INTERACTIVE TOOLS TO LEARN GEOSTATISTICAL BASIC CONCEPTS

*A. Vervoort[1], A. Govaerts[2] and P. Darius[3]

[1]Department of Civil Engineering, KU Leuven
Kasteelpark Arenberg 40
3001 Leuven, Belgium
*(*Corresponding author: andre.vervoort@bwk.kuleuven.be)*

[2]formerly Department of Civil Engineering, KU Leuven
Kasteelpark Arenberg 40
3001 Leuven, Belgium

[3]MeBios and Leuven Statistics Center, KU Leuven
Kasteelpark Arenberg 30
3001 Leuven, Belgium

INTERACTIVE TOOLS TO LEARN GEOSTATISTICAL BASIC CONCEPTS

ABSTRACT

The concepts of geostatistics are commonly difficult to understand for students, even for those with a strong mathematical background. A main problem is certainly the link between the variation of the parameter value in space or time, and the calculated experimental semivariogram as a function of the interdistance, but over the entire area. Another problem is to differentiate between directional and omni-directional semivariograms and the way they have to be interpreted. Furthermore first time users of geostatistics often do not want to struggle through different books with the whole geostatistical background before using it. Too often this leads to using geostatistics as a black box with all the risks linked to it. That is the main reason that we developed some interactive tools based on the free public domain software R. We aim by developing three different modules to assist individual first time users, but also regular students and teachers to get a feeling of the importance of the various aspects in conducting a good geostatistical study. The modules show directly e.g. the impact of deleting, moving or adding data points, of changing parameters of the semivariogram model and of the geometrical parameters during kriging. The first module allows making the link between the variation of the parameter values in one direction and the calculated experimental semivariogram, and this for different given datasets. The second module covers two dimensional datasets and apart from calculating the semivariogram values, models can be fitted to the calculated data. A distinction can be made between omni-directional and directional semivariograms, including the effect of the tolerance on the angle and lag-distance. The third module focuses on ordinary kriging of point values in two dimensions. The aim of the interactive tools developed is limited to educational purposes only, i.e. to visualize some basic concepts for different pre-selected types of spatial variation. They do not intend to do a geostatistical study on a certain dataset (for this purpose there exists already a broad spectrum of good software packages). The original intention of developing the three modules was aimed at users with no or a limited amount of geostatistical experience. However, these modules seem also very useful to remain critical for experienced users. One should never forget that in a real application one does not have a complete view of the phenomenon studied. The geostatistical interactive modules can be accessed through: www.bwk.kuleuven.be/geostatistics . We compiled a manual to facilitate the use of the modules and to guide the user we formulated several questions to be solved for each module.

KEYWORDS

Geostatistics, educational tool, semivariogram, kriging, public domain software R

INTRODUCTION

The concepts of geostatistics are commonly difficult to understand for students, even for those with a strong mathematical background. A main problem is certainly the link between the variation of the parameter value in space or time, and the calculated experimental semivariogram as a function of the interdistance, but over the entire area investigated. Furthermore first time users of geostatistics often do not want to struggle through different books with the whole geostatistical background before using it. Too often this leads to using geostatistics as a black box with all the risks linked to it. This is in particular a risk when using integrated software packages, like for example GIS-software. A geostatistical study should not be a trial and error process in which some buttons are hit at random or tried out to see what the effect could be. The risk is that one aims to get the nicest map but not the best result.

Although kriging claims to be the best linear unbiased estimator, it only fulfils this role when the spatial variation is properly translated in a correct semivariogram model. And the latter is significantly affected by the sampling campaign and the way the semivariogram is calculated and modelled afterwards. Quite often the data is blindly imported in a software program, which calculates an experimental semivariogram, determines the 'best' semivariogram model and computes kriging estimates. These results are then considered as the only and best results, and often get a high quality label as they are based on geostatistics, while the opposite is true.

The developed R-based interactive modules aim to assist students and teachers of geostatistical courses. Individual first time users can use these modules too, i.e. to get a feeling of the importance of the various aspects in conducting a good geostatistical study. In this way, valuable experience can be gained prior to starting real estimation projects, without being confronted with the consequences of a bad study. The geostatistical interactive modules can be accessed through: www.bwk.kuleuven.be/geostatistics . We compiled a manual to facilitate the use of the modules and to guide the user.

R-BASED INTERACTIVE MODULES

The statistical package R (www.r-project.org), which is free and open source software, is used to develop the modules presented in this paper. The benefits of this package are: (1) it is available for free; (2) it is available for multiple operating systems; (3) R has a large active community and all the programs written in R are shared with the other R users. After a thorough study of the capabilities of R, it was clear that R has sufficient opportunities to develop interactive, visually appealing modules. In this case the tcltk-package of R was used to develop the modules. In addition, the already existing 'packages' in R can be re-used for the mathematical and (geo)-statistical calculations.

In this paper, we describe a number of generally accessible R-based interactive modules, developed by us, which can be used as part of a geostatistical or GIS-course, but also by an individual user getting familiarized with the key concepts of geostatistics. These modules should allow the student or practitioner to see directly the impact of deleting, moving or adding data points, of changing parameters when calculating an experimental semivariogram (e.g. lag distance, lag tolerance, directional tolerance, etc.), of changing parameters of the semivariogram model and of the geometrical parameters during kriging. The modules are compiled in such a way that they work completely autonomously. This means a.o. that it is not required to have a dataset of your own. Those modules are therefore not meant to do a geostatistical study on a certain dataset (for this purpose there exists already a broad spectrum of good software packages). After being familiarized with the key concepts of geostatistics it should be easier for the user to do a good quality geostatistical study on their own dataset and to do a good interpretation of these results.

BASIC CONCEPTS OF GEOSTATISTICS

A geostatistical study always starts by determining the semivariogram(s), based on a number of sampling points (e.g. Journel & Huijbrechts, 1978 and Webster & Oliver, 2007). The experimental semivariogram is defined as a function of the lag distance h. It equals half the quadratic average of the difference between data at a particular lag distance h:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} \left\{ F(x_i + h) - F(x_i) \right\}^2$$

where x = the coordinates in space; N(h) = the number of data pairs separated by h; F(x) the parameter value in x. Hence, the experimental semivariogram describes the correlation of two values of the same parameter in space (or time). Each individual point of a semivariogram must be based on enough data pairs. The obtained semivariogram is fitted by a model. Most often the range, the sill and the nugget value are the three most important parameters describing these models. The range is the maximum lag distance for which the values at the two points are still (partly) dependent of each other. The sill is the plateau that the semivariogram reaches at the range. The nugget is the vertical jump from value 0 at the origin to the semivariogram value at extremely small separation distances. Four models are available in the interactive modules (Govaerts & Vervoort, 2011): spherical, bounded linear, power and periodic. As the individual estimated points of a semivariogram are significantly affected by the sampling geometry (size of the sampling campaign and absolute and relative location of the individual samples), the final chosen semivariogram model is also influenced by these parameters. The interactive modules allow to get a better understanding of the link between sampling geometry and semivariogram model for different spatial variations.

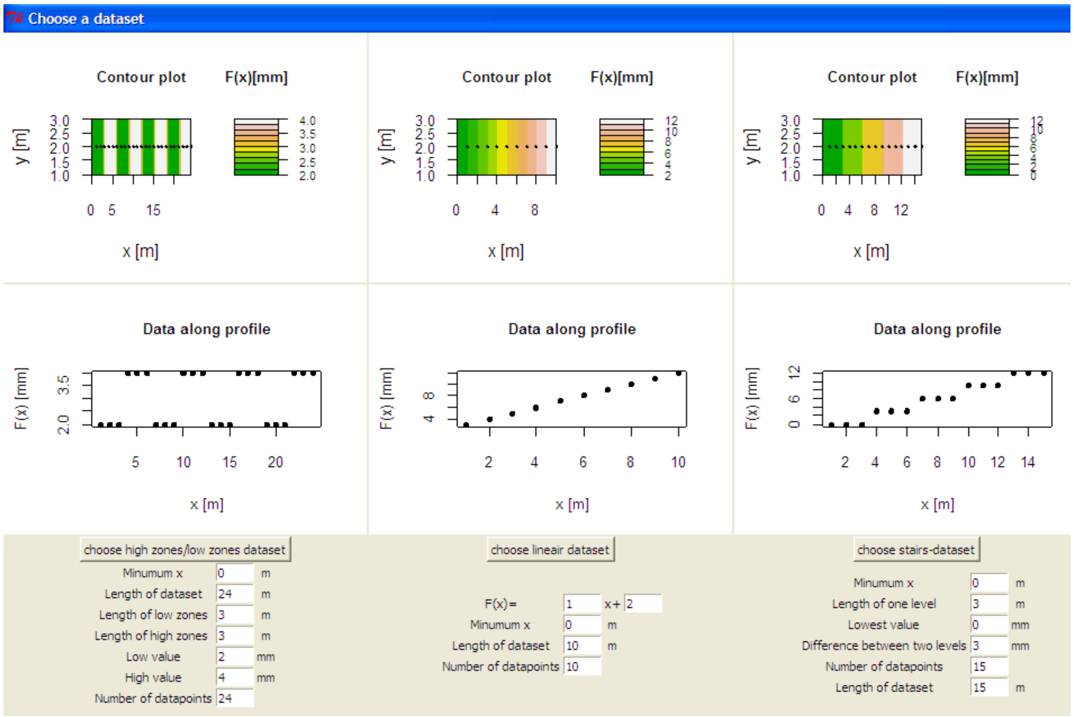
The information of the semivariogram model is used to estimate or krige the parameter at a specific location. Geostatistical estimation, or kriging, refers to techniques that provide the best linear unbiased estimators (BLUE) of unknown properties (Journel & Huijbrechts, 1978). In other words, kriging determines the weights for which the estimation variance is the smallest. In comparison to classical statistical interpolation techniques (e.g. inverse distance weights, 1/n, etc.), the spatial information is explicitly taken into account to determine the weights. Referring to the remarks formulated above, it is logic that the quality of kriging depends on a well defined semivariogram model, which needs an accurate estimation of the individual points of the experimental semivariogram. Apart from the quality of the semivariogram, the geometry between the samples and point(s) to be estimated plays a role also.

PRESENTATION OF THREE DEVELOPED MODULES

Module 1, Experimental Semivariogram in 1D

The first module focuses on the experimental semivariogram of different datasets in 1 dimension. The main aim of this module is to make the link between some simple variations of the parameter value along a line and the resulting calculated experimental semivariogram. Although this module is limited to 1D, the contour plots in 2D are also presented, so that the user gets already familiarized with this. Making the link between contour plots and a classic graph F(x) vs. x is not always that easy for students, apart from the link with the experimental semivariogram. In a first screen the user can choose one out of three datasets (Figure 1): (1) a dataset with an alternation of high zones (one constant value) and low zones (one constant value); (2) a linear rising dataset; (3) a stepwise rising dataset. For each dataset it is possible to adjust different parameters describing the datasets (e.g. minimum x, length of the dataset and number of data points). Afterwards a second screen displays the selected dataset (Figure 2). It was decided not to show the experimental semivariogram immediately; it is only visible after ticking a box. This encourages the user to first reflect on how this semivariogram could look like or how it would change if one for example delete or add a sample point. One can make some small hand calculations (e.g. number of pairs for each lag distance or semivariogram value). Changing the x or the F(x) value of a data point or deleting a data point can be easily done by the user on this screen and one immediately sees the effect on the experimental semivariogram.

As an illustrative example, the number of samples for the same phenomena is changed in Figure 2. The phenomena itself is an alternation of high and low zones, each with a width of 3 m and the total interval studied is 24 m. First, 24 samples are taken at an equal distance of 1 m (Figure 2.a). This results in 3 samples for each zone of either a high or low value. The calculated experimental semivarogram is logically a perfect cyclic one with a periodicity of 6 m. Such a sampling strategy is the most preferable one, at least if one knows the phenomena beforehand. In reality this is normally not the case. Therefore two other sampling strategies are presented, i.e. 20 samples (Figure 2.b) and 10 samples (Figure 2.c), both at equal distance and placed over the entire length. For 20 samples, this results in still 3 samples for the high zones, but systematically there are only 2 samples in the low zones. However, the cyclic character of the semivariogram is still clearly visible with the same periodicity, but a smaller maximum variance is observed. If the number of samples is further decreased (i.e. to 10), it becomes difficult to see the alternation in the F(x)-curve and certainly to see the periodicity in the semivariogram.



Minumum x 0 m
Length of dataset 24 m
Length of low zones 3 m
Length of high zones 3 m
Low value 2 mm
High value 4 mm
Number of datapoints 24

F(x)= 1 x+ 2
Minumum x 0 m
Length of dataset 10 m
Number of datapoints 10

Minumum x 0 m
Length of one level 3 m
Lowest value 0 mm
Difference between two levels 3 mm
Number of datapoints 15
Length of dataset 15 m

Figure 1 - First screen of Module 1, selection of one of the three dataset provided (i.e. simple variations in 1 direction)

For each of the three basic datasets, a list of questions is provided in the manual (Govaerts & Vervoort, 2011). For example for the dataset with the alternating zones, some of these questions are:

- What is the effect if one adds the same value to the high value and the low value?
- How does the experimental semivariogram of a dataset with high and low zones look like if there are several high zones and several low zones with a different length?
- How does the experimental semivariogram of a dataset with only one low zone and one high zone, with the same length look like? (e.g.: minimum x=0 m, maximum x=20 m, length of low zone=10 m, length of high zone=10 m, number of data-points=20).
- Look at the effect of changing or deleting a data-point.

Module 2, Modelling of Experimental Semivariogram in 1D and 2D

Module 2 shows the experimental semivariogram of different datasets in 1D or 2D (Vervoort, Govaerts, & Darius, 2011). There are four datasets readily available. The four datasets are (Figure 3): a first artificial dataset which is constant in the y-direction and linear in the x-direction; a second artificial dataset which is constant in the y-direction and showing an alternation of high and low zones in the x-direction; a realistic lognormal dataset (based on simulated lead contamination dataset using data from Houlding (2000)); a realistic normal dataset (based on data of elevation measurements from Isaaks and Srivastava (1989)). Each dataset contains 10,000 points at a regular grid within a zone of 10 m x 10 m. Hence, the separation distance is only 0.1 m along x and y. Every point is clickable. For simplicity, mm is used as unit for all four datasets, as there is no direct link with the original application. The user can also change the datasets, if he or she wishes to do this.

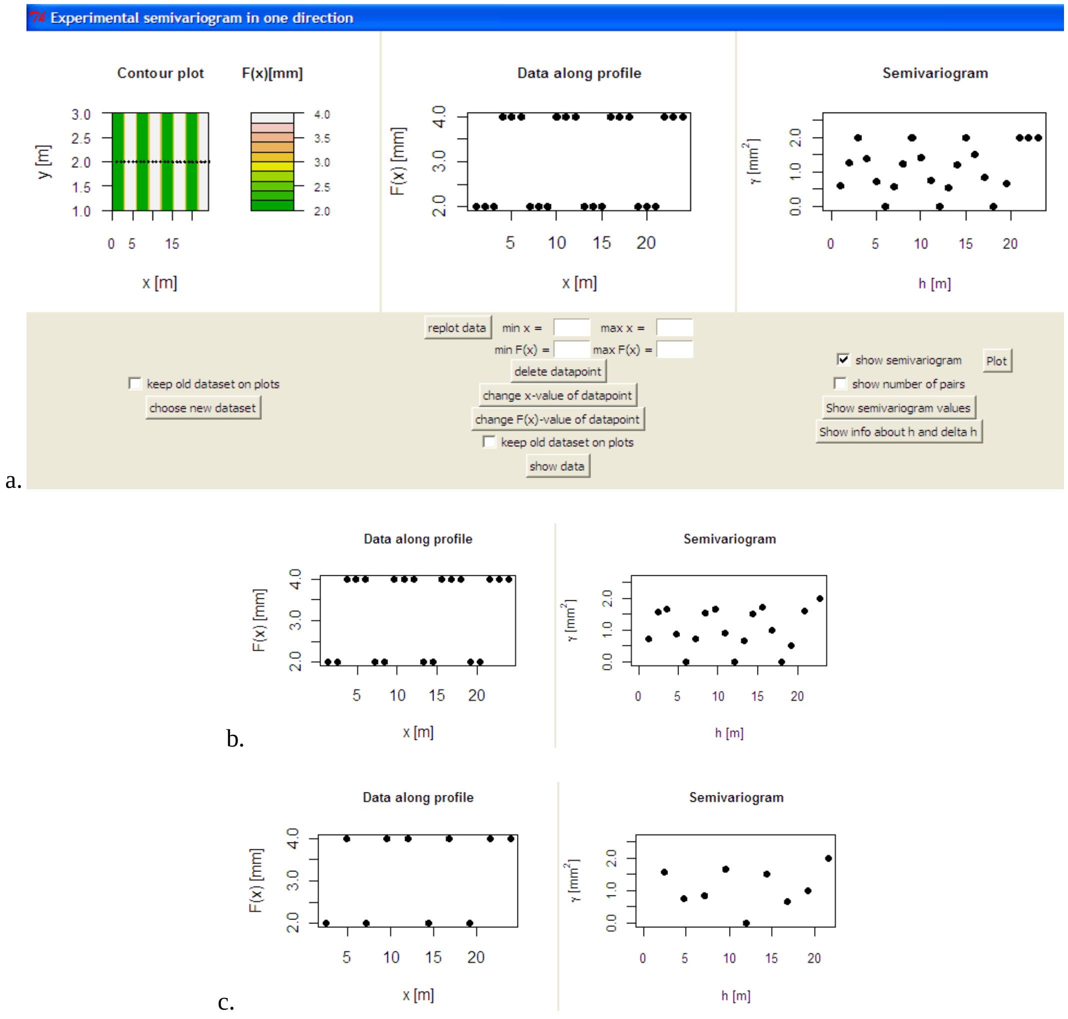


Figure 2 - Second screen of Module 1, computation of the experimental semivariogram in 1 direction for a phenomena with an alternation of high and low zones with a width of 3 m over a total length of 24 m: a. 24 samples at equal distance; b. 20 samples at equal distance over the total length of 24 m; c. 10 samples at equal distance over the total length of 24 m

Unlike the first module, where everything is kept as simple as possible, the sampling points are not determined in advance at certain positions with equal spacing. The user can choose the positions of the sampling points (Figure 4). The positions can be along a regular mesh, at random, or individually determined. The user can also choose the way that the semivariogram is calculated (directional vs. omni-directional, tolerance on angle, lag-distance and maximum lag). In this way it is shown to what extent the number and positions of the samples affect the calculated experimental semivariogram and in which way the mentioned parameters have an influence on the calculation of the semivariogram. As for Module 1, one can also extract a table with all relevant values of the experimental semivariogram (i.e. lag-distance, calculated values and number of pairs for each lag-distance). A second important element of this module is to model the experimental semivariogram. This modelling is done in real-time on the screen. The user selects a particular type of model and the values of its parameters. The model is plotted and one can immediately see whether it matches well the calculated experimental semivariogram values or not. One can then start adjusting these parameters until a good fit is reached. To assist the user the variance of the samples is also plotted, as it gives an indication of the sill value (if applicable). A help screen is available with information on the various available semivariogram models. In Figure 4, two different sampling locations are considered, i.e. 200 data points at random and 200 data points on a regular grid of 20 on 10 points. The omni-directional semivariogram is calculated. In this particular case, the variance for both data points is similar, as is the variation of the semivariogram.

Again a list of questions is provided in the manual (Govaerts & Vervoort, 2011), e.g.:

- Take 100 regular gridded samples. Does the experimental semivariogram look different than with random sampling?
- Take samples on a line with either constant y, or constant x, or x=y. Which models do describe the different semivariograms?
- Take 100 regular gridded samples: Which models do describe the omnidirectional semivariogram and the directional semivariograms NS, EW, NE-SW and NW-SE?
- Define the two major directions of the dataset.

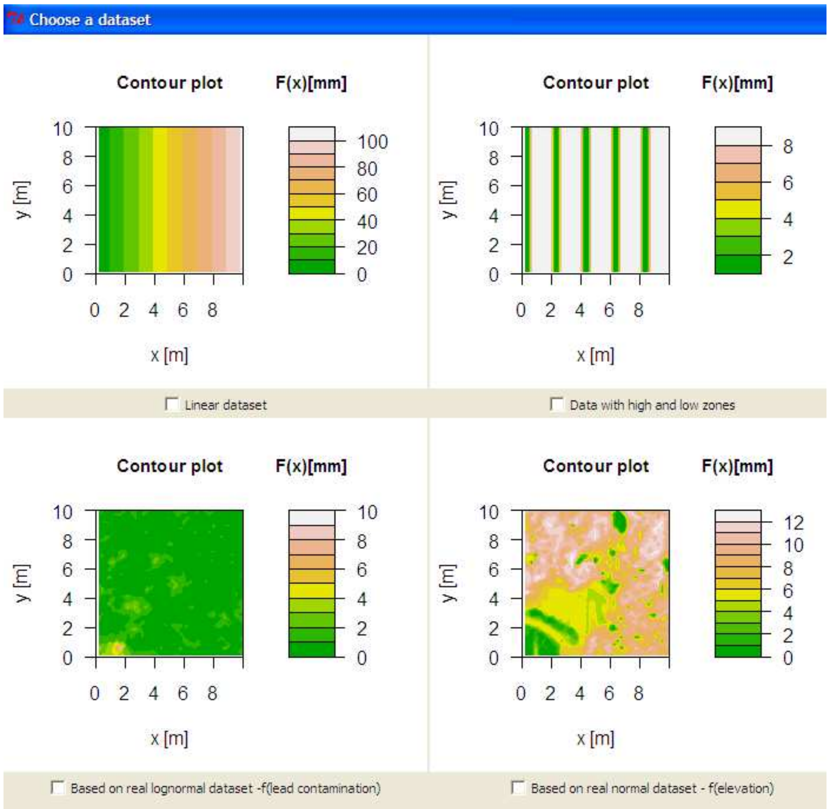
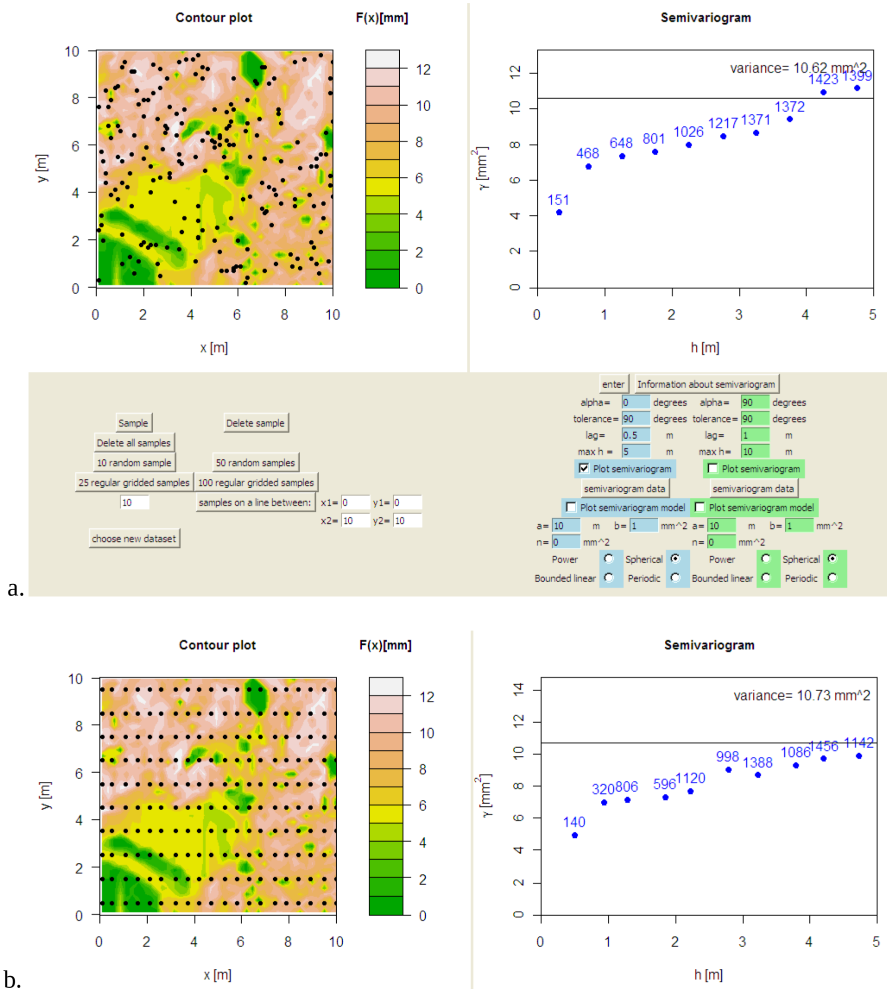


Figure 3 - First screen of Module 2, selection of one of the four dataset provided

Module 3, Estimating Values in 1D and 2D

The third module is about kriging. For the moment, only points can be estimated. The module contains the two real datasets of the second module. The parameters of a suitable semivariogram model should be noted when using Module 2 and introduced in the second screen of Module 3. In Module 3, one chooses new samples, which can be again along a regular mesh, at random, or individually determined (see Figure 5.a). One also selects an unknown point, but this can be repeated a large number of times in successive steps. As the spatial frequency is 0.1 m in the x and y-direction, each possible position of an



10

x1= 0 y1= 0
x2= 10 y2= 10

alpha= 0 degrees alpha= 90 degrees
tolerance= 90 degrees tolerance= 90 degrees
lag= 0.5 m lag= 1 m
max h = 5 m max h= 10 m
☑ Plot semivariogram ☐ Plot semivariogram
☐ Plot semivariogram model ☐ Plot semivariogram model
a= 10 m b= 1 mm^2 a= 10 m b= 1 mm^2
n= 0 mm^2 n= 0 mm^2
Power ○ Spherical ● Power ○ Spherical ●
Bounded linear ○ Periodic ○ Bounded linear ○ Periodic ○

Figure 4 - Second screen of Module 2, sampling and extimating an omni-directional semivariogram: a. 200 samples at random; b. 200 samples on a regular 20 x10 grid

unknown point by clicking corresponds with a known value. This means that it is possible to compare the estimated value by kriging to the true value. The user can also compare the variance of all possible points, the variance of all the selected samples (calculated within the module) and the kriging or estimation variance (i.e. of the error). And these values can then be situated against the selected value for the sill value for a spherical or bounded linear model. By repeating such estimations, one gets a good insight in the meaning and limitations of kriging: what does it really mean that kriging is called the best linear unbiased estimator, that the average error is equal to zero, but that the individual error is in most cases different from zero and that kriging results in an estimation error. Typical questions that can be addressed in this module are (Govaerts & Vervoort, 2011):

- Take 25 regular samples, and choose the position of the unknown point. Use the semivariogram parameters of the directional semivariograms of the major directions you defined using Module 2. Compare the error and the kriging variance with the results of taking an average. Which results are the best and why? Try the same for different positions of the unknown point. Is the conclusion the same for all positions of the unknown point? Why (not)? Overall, is kriging better than taking the average?
- Compare the results of an isotropic model with the results of the anisotropic model.

- Compare the results for the case where there are 100 gridded samples.

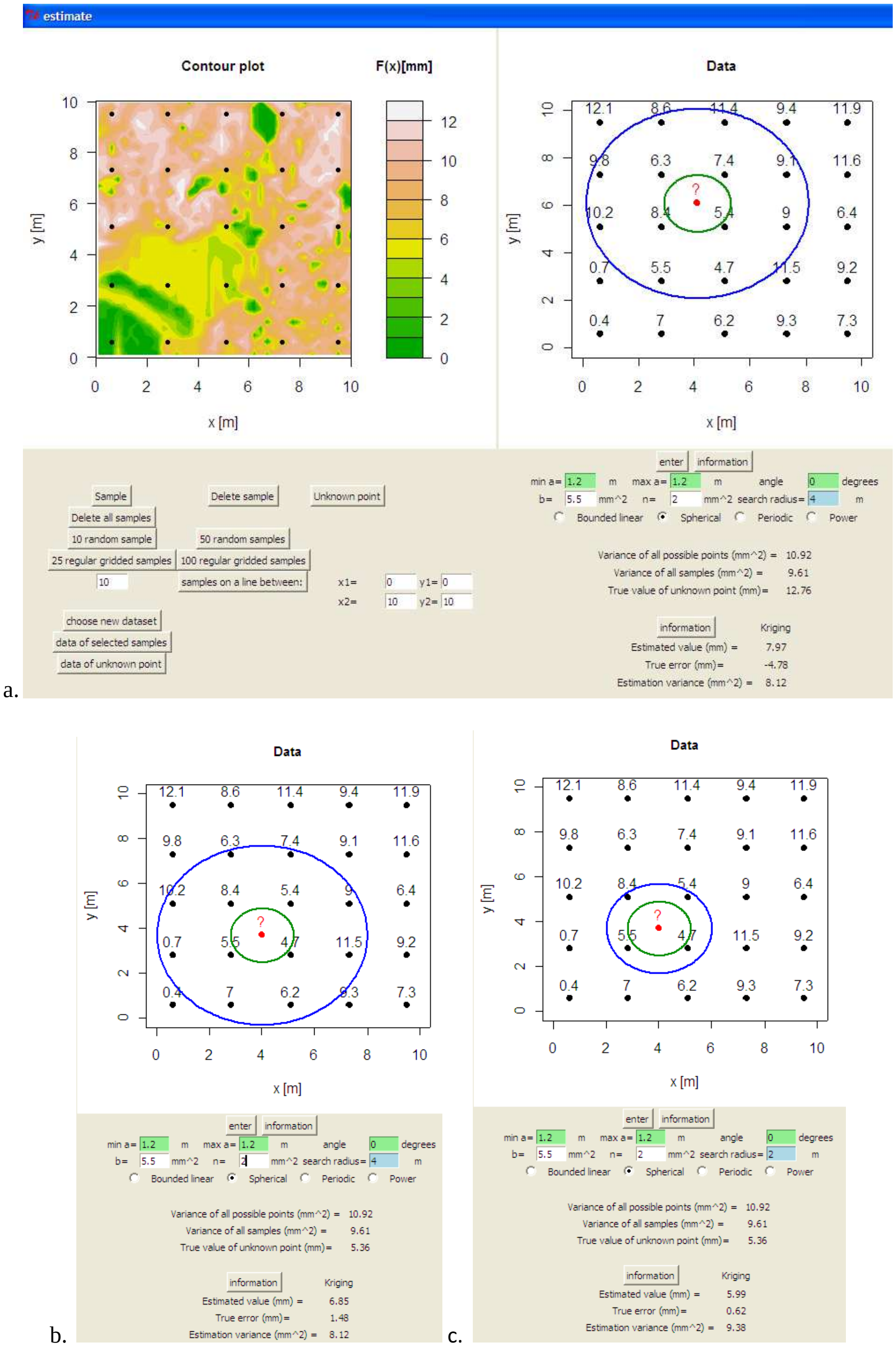
Again, the user can make an extensive sensitivity analysis, e.g. what is the effect of the various parameters of the semivariogram model (e.g. no nugget in comparison to a 50% nugget of the sill value), the effect of the position of an unknown point in comparison to a grid of samples, the effect of increasing or decreasing the search radius, and the effect of increasing the number of samples. When one of these parameters is changed, the results are automatically recalculated.

To illustrate some of these possibilities, twenty-five regular spaced points (5 x 5) are sampled in Figure 5. An estimation of the experimental semivariogram in Module 2 based on a very large number of samples resulted for small lag-distances in a spherical semivariogram with a range of 1.2 m, a nugget effect of 2.0 mm² and a sill value for the spherical part of 5.5 mm². In the three examples of Figure 5, an unknown point situated centrally between 4 samples was considered. In Figure 5.a, this unknown point is situated in a zone with relatively high values (the true value of the unknown point is 12.76 mm). In Figure 5.b and 5.c, the unknown point is situated in a zone with lower values on average (the true value of the unknown point is 5.36 mm). In Figure 5.a and 5.b, the search radius is 4 m (i.e. estimation is based on 12 samples), while in Figure 5.c, it is only 2 m (i.e. estimation based on 4 samples). One should be careful to derive too general conclusions from a limited number of cases, but the estimation variance of the first two cases is the same (8.12 mm²), which is logic as the geometry is the same, even that it is situated somewhere else. The estimation variance for the third case is larger (9.38 mm²), as less samples are considered, resulting in a poorer estimation. The estimation itself is also further away from the true value. If one would increase the search radius to 6 m, instead of 4 m, the estimation is based on 20 samples and the estimation variance is only 7.88 mm².

EXPERIENCE

Experience in lecturing basic geostatistical courses during the past 20 years to university students of different background (mining, civil, agricultural and environmental engineering and geology, geography and biology) has shown that it is not always that easy to teach the basic concepts of the successive steps in a geostatistical study (Govaerts, Vervoort, & Darius, 2011). For this, insight is needed, which can only be gained by practice. The problem starts already by making the link between the full variation of a parameter in space or over time, and a plot of the variation of a limited number of samples in 1 or 2D. Secondly, translating this variation of these samples in an experimental semivariogram as a function of the lag distance is not that easy. The whole concept of working in a new dimension, being the interdistance, is not easy. The first two modules help a new user to gain experience and insight. The translation of the calculated experimental semivariogram values into a semivariogram model is of course important, but it is not the most difficult step. However, it is important for the user to understand that there is a lot of uncertainty in the calculated values and hence in the model, especially when there are a limited number of samples. Module 2 helps to illustrate this well, as one can easily add additional samples and see the effect on the semivariogram. Finally, it is not that easy to understand the meaning of an average error equal to zero, combined with an error variance different from zero. In other words, even that kriging results in the best linear unbiased error, one could claim that mostly the most expected estimation is different from the true value. For a good interpretation of the error variance it is important to compare the latter to the variance of all samples, which is possible using the third Module 3. The latter module helps also to quantify the effect of the various parameters in an estimation procedure (sensitivity analysis of the model parameters, effect of search radius and other geometrical parameters, etc.).

Experience when using the modules in a MSc-course has shown that these modules have a significant added value to get familiarized with the basic concepts of geostatistics, but that a certain supervision when using the modules is needed. A manual has been prepared where specific problems and questions are presented, which can be solved by using the different modules (see Govaerts & Vervoort, 2011 and www.bwk.kuleuven.be/geostatistics). Experience showed that students have the tendency to read the questions, try it out immediately by clicking the various buttons, and concluding that it all looks logic. The added value of working in this way is very limited. One has to force the students to first think about the various questions and to estimate the answers, before applying the various modules. Often, the most is learned when students can further reflect on the difference between the answer they thought was right and



a.

10 samples on a line between: x1= 0 y1= 0 x2= 10 y2= 10

min a= 1.2 m max a= 1.2 m angle 0 degrees
b= 5.5 mm^2 n= 2 mm^2 search radius= 4 m
○ Bounded linear ● Spherical ○ Periodic ○ Power

Variance of all possible points (mm^2) = 10.92
Variance of all samples (mm^2) = 9.61
True value of unknown point (mm)= 12.76

Kriging
Estimated value (mm) = 7.97
True error (mm)= -4.78
Estimation variance (mm^2) = 8.12

b.

min a= 1.2 m max a= 1.2 m angle 0 degrees
b= 5.5 mm^2 n= 2 mm^2 search radius= 4 m
○ Bounded linear ● Spherical ○ Periodic ○ Power

Variance of all possible points (mm^2) = 10.92
Variance of all samples (mm^2) = 9.61
True value of unknown point (mm)= 5.36

Kriging
Estimated value (mm) = 6.85
True error (mm)= 1.48
Estimation variance (mm^2) = 8.12

c.

min a= 1.2 m max a= 1.2 m angle 0 degrees
b= 5.5 mm^2 n= 2 mm^2 search radius= 2 m
○ Bounded linear ● Spherical ○ Periodic ○ Power

Variance of all possible points (mm^2) = 10.92
Variance of all samples (mm^2) = 9.61
True value of unknown point (mm)= 5.36

Kriging
Estimated value (mm) = 5.99
True error (mm)= 0.62
Estimation variance (mm^2) = 9.38

Figure 5 - Second screen of Module 3, kriging of an unknown point: a. Unknown point situated centrally between 4 samples with a search radius of 4 m; b. Idem; c. Same point as b, but with a search radius of only 2 m.

the answer given by one of the three modules. This requires some self-discipline, i.e. to go through a difficult and sometimes confronting process. However, once the students realize that this is the right process, the modules become very useful for teaching a geostatistical course.

ACKNOWLEDGEMENTS

The financial support by the Office for Educational Policy of the KU Leuven (Project OWP2009/06) is greatly appreciated. The MSc students of the KU Leuven in the programmes Geotechnical and Mining Engineering (course H0C92A, Geostatistics) and Earth Observation (course I0D56A, Practical Geostatistics) are thanked for providing the necessary feedback and pointing out some basic programming bugs in the first version of the modules. The modules have also been used recently by the students of the Anton de Kom University of Surinam, who follow a geostatistical course in the MSc on Sustainable Management of Natural Resources; they also provided useful feedback.

REFERENCES

Govaerts, A., Vervoort, A., & Darius, P. (2011). Interactive modules for the visualization and teaching of geostatistical concepts. *Proceedings of INTED2011; International Technology, Education and Development Conference*, Valencia, Spain, 7-9 March 2011, 223-230.

Govaerts, A., & Vervoort, A. (2011). Interactive tools for the visualisation of geostatistical concepts, Manual. (KU Leuven, www.bwk.kuleuven.be/geostatistics).

Houlding, S. W. (2000). *Practical geostatistics, modeling and spatial analysis.* Springer – Verlag Berlin Heidelberg New York.

Isaaks, E. H., & Srivastava, R. M. (1989). *An introduction to applied geostatistics.* Oxford University Press, New York.

Journel, A. G., & Huijbrechts C. J. (1978). *Mining geostatistiscs.* Academic Press.

Vervoort, A., Govaerts, A., & Darius, P. (2011). Learning geostatistics through interactive modules based on R-software. *Proceedings of EDULEARN11,* Barcelona, Spain, 4-6 July 2011, 854-863.

Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists (2nd Edition)*. John Wiley & Sons.