample size and number of strata. Figure 13 shows the result of a stratification into 25 strata with two samples drawn randomly from each stratum, resulting in a total sample size of 50.

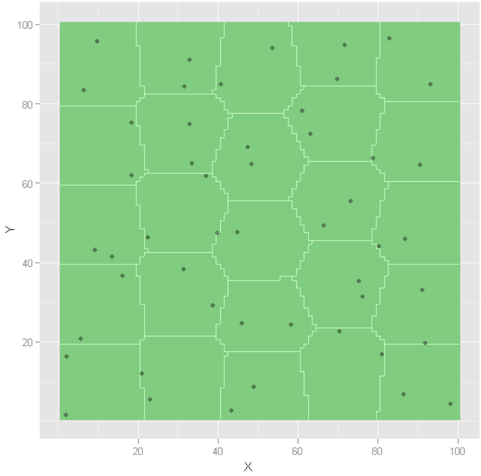


Figure 13: Compact Geographical Stratification into 25 strata using Spcosa (Walvoort et al.,2008). Using a Stratified Random Sampling pattern with 2 samples/stratum.

Table 2 gives the estimates of the (global) mean, variance of the mean and confidence intervals, calculated according to the equations given for inference in the case of Stratified Random Sampling (section 2.3.1). The variance, or spatial variance as it is called in the *spcosa* documentation, is calculated according to equation 7.16 from De Gruijter et al. (2006) and estimates the same parameter as the variance in Table 1 for SRS.

To test the influence of the strata sizes, and so the degree of spatial coverage, on the outcome, three different results are calculated using the same total sample size but with different amounts of strata.

Table 2: Inference results from Compact Geographical Stratification, CGS1-CGS3 represent different strata sizes.

	Nr of strata	Sample size in strata	Mean	Variance	Var of mean	95% conf. int.		Error
						left	right	
CGS1	25	2	44.98	72.81	0.32	44.83	45.14	0.35
CGS2	10	5	45.61	72.71	1.01	45.33	45.88	0.98
CGS3	5	10	44.39	84.02	1.12	44.11	44.67	-0.24

Centred Grid Sampling + Block kriging

Just like random sampling, a regular sampling pattern can be implemented in R by means of the *spsample* function. This is done by setting the pattern at regular instead of random. Figure 14 gives such a regular sampling pattern, this pattern shows 49 sample locations instead of 50 because otherwise a regular pattern with even spacing in x and y direction would not be possible.

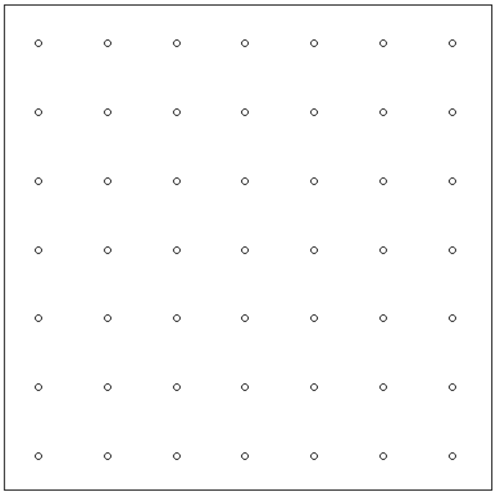


Figure 14: Example of Centred Grid Sampling with a sample size of 49 (7x7).

Before Block Kriging can be applied, a model of spatial variation must be estimated from the sampled data. This was done through the construction of an experimental variogram and fitting of a suitable function.

Displaying this data in an experimental variogram allows us to estimate such a function (Figure 15). A Spherical function was chosen, based on the fact that the model from the Random Field is also based on a Spherical function. The other parameters, the nugget range and sill were estimated visually and fitted with the *fit.variogram* function from Gstat.

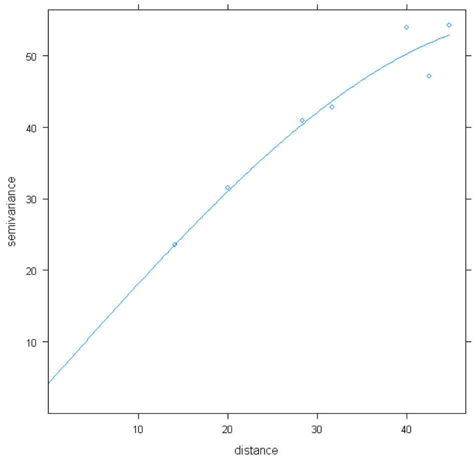


Figure 15: Experimental variogram including a fitted spherical model with: nugget = 4.17, psill = 51.00, range = 54.32.

Using this model in the implementation of Block Kriging (BK) with the R package Gstat, according to the equations given in section 2.3.1 gives the results displayed in Table 3.

Table 3: Prediction results from Block Kriging

Sample size	Mean	Variance[1]	Var of mean (kriging var)	95% conf. int.		Error
				left	Right	
49	44.8	-	0.19	44.68	44.92	0.17

[1] The variance is not given here because the Block Kriging function only gives the (kriging) variance of the calculated mean (Equation 21).

Stratified Random Sampling

The implementation of Stratified Random Sampling is mostly the same as with SRS. The difference being that the sampling must now be done for each stratum separately (Figure 16).

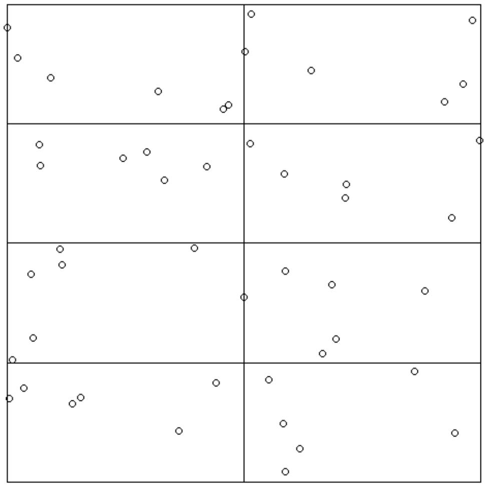


Figure 16: Example of a Stratified Random Sampling pattern with 6 samples/stratum.

This was done by placing the polygons, the strata as defined in 3.1.1, in a loop in which the sampling takes place.

From those sampling results the mean and variance are estimated for each stratum and the results are stored. Loading these results in one data frame makes the estimation of the global mean and variances according to the equations for inference from Stratified Random Sampling in section 2.3.1 possible.

The result for a total sample size of 48 with a sample size of 6 units in each stratum is given in Table 4.

Table 4: Inference results from Stratified Random Sampling

Sample size	Mean	Variance[2]	Var of mean	95% conf. int.		Error
				left	Right	
48	44.57	73.45	0.34	44.41	44.73	-0.06

[2] Calculated according to same equation as Compact Geographical stratification.

Comparison

The past section has focussed on implementing the strategies for scenario 1 and giving the results for each strategy separately. It is however also interesting to see how the strategies compare to one another as they all estimate the same parameter (the spatial mean), Figure 17 and Table 5 gives such a comparison.

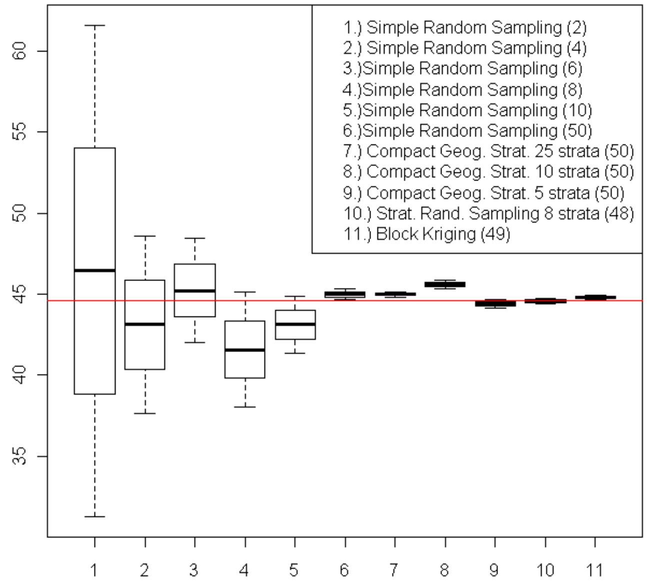


Figure 17: Mean and 95% confidence intervals for strategies of Scenario 1. The red line through the figure represents the true mean of the Test Field and the number in between brackets in the legend indicates the sample size.

Table 5: Performance criteria for methods of scenario 1. Number in between brackets indicates the sample size.

Strategy	ME	RMSE[3]
Simple Random Sampling (2)	1.80	8.77
Simple Random Sampling (4)	-1.52	8.72
Simple Random Sampling (6)	0.59	8.60
Simple Random Sampling (8)	-3.05	9.11
Simple Random Sampling (10)	-1.52	8.72
Simple Random Sampling (50)	0.36	8.59
Compact Geog. Strat 25 strata (50)	0.35	8.59
Compact Geog. Strat 10 strata (50)	0.97	8.64
Compact Geog. Strat 5 strata (50)	-0.24	8.59
Strat. Rand. Sampling 8 strata (48)	-0.07	8.58
Block Kriging (49)	0.17	8.59

[3] Calculated by considering the error to be: Estimated mean – Real value per point.

3.1.3 Scenario 2: Chloropleth map

Stratified Random Sampling

Sampling in this case is done by the same procedure as Stratified Random Sampling in scenario 1. This results in a pattern similar to that of Figure 16 when using the same sample size. The difference between the two strategies is of course the goal of the survey; where in Scenario1 the goal was to estimate the global mean; in this scenario we are interested in the strata mean (Figure 18).

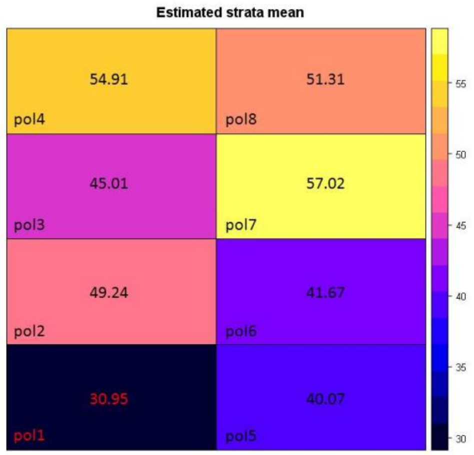


Figure 18: Estimated mean of the strata, obtained through Simple Random Sampling within the strata with a sample size of 6/stratum.

Comparing Figure 18 with Figure 11,which contains the real mean values, tells us that generally the estimates are fairly good. However when looking in more detail, Table 6 shows that there are a few outliers in terms of error; strata 6-8. Calculating the overall ME, RMSE and MAE from the errors of the strata (estimated mean of strata -real mean of strata) yields: ME: 1.64, RMSE: 2.28, MAE: 1.64.

Table 6: Inference results for strata separately using Stratified Random Sampling with a sample size of 6/stratum. The real mean per stratum is given as reference.

	Real mean	Estimated mean	Error	Variance	Var of mean	95% Confidence interval	
						left	right
pol1	30.34	30.95	0.61	4.59	0.77	30.08	31.83
pol2	47.59	49.24	1.65	17.55	2.93	47.53	50.95
pol3	44.83	45.01	0.18	15.04	2.51	43.43	46.59
pol4	54.24	54.91	0.67	25.18	4.20	52.86	56.96
pol5	40.01	40.07	0.06	5.98	1.00	39.08	41.07
pol6	38.62	41.66	3.04	28.17	4.69	39.50	43.83
pol7	52.01	57.02	5.01	16.52	2.75	55.36	58.68
pol8	49.43	51.31	1.88	13.78	2.30	49.79	52.82

3.1.4 Scenario 3: Continuous map

Most of the previous strategies, being design based, were relatively straightforward; a certain sampling method was chosen and a matching set of inference techniques was used to obtain the desired values. This scenario however uses a model based strategy; Ordinary Kriging.

Using a model based method for spatial prediction, without a prior model or secondary information, means making assumptions about the spatial variation and how to capture it best, see section 2.3.2.

Applying this prediction method to the Test Field allows us to observe the effect of certain elements in the design of the survey and the estimation of the model function on the outcome.

Capturing the spatial variation is done by sampling, based on literature (sections 2.2.2 and 2.3.2) some assumptions were made about what type of sampling is preferred for prediction methods and how large the sample size should be. Concluding from these sections Squared Grid Sampling with a simple square grid and a sample size of at least 50 seemed most suitable. By using the above described sampling method as a reference and varying in the following three points an indication about the importance of each factor is obtained:

1. Sampling method (random vs. regular).
2. Grid shape (square vs. triangular)
3. Sample size

Sampling parameters

This section gives prediction and variance maps + the function used in the calculation for the following sampling situations:

- Reference situation; sample size 49 and regular square grid. (Figure 19)
- Random Sampling; with sample size 50. (Figure 20)
- Centred Grid Sampling with triangular grid; sample size 56. (Figure 21)
- Centred Grid Sampling with square grid; sample size 100. (Figure 22)
- Centred Grid Sampling with square grid; sample size 196. (Figure 23)

The sample size are not rounded numbers because of the square shape of the area. To get optimal coverage with a sample size of around 50 results in slightly differing sample sizes.

Table 7 gives an overview of how OK performs under the circumstances mentioned above.

Table 7: Mean, maximum variance and performance criteria of OK point predictions for 5 different sampling situations: row 1 matches fig. 18, row 2 matches fig. 19, row 3 Matches fig.20, row 4 matches fig.21 and row 5 matches fig. 22.

Sample design	Sample size	Mean	Maximum kriging variance	ME	RMSE	MAE
Reg	49	44.80	31.33	0.17	4.11	3.20
Random	50	44.81	56.43	0.17	5.21	3.82
Triangular	56	44.50	38.35	-0.13	3.82	3.02
Reg	100	44.72	26.75	0.08	3.60	2.78
Reg	196	44.60	19.78	-0.04	3.08	2.33

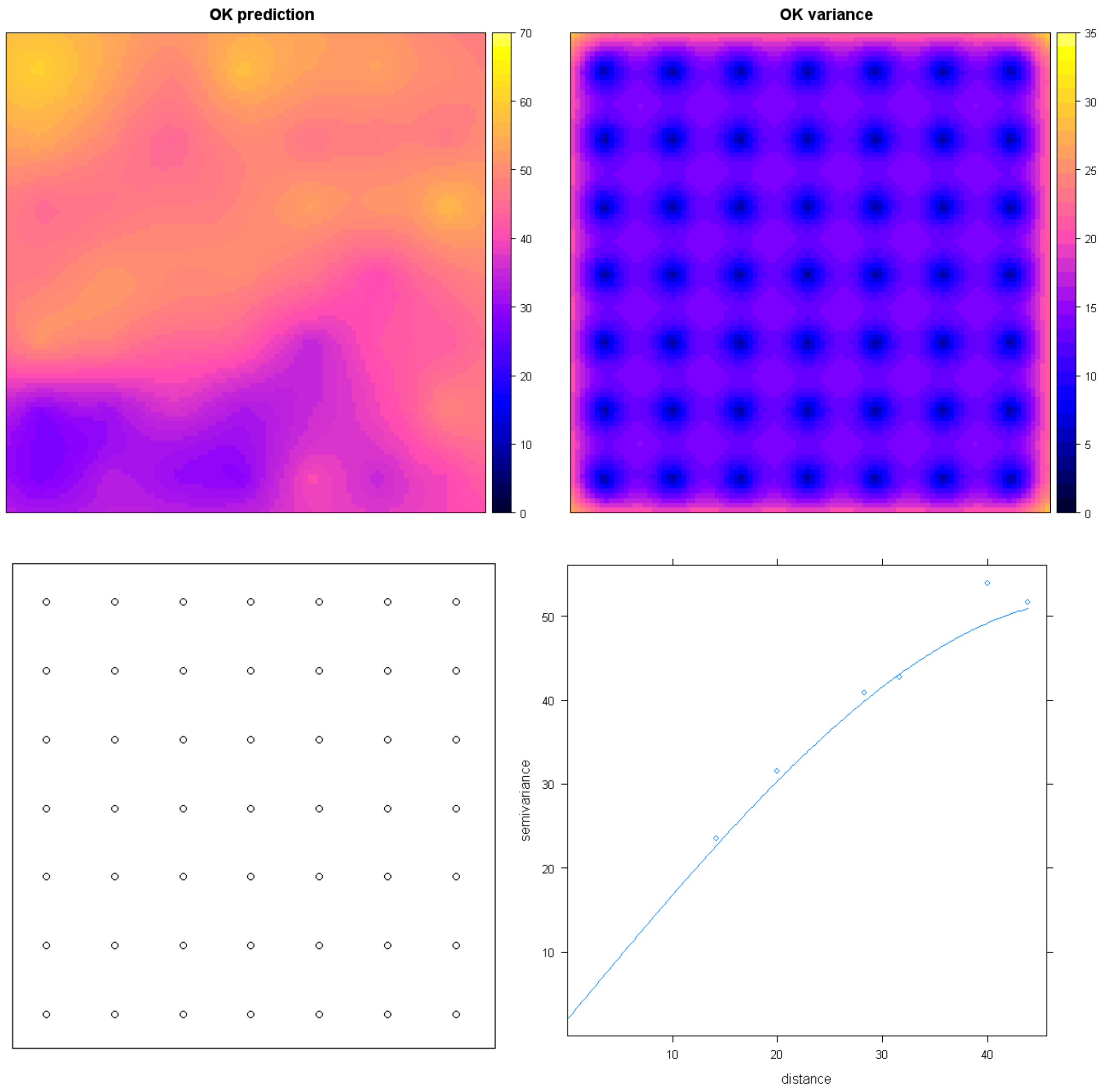


Figure 19: Point prediction map and Kriging variance map, using Centred Grid Sampling with a sample size of 49 and a spherical model with nugget 3.65, psill 50.59 and range 52.53.

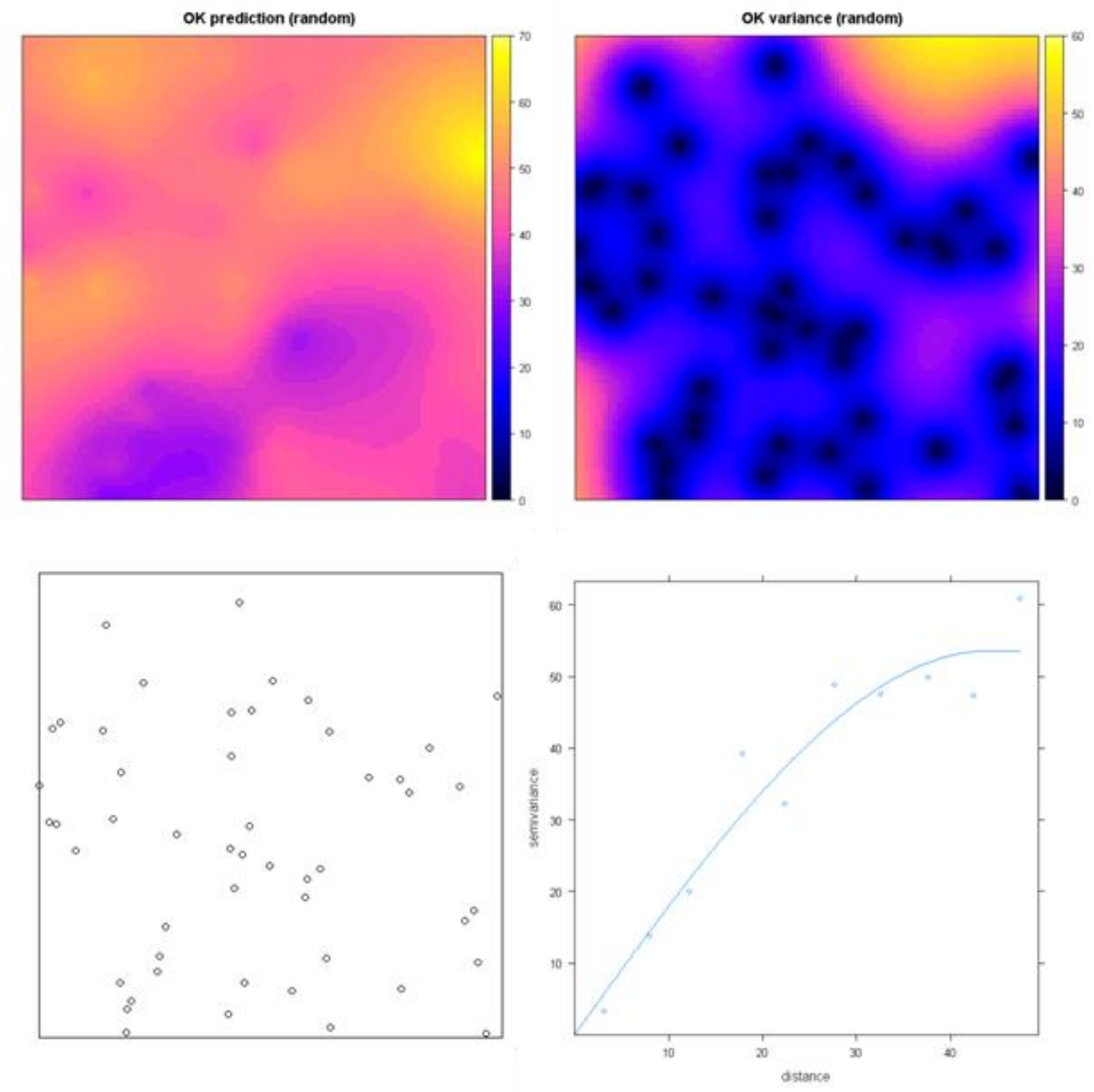


Figure 20: Point prediction map and Kriging variance map, using Simple Random Sampling with a sample size of 50 and a spherical model with nugget 0, sill 53.60 and range 44.10.

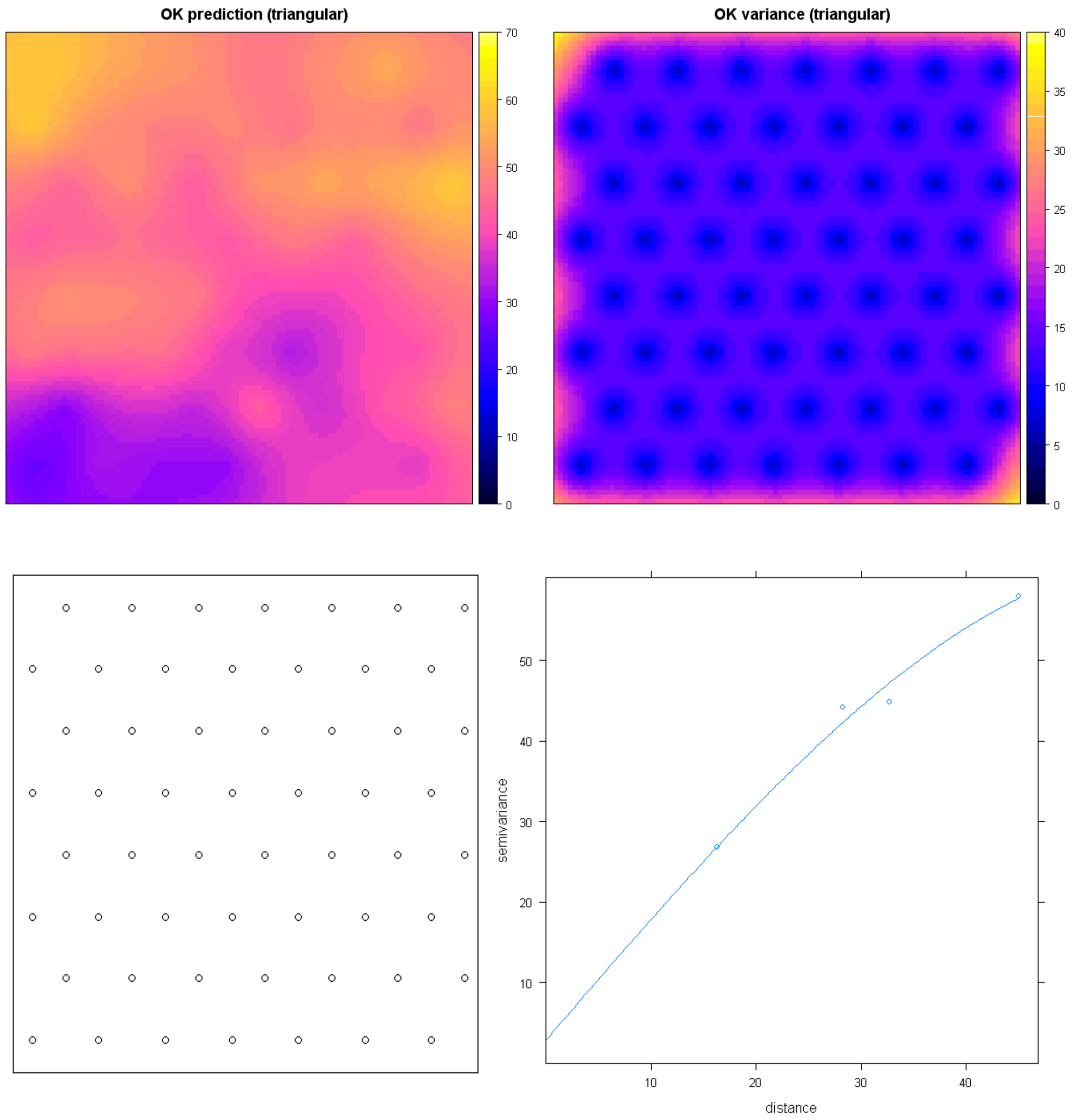


Figure 21: Point prediction map and Kriging variance map, using triangular Centred Grid Sampling with a sample size of 56 and a spherical model with nugget 2.88, psill 59.48 and range 59.02.

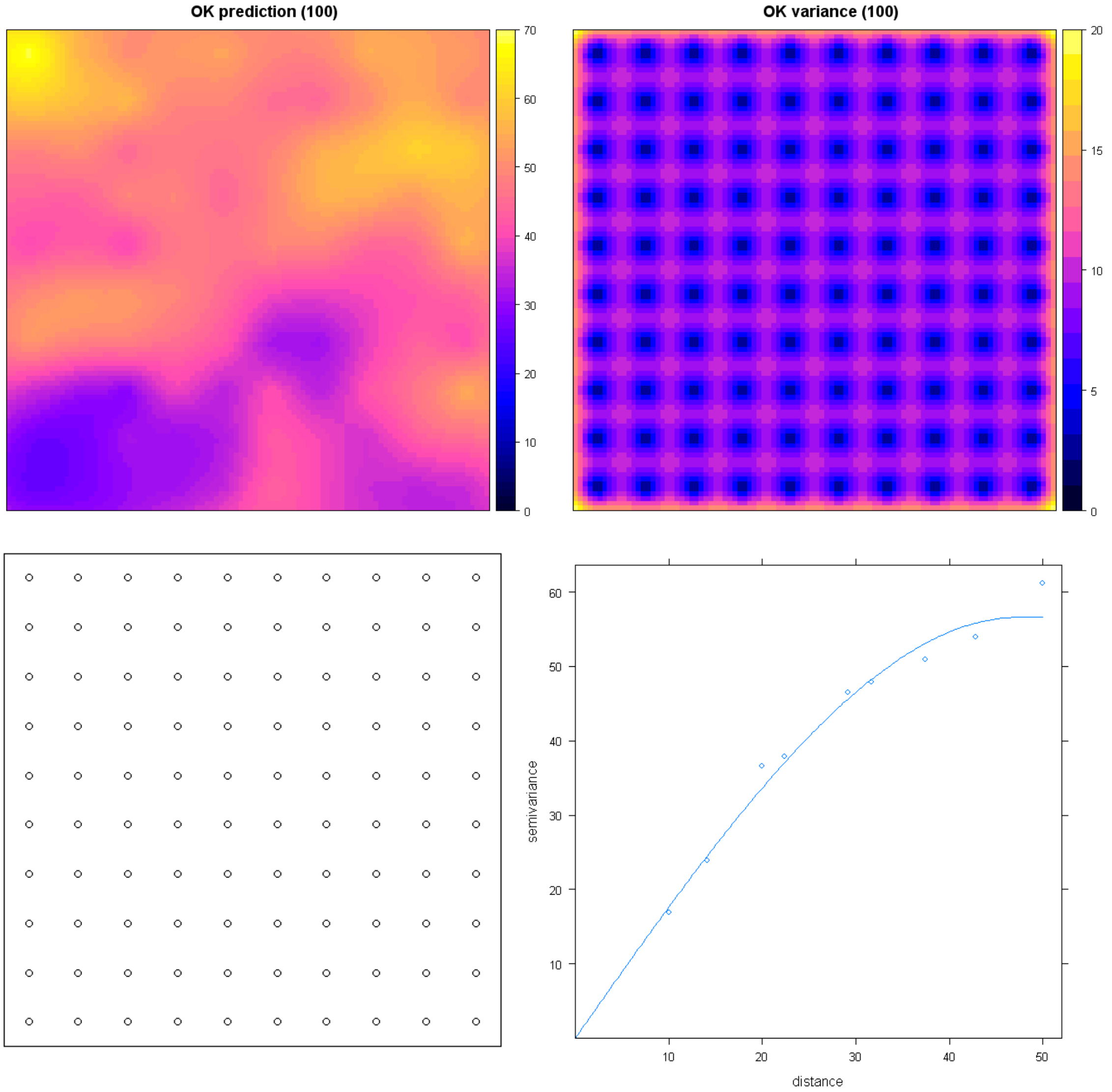


Figure 22: Point prediction map and Kriging variance map, using squared Centred Grid Sampling with a sample size of 100 and a spherical model with nugget 0, psill 56.61 and range 47.44

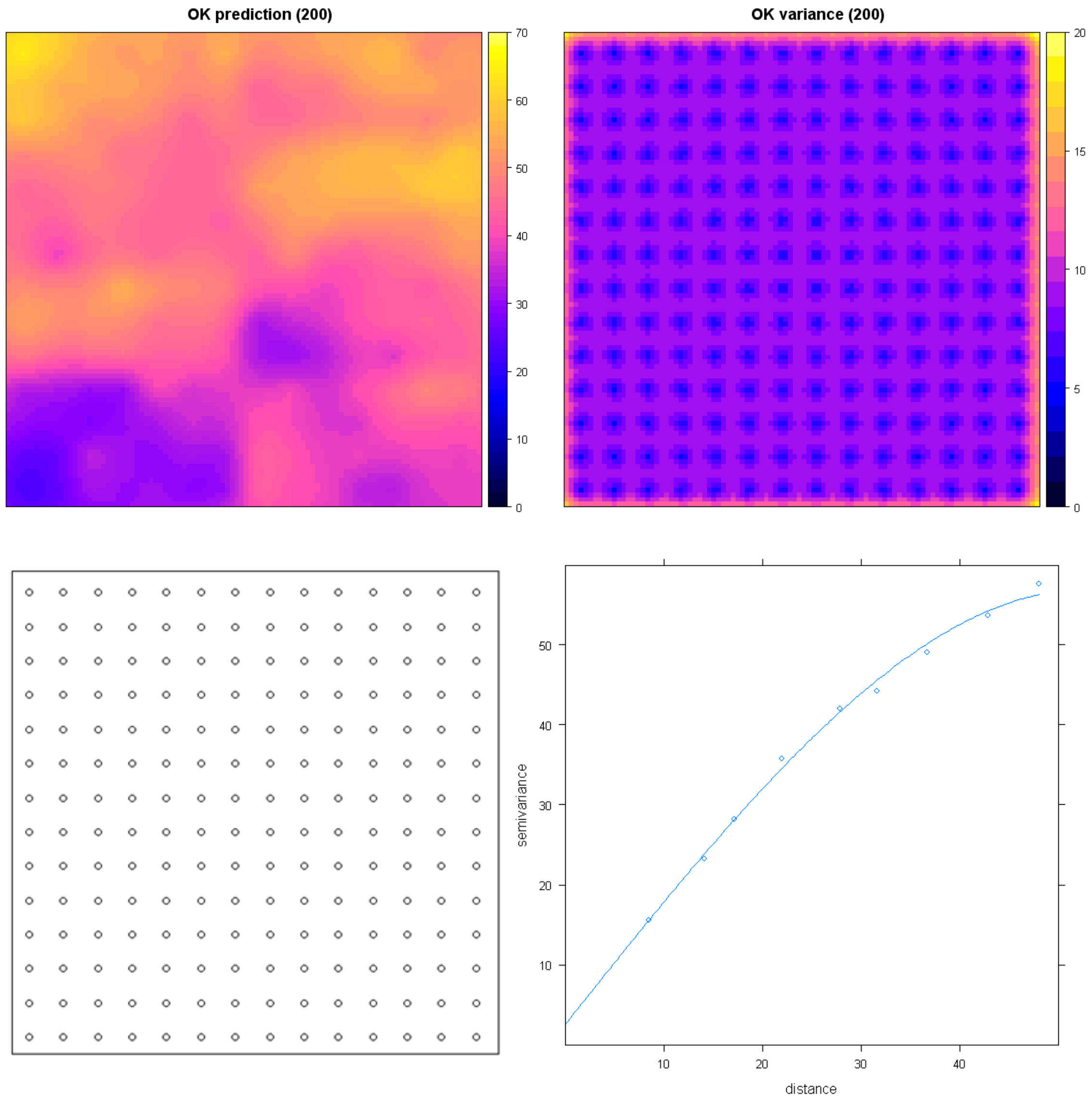


Figure 23: Point prediction map and Kriging variance map, using squared Centred Grid Sampling with a sample size of 196 and a spherical model with nugget 2.61, psill 54.28 and range 52.76.

3.1.5 Scenario 4: Continuous map using spatially exhaustive secondary information

In this section the two prediction methods using secondary information about a trend will be demonstrated; Universal Kriging and Regression Kriging.

Universal Kriging

In this section Universal Kriging is applied first using only the y-coordinate (trend) as secondary variable (Figure 24, Figure 25, Figure 26 and Table 8) and secondly using both the y-coordinate as the strata as secondary variables (Figure 27, Figure 28, Figure 29 and Table 9).

An easy to apply option for doing so is provided in the Gstat extension for R (Pebesma, 2004); the *Krige* function contains an option formula where the user can define the dependant variable as a linear function of an independent variable, which in this case would be the y-coordinate and the strata.

Table 8: Mean, maximum variance and performance criteria for UK (only trend) with 3 different sampling situations: sample size 49 matches Figure 24 , 100 matches Figure 25 and 196. Matches Figure 26.

Sample size	Mean	Maximum variance	ME	RMSE	MAE
49	44.81	25.51	0.17	4.07	3.13
100	44.72	18.30	0.08	3.57	2.67
196	44.61	18.25	-0.02	3.10	2.34

Table 9: Mean, maximum variance and performance criteria for UK (trend and strata) with 3 different sampling situations: sample size 49 matches Figure 27Figure 24 , 100 matches Figure 28 and 196. Matches Figure 26.

Sample size	Mean	Maximum variance	ME	RMSE	MAE
49	45.11	16.53	0.47	3.71	2.95
100	44.60	10.76	-0.04	2.58	2.06
196	44.49	10.40	-0.15	2.45	1.95

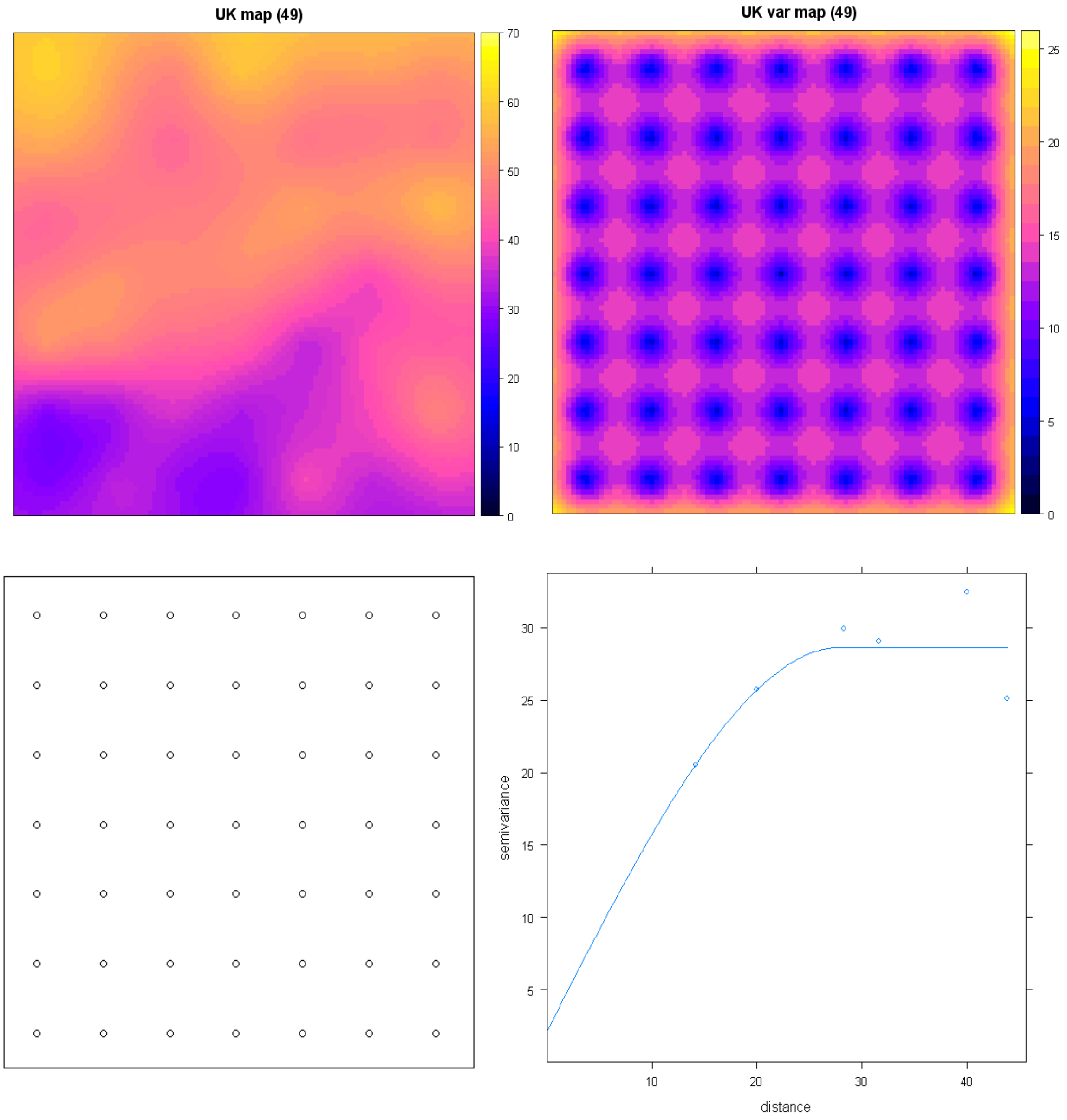


Figure 24: Point prediction and Kriging variance map with UK, using squared Centred Grid Sampling with a sample size of 49 and a spherical model with nugget 2.17, psill 26.49 and range 28.0.

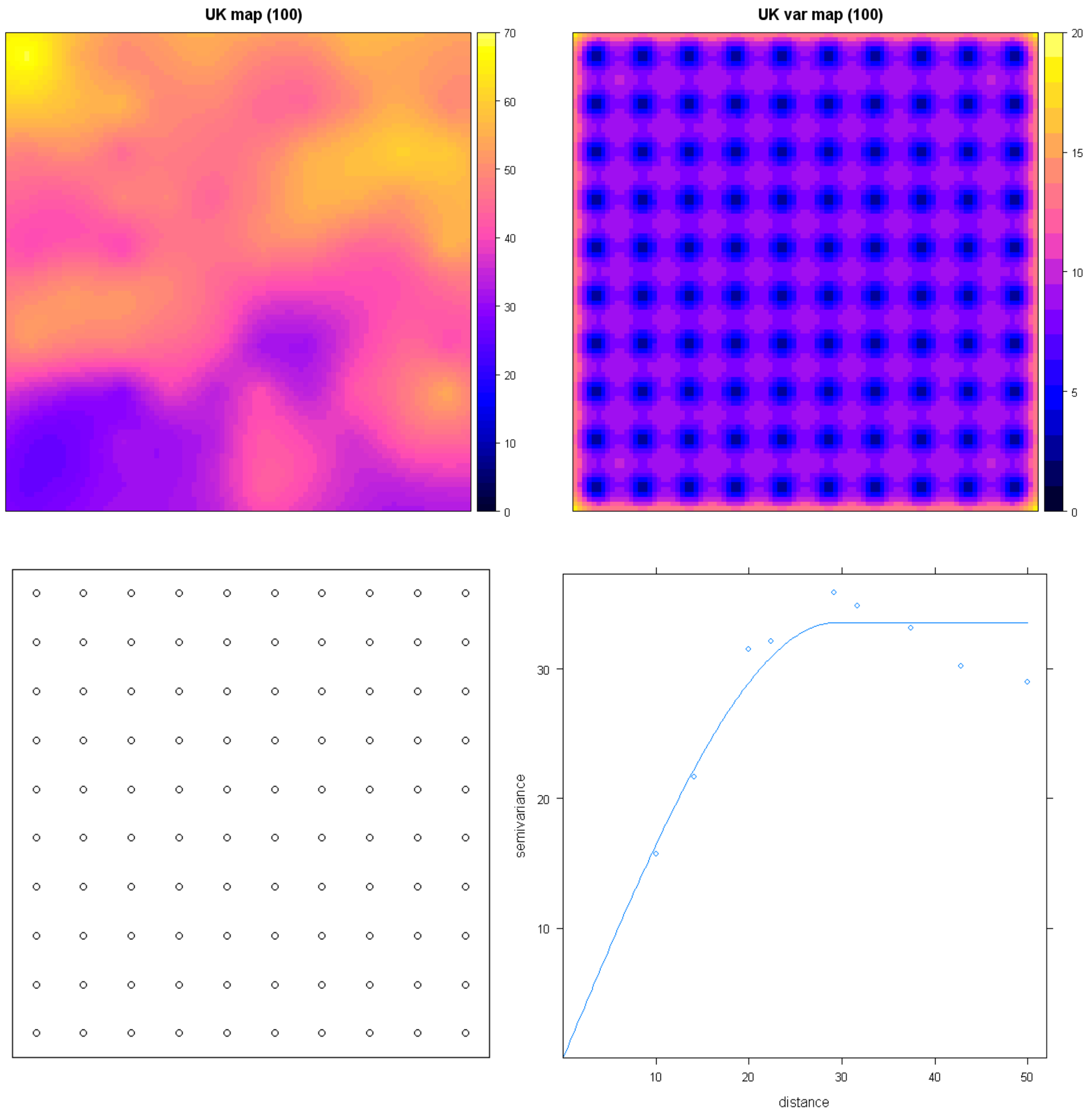


Figure 25 Point prediction map and Kriging variance map with UK, using squared Centred Grid Sampling with a sample size of 100 and a spherical model with nugget 0 , psill 33.60 and range 29.44.

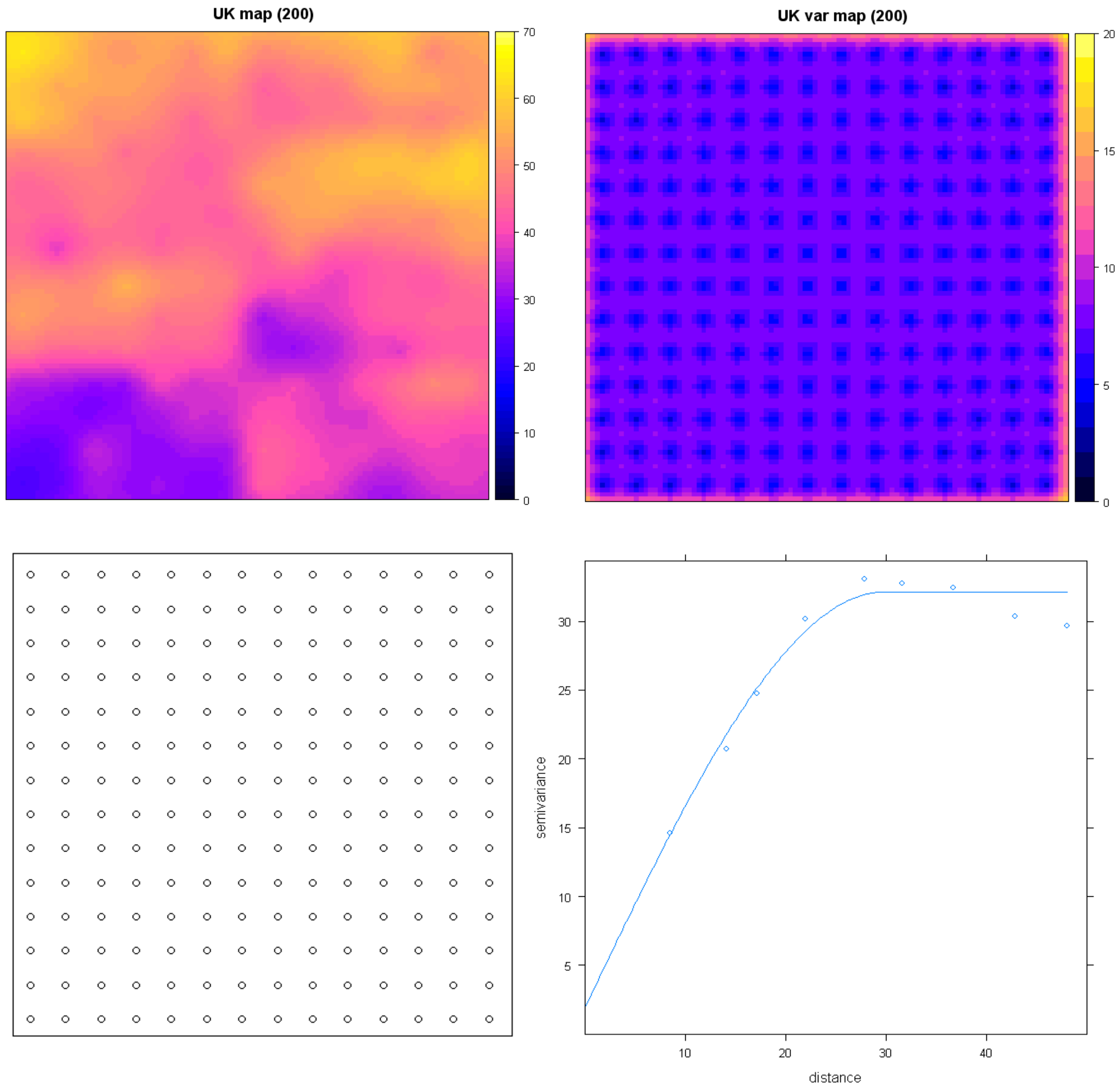


Figure 26: Point prediction map and Kriging variance map with UK, using squared Centred Grid Sampling with a sample size of 196 and a spherical model with nugget 1.95 , psill 30.17 and range 29.75.

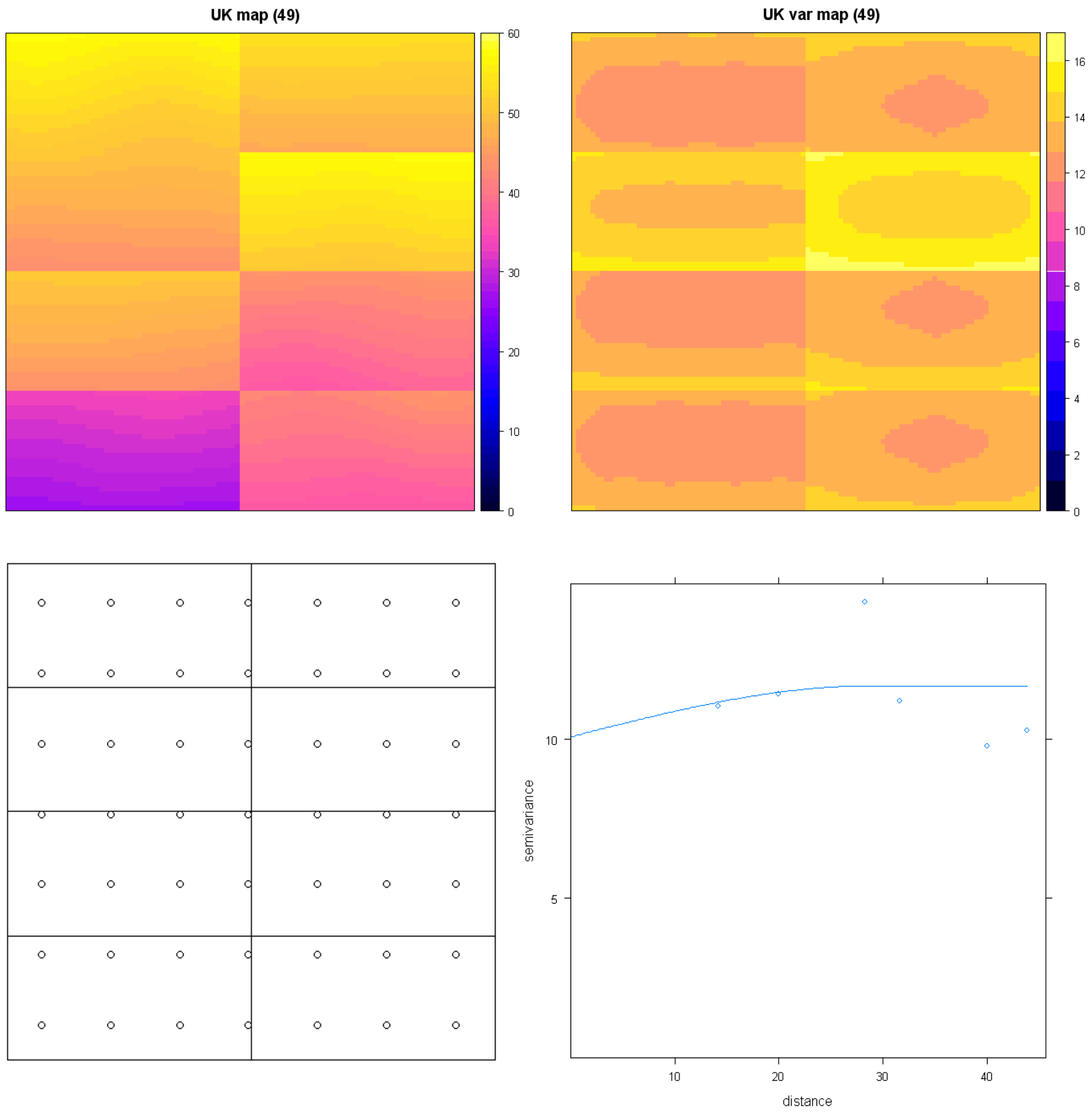


Figure 27: Point prediction map and Kriging variance map with UK (strata +trend), using squared Centred Grid Sampling with a sample size of 49 and a spherical model with nugget 1.6 , psill 10.08 and range 28.76.

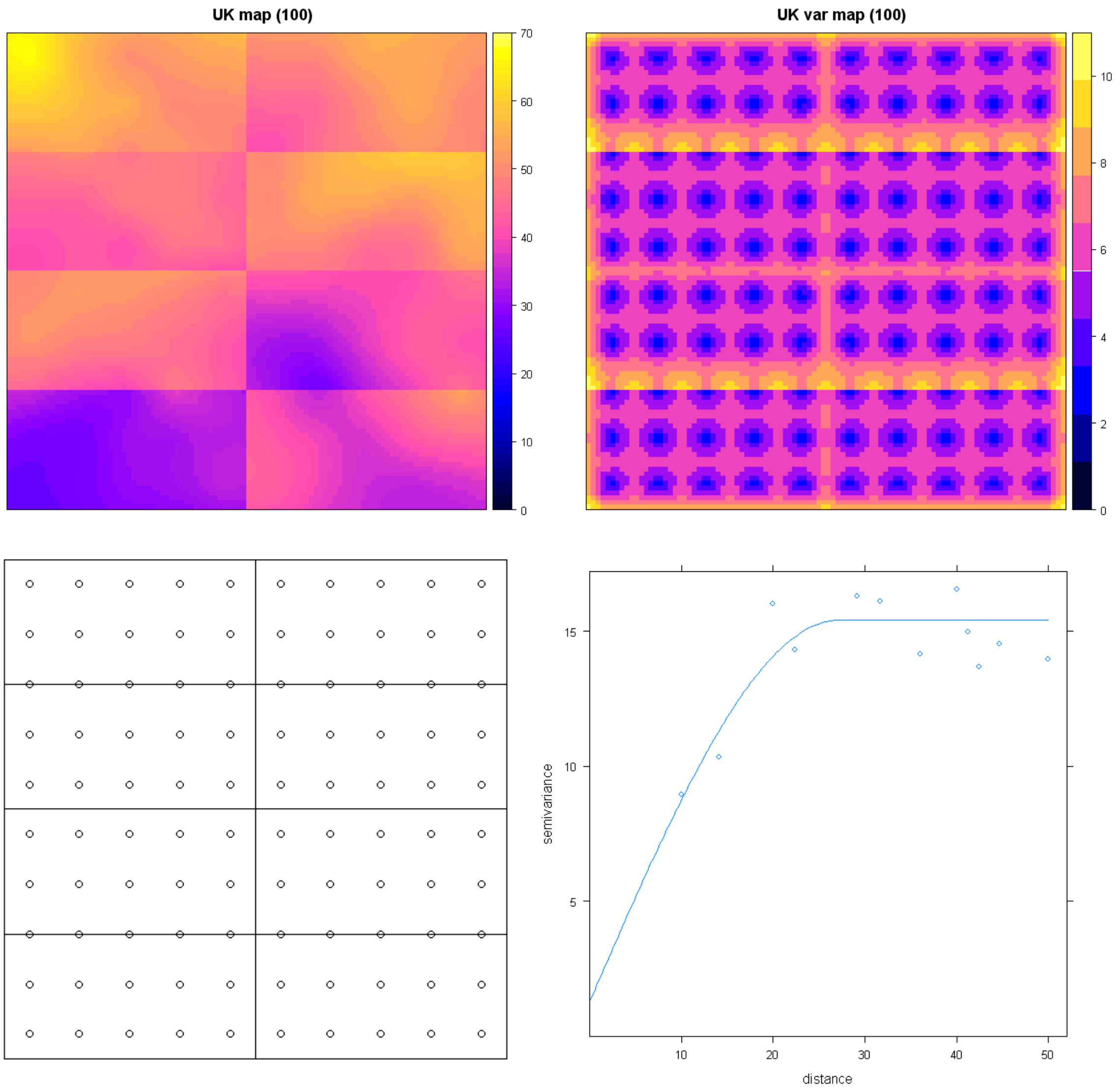


Figure 28: Point prediction map and Kriging variance map with UK (strata +trend), using squared Centred Grid Sampling with a sample size of 100 and a spherical model with nugget 1.28 , psill 14.13 and range 27.10.

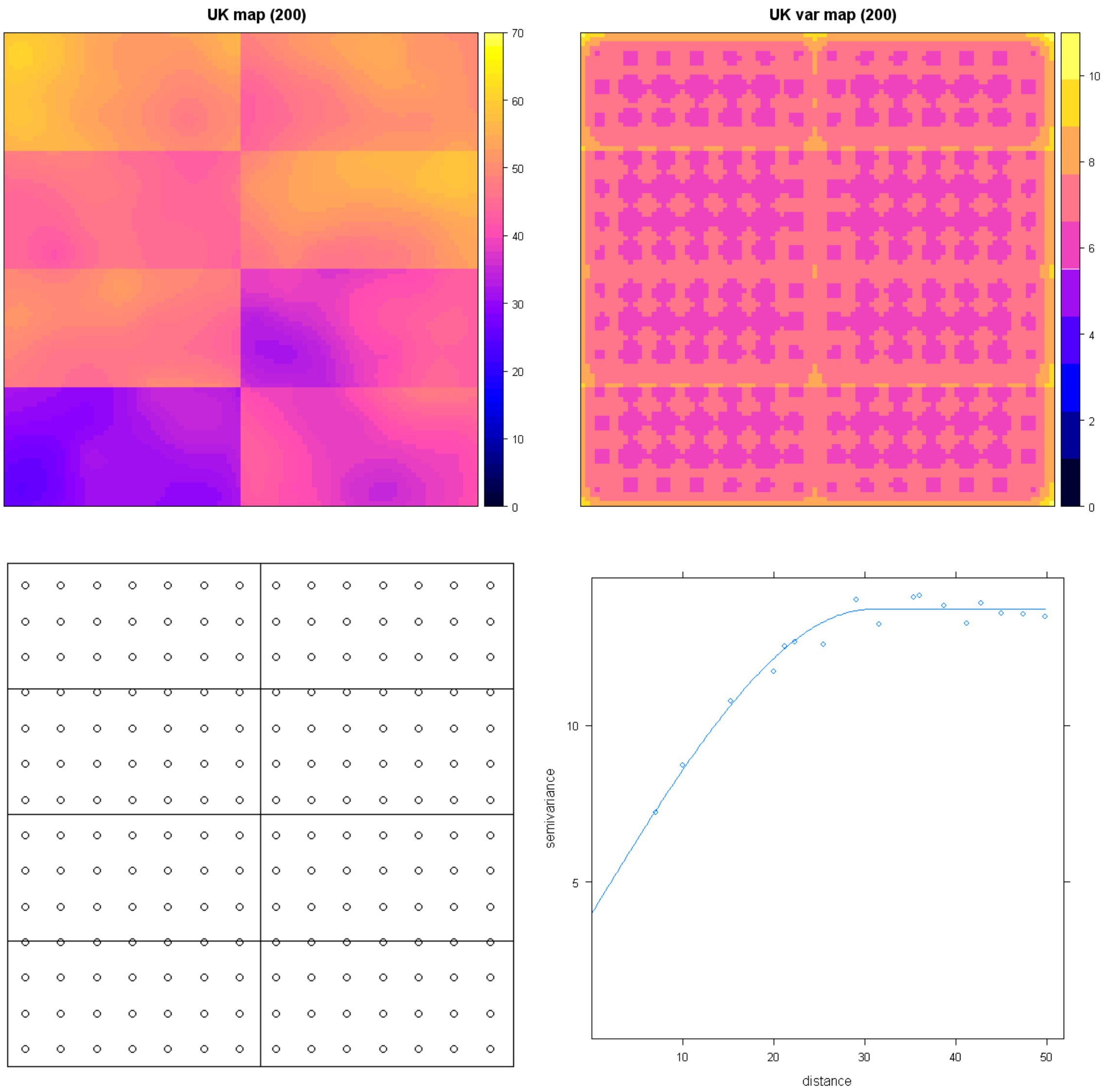


Figure 29: Point prediction map and Kriging variance map with UK (strata +trend), using squared Centred Grid Sampling with a sample size of 196 and a spherical model with nugget 4.0 , psill 9.69 and range 30.72.

Regression Kriging

As described in the methodology section on Regression Kriging the processes of estimating the trend and Kriging the residuals are strictly separated.

This separation comes in handy when applying it to the Test Field as the trend function is known to exactly; 0.25*Y-coordinate.

Although a good prediction would be expected when the exact function is used, comparing Table 10 with Table 7 shows that applying RK to the Test Field did not yield results better than OK.

Table 10: Regression Kriging results from the Test Field for squared Centred Grid Sampling with sample sizes; 49,100 and 196.

Sample size	Mean	Maximum variance (OK variance of residuals)	ME	RMSE	MAE
49	44.81	22.87	0.18	4.10	3.14
100	44.72	17.75	0.08	3.59	2.70
196	44.61	17.33	-0.02	3.11	2.36

A possible cause for this lack of improvement is that the strata which are also still present in the Test Field cause much extra spatial variation and thereby "overshadow" the effect of the trend. This is why another experiment was conducted using a new Test Field without the strata (Figure 30).

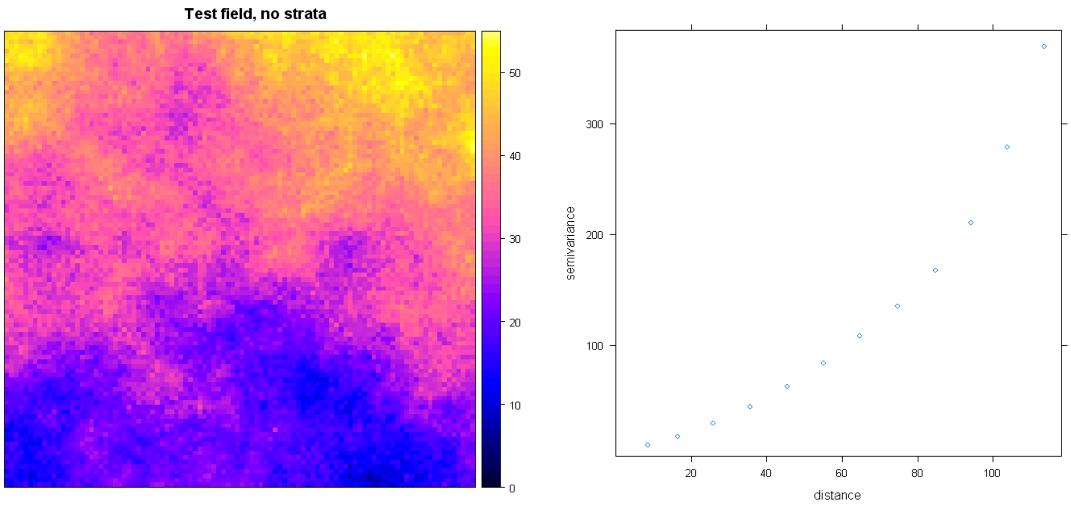


Figure 30 Left; New Test Field without strata. This new field is a summation of figures 6 and 8. It has a global mean of 30.63. Right; semivariogram of the New Test Field.

Performing RK on this field (Figure 31, Figure 32, Figure 33) yields the results as shown in Table 11.

Because the exact nature of the trend will never be known to the surveyor, like in the example above, a *lm* (linear regression) function is used here to construct a linear model using the y coordinates an independent variable.

Table 11: Mean, maximum variance and performance criteria for Regression Kriging (only trend) on the New Test Field without strata using the y-coordinates to construct a linear model. Sample size 49 corresponds to figure 27, 100 to figure 28 and 196 to figure 29.

Sample size	Mean	Maximum Variance	ME	RMSE	MAE
49	30.63	17.35	-0.01	2.62	2.08
100	30.57	11.18	-0.07	2.55	2.03
196	30.53	10.54	-0.10	2.33	1.86

Another option is to include the strata into the regression analyses just as with UK, using the original Test Field to sample from. This results in Table 12. The figures containing the maps and variances are left out here because they are almost identical to those of UK (Figure 27, Figure 28 and Figure 29).

Table 12: Mean, maximum variance and performance criteria for Regression Kriging (trend and strata) on the (orignal) Test Field using the y-coordinates and the strata to construct a linear model.

Sample size	Mean	Maximum Variance	ME	RMSE	MAE
49	45.10	17.65	0.46	3.71	2.95
100	44.60	11.22	-0.03	2.60	2.07
196	44.48	9.89	-0.15	2.51	2.00

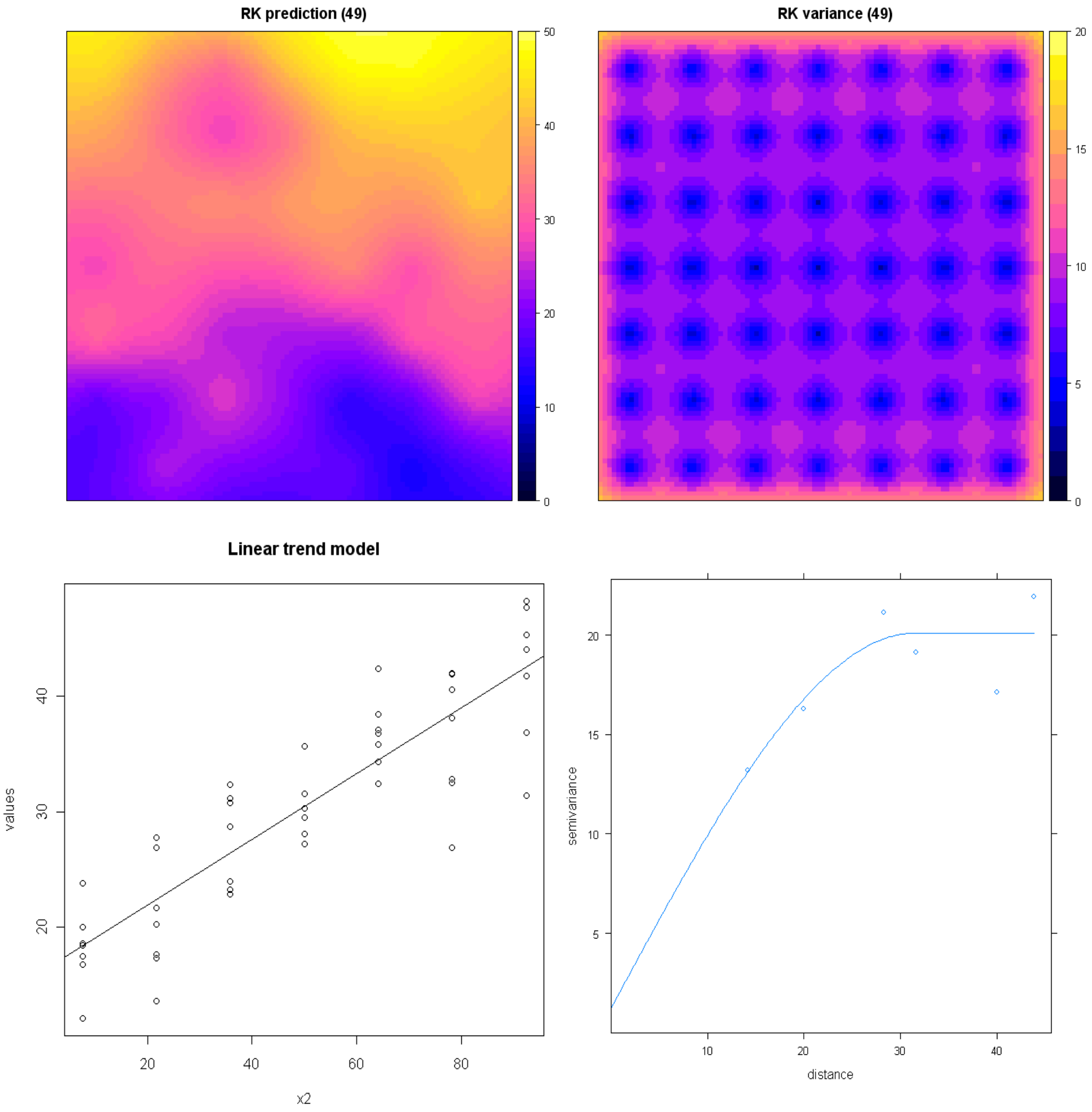


Figure 31: Regression Kriging prediction and variance maps using square Centred Grid Sampling with a sample size of 49. The variogram is modeled using a Spherical function with; nugget 1.23, psill 18.84 and range 31.39. The trend is linear with, intercept at y-axis 16.22 and a slope of 0.28.

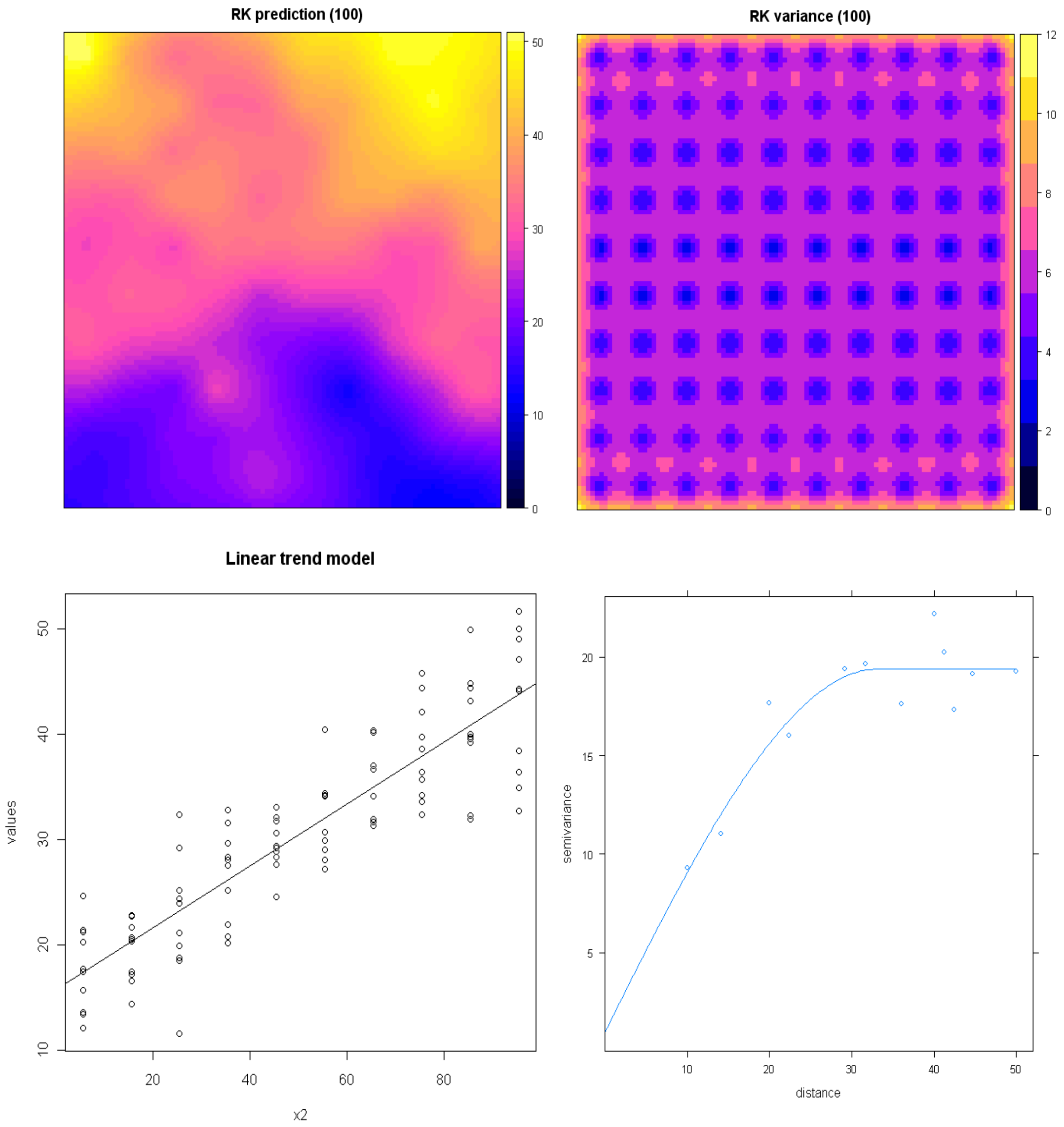


Figure 32: Regression Kriging prediction and variance maps using square Centred Grid Sampling with a sample size of 100. The variogram is modeled using a Spherical function with; nugget 0.98, psill 18.41 and range 33.14. The trend is linear with, intercept at y-axis 15.78 and a slope of 0.29.

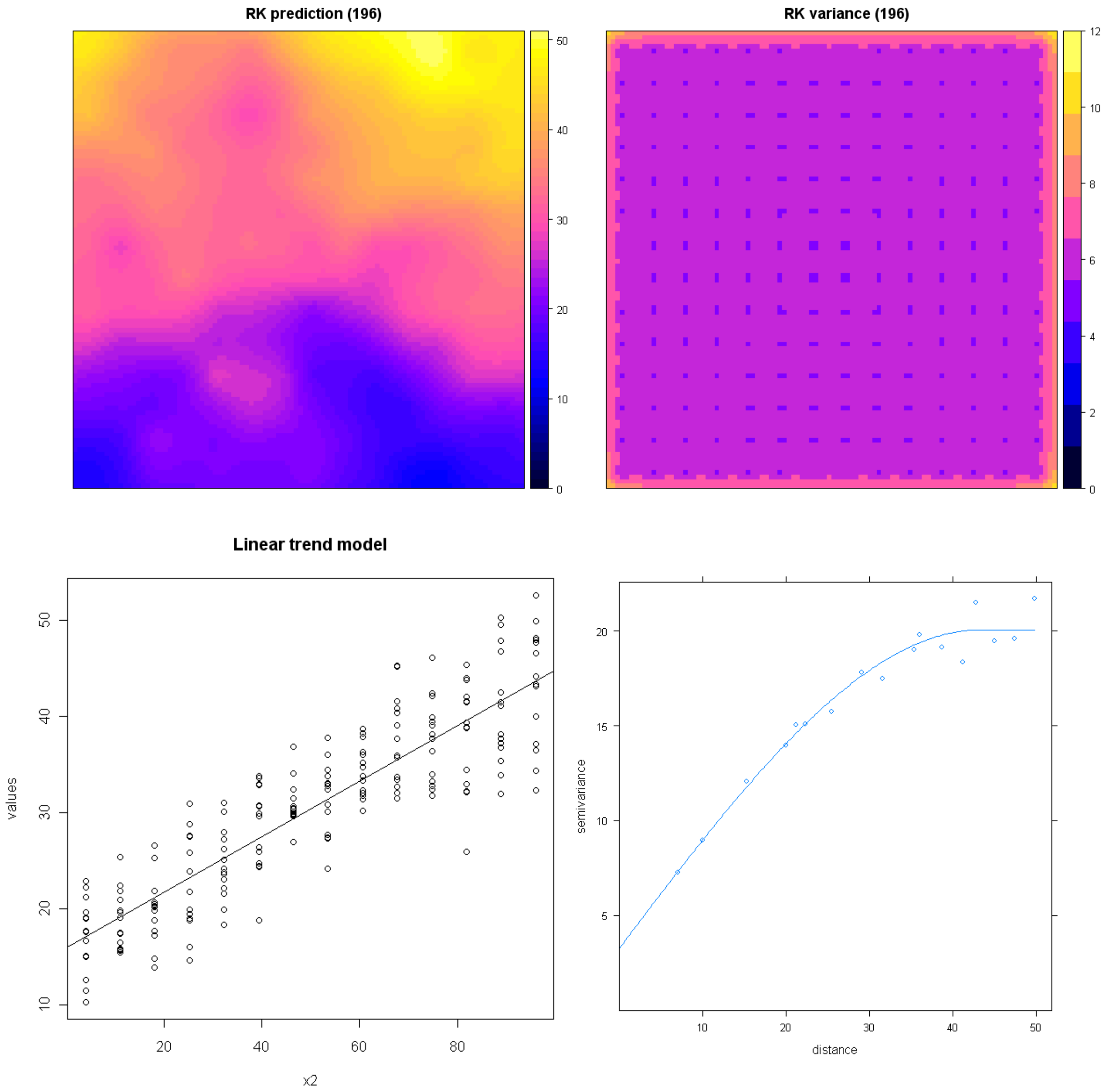


Figure 33: Regression Kriging prediction and variance maps using square Centred Grid Sampling with a sample size of 196. The variogram is modeled using a Spherical function with; nugget 3.28, psill 16.78 and range 43.32. The trend is linear with, intercept at y-axis 15.89 and a slope of 0.29.

3.2 *Hoeksche Waard case*

This case applies Simple Random Sampling and Stratified Random Sampling to an actual dataset. By doing this we are able to compare the possible gain in efficiency by stratification for scenario 1 and show what results would look like for scenario 2.
The dataset used for this method is one where different soil parameters are measured (pH, soil moisture, Na, Mg etc.) on four different farms in the Hoeksche Waard which is an area in the western part of The Netherlands, near to Rotterdam. The fields were divided into different regions according to management practices defined by the farmers themselves (Heijting et al., 2010).

Data preparation

The data were provided in excel sheets with the sample number in the first column, the zone (i.e. stratum) in the second, followed by the x-coordinates, the y-coordinates and the measured soil parameters. By importing these files together with the boundaries of the management zones (the strata) we are able to compare the results between treating the data as stratified or as non-stratified.
In the sections below only one of these fields is used, Field K, which has 25 sample locations divided evenly between 5 management zones (Figure 34).
Only one variable is used in the calculations of the following example; Potassium (K).

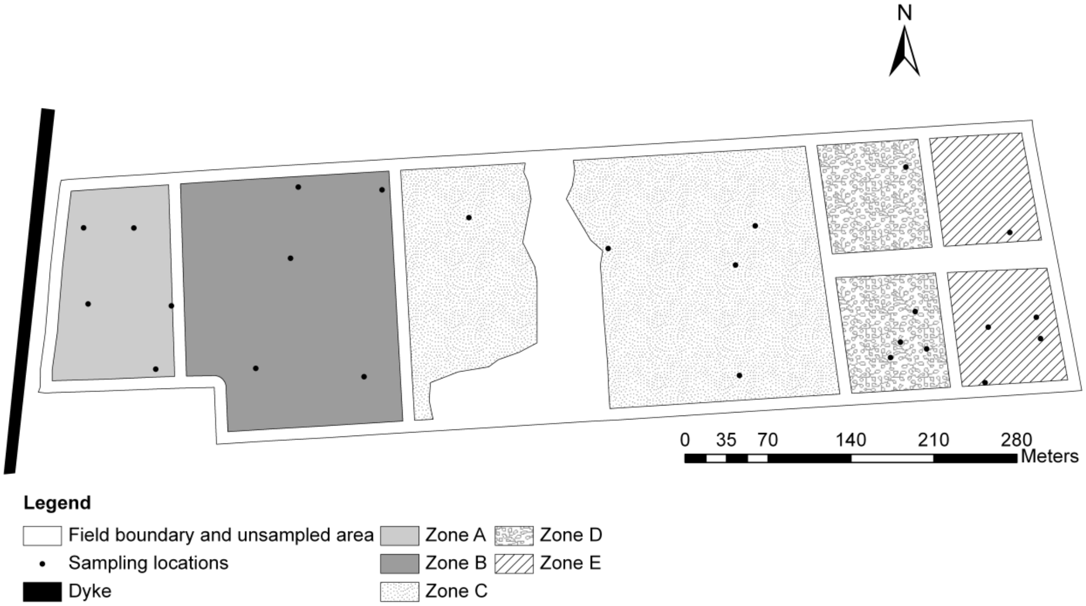


Figure 34: Agricultural "Field K" displaying the boundaries of management zones and all 25 sampling locations. (Heijting, et al., 2010)

Simple Random Sampling

Choosing this strategy means answering no to question 1 in the decision tree of Figure 6, which in term means that the user is really quite uncertain about whether the affordable sample size is big enough to cover the whole area or sample all possible strata.

The procedure of implementing Simple Random Sampling on the real dataset as described above is almost the same as with the Test Field in the previous sections. The only difference is that the data were already sampled.
This means that we cannot obtain a real random sample from the data like with the Test Field. Instead values of K are randomly selected from the table containing all sampled values and inference calculations are based on those values resulting in Table 13. Plotting the sample locations used results in Figure 35.

As was mentioned in section 2.3.3, validation in real cases can be done through a process called cross-validation. Hengl, et al., (2009) describe the process in case prediction; with Leave one out cross-validation each sample point is left out and predictions are made for that point, from which the error can be calculated. A kind of leave one out cross-validation can also be used with mean estimation, by estimating the mean with all points - 1 so that the error can be calculated as: estimated mean – real value point. This is done for a sample size of 25, resulting in:

ME = 0
RMSE = 15.29
MAE = 12.90

Table 13: Inference results from Simple Random Sampling on the Hoeksche Waard case

Sample size	Mean	Variance	Var of mean	95% conf. int.	
				left	right
2	124.00	128.00	64.00	99.66	148.34
4	81.50	57.00	14.25	76.26	86.74
6	95.00	188.80	31.47	89.40	100.60
8	94.75	296.79	37.10	89.78	99.72
10	95.80	147.96	14.80	93.09	98.51
15	97.67	200.67	13.38	95.65	99.68
20	96.00	250.42	12.52	94.35	97.65
25	96.32	224.39	8.98	95.09	97.55

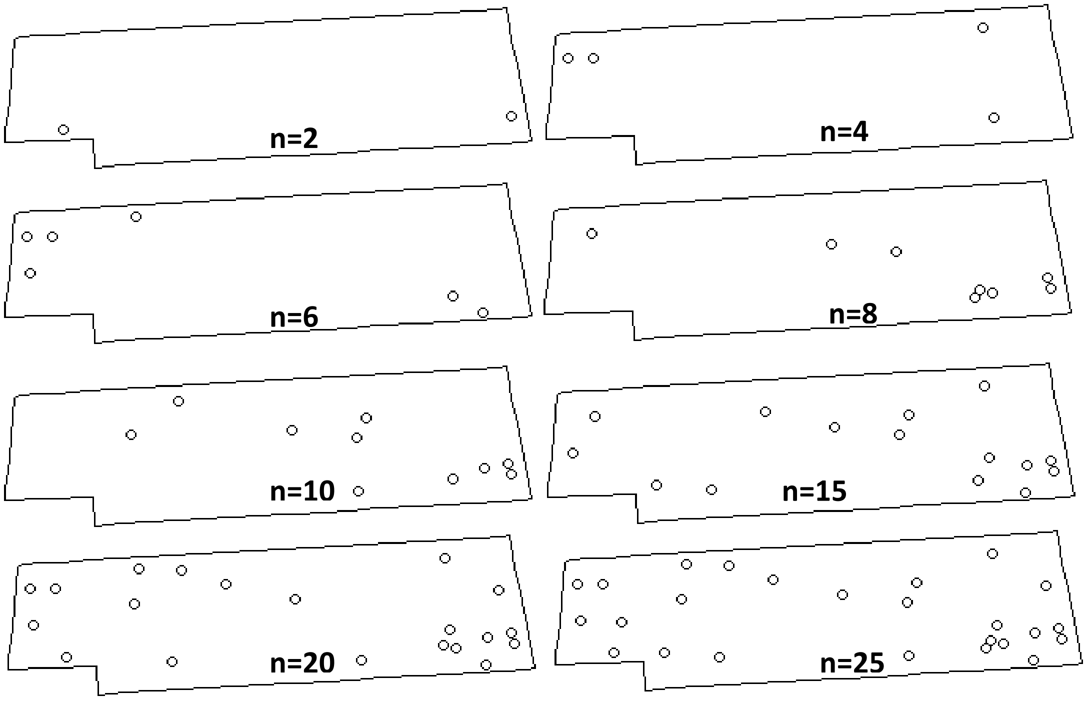


Figure 35: Simple Random Sampling pattern for different sample sizes 2,4,6,8,10,15, 20 and 25.

Stratified Random Sampling

Choosing this strategy means that the user can afford more than only a small sample size, meaning that question 1 in Figure 6 is answered with yes (at least 10 in this case). The aim is not to create a map no to question 2. There is information with which the area can be stratified (management zones); yes to question 3a. Finally the goal is to estimate the spatial mean; question 4b.

The inference results can be seen in Table 14 and the sample locations and field boundaries in Figure 36.

Just as with Simple Random Sampling an adjusted type of cross-validation can take place to for validation. As in the case of stratification the assumption is that the strata are more homogenous (more similar values) the error is measured per stratum: estimated mean stratum – real value point in stratum. Which, for a sample size of 25 results in: ME = 0, RMSE = 15.51 and MAE = 13.78.

Table 14: Inference results from Stratified Random Sampling

Sample size	Mean	Variance	Var of mean	95% conf. int.	
				left	right
10	94.46	101.23	13.06	91.91	97.00
15	94.90	41.13	10.65	93.11	96.70
20	94.64	20.67	4.85	93.61	95.66
25	93.95	32.50	4.81	93.04	94.85

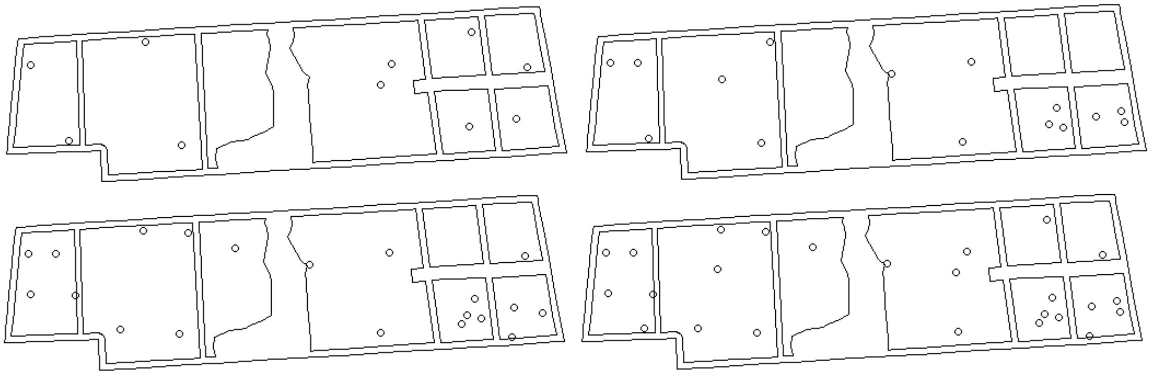


Figure 36: Stratified Random Sampling pattern for different sample sizes 2,3,4 and 5 per stratum.

Comparison

The results from Table 13 and Table 14, the mean and 95% confidence intervals of simple and stratified sampling are depicted in the boxplots in Figure 37.

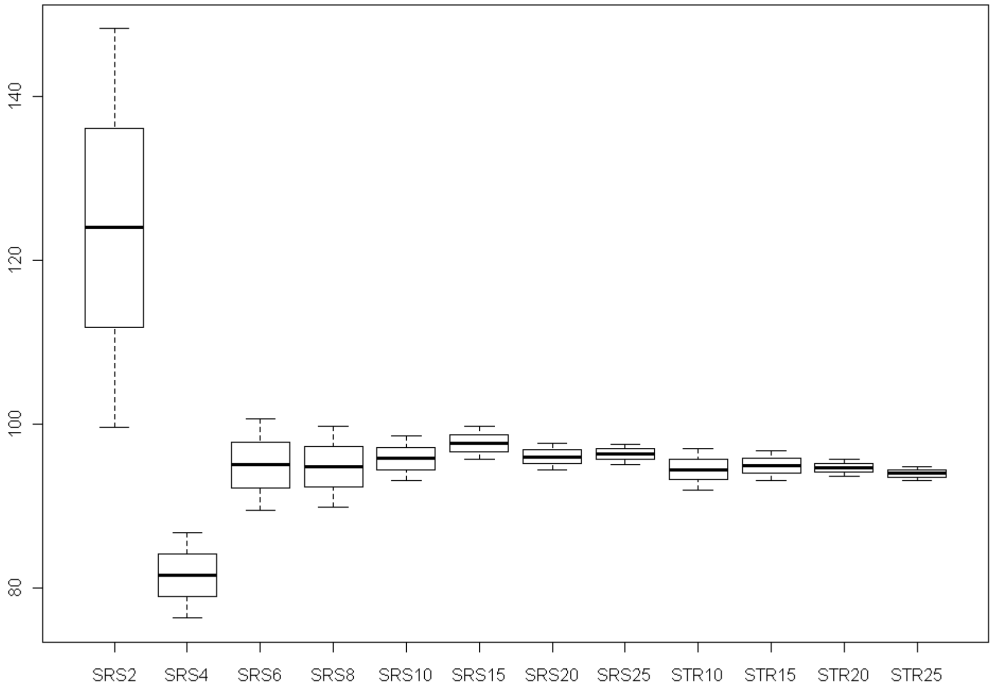


Figure 37 Mean and 95% confidence intervals for Simple Random Sampling and Strafied Random Sampling with different sample sizes from case 1. SRS2-SRS20 represent Simple Random Sampling with sample sizes; 2,4,6,8,10,15, 20 and 25. STR10-STR25 represent Stratified Random Sampling with sample sizes; 10,15,20 and 25.

4. Discussion

The objective of this thesis was to find suitable interpolation and inference methods for the educational project the Learning Map.
It was decided to follow the approach of building a decision framework guiding the user towards the use of a specific strategy instead of adopting a one fits all strategy. This decision framework is presented in Figure 6 and is in fact the result of literature research from section 2, resulting in one or more strategies (sample design + inference/interpolation) for each scenario.

Although there are similarities between this decision framework and others, like that of De Gruijter, et al., (2006) for selecting sampling strategies in case of inference of the spatial mean, that of Li and Heap (2008) and Hengl, et al. (2009) for selecting appropriate spatial prediction methods and that of Pebesma, et al. (2010) for automated interpolation, it is unique in the sense that it focuses neither on inference nor mapping alone.

The key question in Figure 6 is about spending resources efficiently, taking into account the factors mentioned in section 2.1 (i.e. goal and secondary information). As can be seen from the figure this can result in either mapping or inference, while the formerly mentioned decision trees already focus on mapping or inference specifically.

Because of the educational purpose of the Learning Map and the intended use by students, the main question is whether the decision tree of Figure 6 and the accompanying strategies answer to the LM objectives as defined in the introduction and of course whether it is practically feasible to apply in the current form.

The first of these two questions, about the LM objectives, can to some degree be dealt with by reflecting back on the 4 LM objectives that were stated in the introduction and relating them to the acquired results, which will be done in the conclusion. The question whether it is practically feasible however is more difficult to answer as it would require testing and feedback from students.

Before going on to the conclusions about the LM objectives and research questions, the results from applying the different strategies to the Test Field (3.1) and the real case (3.2) will be discussed per scenario as was done throughout this thesis.

Scenario 1: Spatial mean of study area

A general observation that can be made is that the results from the application of different sampling methods and sizes on the Test Field are to a large degree as was to be expected from literature. Larger sample sizes and spatial coverage lead to better estimates of the mean and less uncertainty (lower variance) because a more representative sample can be taken (De Gruijter, et al., 2006).

From Table 1, where the results of Simple Random Sampling are given it can be seen that the variance of the mean and thus the uncertainty about the estimate is very high for small

sample sizes, but even with a sample size of 50 it still gives a very high variance of the mean (up to 5 times as high) when comparing it to the results from other strategies for mean estimation which can be seen in Table 2, Table 3 and Table 4.
However when looking at Table 5 which gives the ME and the RMSE and the errors in Table 2, Table 3 and Table 4 (estimated mean – real mean), the difference with Simple Random Sampling and the other types of inference are not that big at all for the same sample size. This means that the added value of other sampling methods (i.e. not SRS) lies largely in a lower uncertainty represented by the variance of the mean rather than a better estimate.

From the variances of the mean obtained with Compact Geographical Stratification in Table 2 it can be seen that increasing the spatial coverage (more strata) indeed gave a more reliable estimate; the variance of the mean with 25 strata was more than 3 times smaller than with only 5 strata. This result corresponds with the description and aim of the method as described by Walvoort, et al. (2010).

Another good indication of the importance of spatial coverage were the results from Block Kriging in combination with Centred Grid Sampling (Table 3) which gave an even smaller variance of the mean and a very accurate prediction of the mean; only 0,17.
A drawback of this strategy is however that the variance given is the (Block) Kriging variance and therefore model based. From Equation 20 it can be seen that this variance is dependent on the variogram (model) and therefore changes when the model changes. This means that it is not p-unbiased and is susceptible to human errors (as the model function is determined by the user), which is exactly why De Gruijter, et al. (2006) indicate model based methods to be less suitable for mean estimation.

From Table 4 it can be seen that stratification is very good way to improve the estimation; both the estimate (error is only 0.06) and the variance of the mean are very good, which in this case is no real surprise as the boundaries of the strata are known exactly.
But also when comparing the results from the Hoeksche Waard case (section 3.2) in case of Simple Random Sampling (Table 11) with that of Stratified Random Sampling (Table 14) the results are very clear; the variance of the mean is significantly lower when using stratification.

In Figure 17 all strategies for scenario 1 are put together. It shows nicely what was already pointed out above; that the variance of the mean (and so the confidence limit) decreases with larger sample sizes and is especially low with good spatial coverage (Compact Geographical Stratification) and when more homogenous areas are separately sampled as with strata (Stratified Random Sampling).
However what is also striking to see here is that even the very small sample sizes give a fairly good idea about the spatial mean. Confirming the idea that when only a small sample size is affordable Simple Random Sampling offers a good solution.

Scenario 2: Chloropleth map

From the results in section 3.1.3 (Table 6) it can be seen that the results were relatively good in the sense that the errors (real mean – estimated mean) were small but that there were also considerable differences in accuracy and uncertainty (variance of mean). However one should be careful when using stratified random sampling. With small sample sizes per stratum the uncertainty about the estimate is high (the variance of the mean in Table 6 are fairly large) and one should therefore be cautious in making decisions based on such maps.

Another point of discussion is the distribution of the sample size over the different strata. In this case the samples were distributed evenly over the strata; it can be seen from the column "Var of mean"   and the confidence intervals in Table 6 that this results in a highly varying uncertainty about the reliability of the estimate.

Scenario 3: Continuous map

In section 3.1.4 Ordinary Kriging (OK) was applied to the Test Field with different sample sizes and sampling designs; it was assumed that there was no model about spatial variation yet and thus that it had to be estimated from the semivariogram.

From the performance criteria in Table 7 and the variograms in Figure 19, Figure 20, Figure 21, Figure 22 and Figure 23 it can be seen that sampling had quite a significant effect on the predictions made by OK and the way in which spatial variation is captured; this can be said for both sample size and sample design.

Especially when comparing the results from the three different sample sizes (49, 100 and 196) used with regular sampling (square grid) it becomes clear that there was a strong relation between sample size and the accuracy of the prediction (lower RMSE and MAE). This result is no surprise because the points used to predict from are closer. As Kriging only considers the distances of points to each other rather than the location to compute the kriging weights (Webster and Oliver, 2007), this results in a more accurate and less uncertain prediction as can be seen from the Kriging variance maps.

So 50 points does seem a minimum when it comes to Kriging as was already suggested in section 2.3.2 and that a sample size of 100 or even more might be more suitable. This is especially the case when considering that the users of the learning map are not likely to have much experience in variogram modelling. Therefore requiring a clear pattern.

This is illustrated by the variograms in Figure 19,Figure 20 and Figure 21 which contain very few points (lags) and might prove difficult to model. Figure 20 where a random sampling pattern was used, on the other hand has more lags at short range but shows a less clear pattern when comparing it to the variograms of the larger sample sizes (Figure 22 and Figure 23).

More surprising are the values in Table 7 that were obtained with a triangular sampling pattern; in terms of RMSE and MAE it shows an improvement of around 15% , which would make it very similar to regular (squared) sampling when considering that the sample also increased with about 14%. The ME however is very different; the mean error for triangular sampling is almost twice as large making it a more biased approach.

A plausible explanation for this is the presence of rectangular strata in the Test Field, in which the triangular configuration does not fit well.

It can be concluded from the section on OK that sampling on a square grid gave a sufficient result especially when dealing with a square area and/or strata because they simply fit better. This is in line with what Webster and Oliver (2007) already described That normally triangular grids give a slightly better result because the maximim kriging variance is minimized within the grid, but that mostly a grid is used that suits the user best. Which in this case is a square grid because of the shape of the strata and the total area.

Scenario 4 Continuous map using spatially exhaustive secondary information

The results from section 3.1.5, where Universal Kriging and Regression Kriging were applied to the Test Field using only the y coordinate in the regression analyses (Table 8 and Table 10) clearly show very little improvement when comparing it to the results with the same sample size using OK (Table 7).

As mentioned in section 3.1.5 a possible explanation for this is the influence of the strata which may counteract the trend at shorter distances. This explanation is supported by the fact that when using a Test Field containing only the GRF and the trend to sample from, the results were much better as Table 11 shows.
Another approach is to include the strata in the regression analyses; this was done for both Universal Kriging (see Figure 27, Figure 28, Figure 29 and Table 9) and Regression Kriging (Table 12). The prediction and kriging variance maps of the aforementioned figures show very clearly the boundaries of the strata and also the results shown in both tables show a significant improvement (in RMSE and MAE) compared to Ordinary Kriging (Table 7). One problem that did arise here was that with a small sample size, 49 in this case, it was difficult to fit a suitable function in the variogram (even more than with OK).

Because UK and RK are very similar methods which was as also argued by Hengl, et al. (2009), it does not make much sense to use both of them in the Learning Map.
Seen the fact that RK requires the users themselves to separate the trend from the residuals as opposed to UK where it is “hidden” within the Kriging process, RK might give users a more thorough insight in the process involved in making prediction maps. This perhaps makes RK a more suitable method for the Learning Map.

5. Conclusions

When looking back at the research objective and the research questions as stated in the introduction it can be said that the objective has been achieved; interpolation and inference methods have been found for interpretation of sampled data in the Learning Map. Furthermore a framework has been developed to guide users of the learning map towards the selection of the most suitable sampling strategy given their survey purpose and limitations in terms of sample size and auxiliary information.

Research questions

The first research question; *Which scenarios are relevant for the learning map project given the project background?,* has been dealt with in section 2.1. Four scenarios were created, covering a range of situations that are thought to be representative for scientific field work but at the same time are limited enough to be implemented in a short period of time.

The second question; *What are criteria on which an interpolation or inference method can be judged?,* was answered in section 2.3.3. Three much used criteria for accuracy quantification were selected; the ME, RMSE and MAE. Both the RMSE and the MAE were used because the RMSE might place too much weight on large errors due to the squaring of errors (Li and Heap, 2008). However it can be seen from the results from the Test Field for scenarios 2,3 and 4 that this does not really pose a problem and that the RMSE alone is good enough.

The RMSE does not seem to be a very good measure for global mean estimation (Table 5) as it indeed placed too much weight on large errors; the ME in this case gives a much better impression of the accuracy.

The third question; *Which interpolation or inference methods are most suitable given the different scenarios?*, was addressed in sections 2.2, 2.3 and 2.4.. The formulation of this research question was in fact not adequate because an important part of the answering deals with sampling (section 2.2) rather than inference or interpolation techniques (section 2.3). The result of this literature research therefore does not give only inference or prediction techniques for each scenario (through Figure 6) but also an accompanying sampling method; the combination of the two being what De Gruijter, et al. (2006) call a sampling strategy. A critical note in relation to the selected strategies is that for continuous map making only geo-statistical (Kriging) methods were included which may give students a somewhat limited idea of mapping methods. However as was already pointed out in section 2.3.2 the advantage is that students are already familiar with those techniques meaning that more time can be spent in the field.

The fourth research question; *Can these interpolation methods be implemented for (automatic) interpolation or inference within a statistical computing environment?,* can be answered with a full yes; all strategies have been implemented using R.

R proved to be a very suitable implementation environment for all strategies because specific packages make it possible to handle spatial data (see section 2.5) efficiently. Furthermore R is very flexible, allowing users to define almost everything themselves in detail (e.g. field boundaries, strata etc.) and offers good visualization. A downside of this

could be that considerable programming, testing and refinement would be needed before application of all these methods on a real case is possible.
The answer to the last research question; *How do these methods perform judged on the formulated criteria?*, can be found in the implementation of the strategies in sections 3.1, 3.2 and the resulting discussion . Because most of the methods which were used are fairly commonly applied and therefore well documented in methodological literature on sampling and inference (De Gruijter, et al., 2006; Cochran, 1977) and geo-statistics (Hengl, et al., 2009) the results were mostly as was expected.

Learning Map educational value

Although no spectacular results were obtained whilst implementing the selected sampling strategies on the Test Field and the Hoeksche Waard case in sections 3.1 and 3.2. They do clearly illustrate the effects of the choices made through the use of the decision framework of Figure 6. By taking note of these effects students will gain more understanding about spatial surveys and field work, thus satisfying the four LM objectives set out in the introduction:

1.) *Are introduced to different types of data collection (sampling) and learn about the consequences this has.*
Four different sampling methods were introduced, three design based and one model based method. Choosing for either has certain implications for the set-up of your survey; design based methods are often more laborious and less efficient in data gathering in order to ensure p-unbiasedness. While model based methods require the user to make a model and therefore are more difficult in the interpretation phase. Important here is the fundamental difference between model based and design based methods and especially what this means for the validity of the outcome.
2.) *Learn about different ways of using sampled data to generate a result that matches the goal.*
What became clear is that sampling and inference or interpolation cannot be seen separately and together they form one sampling strategy.
Through the introduction of the scenarios and incorporating those in the framework of Figure 6, three different goals are included covering the whole range of possible outputs in terms of spatial resolution. Also students learn about ways in which an estimation or prediction can be enhanced by using secondary knowledge.
3.) *Are able to use data collected by others as to show the advantage of a real time measurement system.*
It is difficult to assess to what extent this learning goal is answered as there is not yet a fully developed scheme stating which groups are doing what and how samples from other groups can be used.
4.) *Interpret their own results and relate this to the choices made in sampling and/or inference/interpolation methodology and compare this to the results of other students.*
This is evaluation could take place in a plenary session after the fieldwork. Looking at the results obtained from the Test Field this discussion/comparison should be possible by looking at parameters like the ME, RMSE and variance of the mean.

6. Recommendations

If it were chosen to continue with the Learning Map in the form as described in this thesis. Which means using these mapping and inference techniques in combination with a decision tree like in Figure 6, there is still considerable work left to be done.

This work first of all consists of practical issues like setting up a (web based) system in which the connection can be made between the students sampling in the field and the server calculating results. Work on such systems has already been done for automatic interpolation, De Jesus, et al. (2007) for instance discuss a web based application of Automap using a Server Oriented Architecture.

Another important issue is what was already touched upon with the assessment of the LM objectives. That in order to be able to illustrate the added value of real-time systems it is important that also the measurements of others can be used or displayed. With the proposed methodology every group will sample according to a specific scheme (regular, random or stratified) making it difficult to include measurements of others. Possibly a prediction map could be added which uses all measurements to illustrate the added value of web based systems.

7. References

Brenning, C. , Dubois, G., 2008. Towards generic real-time mapping algorithms for environmental monitoring and emergency detection. *Stochastic Environmental Research and Risk Assessment,* 22 , pp. 601–611.

Brus, D.J., De Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). *Geoderma,* 80, pp. 1-44.

Burrough, P.A., McDonnel, R.A., 1998. Optimal interpolation using geostatistics. In: Principles of geographical information systems. Oxford University Press, pp. 132-161.

Cochran, W.G., 1977. Sampling Techniques. Chichester: John Wiley & Sons Ltd.
Brus, D.J., De Gruijter J.J. and van Groenigen, J.W. , 2007. Designing spatial coverage samples using the k-means clustering algorithm. In: Chapter 14 of Digital Soil Mapping; An Introductory Perspective. Amsterdam: Elsevier.

De Bruin, S., Wielemaker, W.G., Molenaar, M., Formalisation of soil-landscape knowledge through Interactive hierarchical disaggregation. *Geoderma,* 91, pp. 151-172.

De Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Berlin: Springer-Verlag.

D'Orazio, M., 2003. Estimating the Variance of the Sample Mean in Two-Dimensional Systematic Sampling. *Journal of Agricultural, Biological, and Environmental Statistics,* 8 (3), pp. 280-295.

EUR , 2005. Automatic mapping algorithms for routine and emergency monitoring data. Report on the Spatial Interpolation Comparison (SIC2004) exercise. In: Dubois G. (ed) EUR 21595 EN. Office for Official Publications of the European Communities, Luxembourg, 150 pp.

Gibbs, J.P., Droege, S., Eagle, p., 1998. Monitoring populations of plants and animals. *BioScience,* 48 (11), pp. 935-940.

Goovaerts, P., 1997. Applied Geostatistics for Natural Resources Evaluation. New York: Oxford University Press, Inc.

Groat, C.G., 2004. Seismographs, sensors, and satellites: better technology for safer communities. *Technology in Society,* 26, pp.169–179.

Heijting, S., De Bruin, S., Bregt, A.K., 2010. The arable farmer as the assessor of within-field soil variation. [online] Precision agriculture . Available at: http://www.springerlink.com/content/q532348x0r752q21

Hengl, T., 2009. A Practical Guide to Geostatistical Mapping. 2nd ed. Amsterdam : University of Amsterdam.

Hiemstra, P.H., Pebesma, E.J., Twenhöfel, C.J.W, Heuvelink, G.B.M., 2007. Automatic real-time interpolation of radiation hazards: a prototype and system architecture considerations. *International Journal of Spatial Data Infrastructures Research*, 3, pp. 58-72.

Jesus, J. de, Dubois, G., Hiemstra, p., 2008. Web-based geostatistics using WPS. GI-Days conference, 17-18 June 2008. http://intamap.geo.uu.nl/~paul/files/paper_GIDays2008.pdf [accessed 13 September 2010]

Kerry, R., Oliver, M.A., 2007. Comparing sampling needs for variograms of soil properties computed by the method of moments and residual maximum likelihood. *Geoderma*, 140 (4), pp. 383-396.

Li, J., Heap, D., 2008. A Review of Spatial Interpolation Methods for Environmental Scientist.[Online] Geoscience Australia. Available at: https://www.ga.gov.au/products/servlet/controller?event=GEOCAT_DETAILS&catno=68229 [accessed 21 September 2010]

McBratney, A.B., Pringle, M.J., 1999. Estimating Average and Proportional Variograms of Soil Properties and Their Potential Use in Precision Agriculture. *Precision Agriculture*, 1, pp. 125-152.

MOBIlearn, 2005. Guidelines for learning/teaching/tutoring in a mobile environment. [Online] available at: http://www.mobilearn.org/download/results/public_deliverables/MOBIlearn_D4.1_Final.pdf [accessed 14 September 2010]

Pebesma, E. J., 2004. Multivariable geostatistics in S: the gstat package. *Computers and Geosciences*, 30 (7), pp. 683–691.

Pebesma, E.J., R.S. Bivand, 2005. Classes and methods for spatial data in R. R News 5 (2).

Pebesma, E., Cornford, D., Dubios, G., Heuvelink, G.B.M., Hristopulos, D., Pilz, J., Stohlker, U., Morin, G., Skøien, J.O., 2010. INTAMAP: The design and implementation of an interoperable automated interpolation web service. *Computers and Geosciences*, 37, pp. 343-352.

R Development Core Team, 2010. R:A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 / http://www.R-project.org

Walvoort, D.J.J., Brus, D.J., De Gruijter, J.J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. *Computers & Geosciences*, 36, pp. 1261–1267.

Webster, R., and Oliver, M.A., 2007. Geostatistics for environmental scientists. 2nd ed. Chichester: John Wiley & Sons Ltd.