Document extract

Title of chapter/article	Geostatistics: A Mathematical Youngster
Author(s)	Ute A. Mueller
Copyright owner	The Australian Association of Mathematics Teachers (AAMT) Inc.
Published in	Mathematics: It's Mine Proceedings of the 22nd Biennial Conference of The Australian Association of Mathematics Teachers Inc.
Year of publication	2009
Page range	17–25
ISBN/ISSN	978-1-875900-66-4

GEOSTATISTICS: A MATHEMATICAL YOUNGSTER

UTE A. MUELLER
Edith Cowan University
u.mueller@ecu.edu.au

Geostatistics is concerned with the mathematical modelling of spatial data that arise in a variety of contexts. The initial applications were related to mining and the data under consideration geological. In the last 15 years the fields of application have broadened considerably, with geostatistics now being applied in such diverse areas as mining, oil and gas exploration, ecology, health and environment. We will discuss the key methods used by means of an example, describe some of the conceptual difficulties and give a brief overview of applications.

Introduction

At 60 years of age, geostatistics is a relatively young field of mathematics. Its origins lie in the need to estimate the size of an ore deposit as accurately as possible and the first applications were to gold deposits in the Witwatersrand in South Africa (Krige, 1951). An early limitation to its application was the data size and it is only with the emergence of fast computers that the development of geostatistical techniques has really taken off. From a mathematical point of view, while we are operating in a stochastic framework, the techniques that are drawn upon come from a variety of mathematical disciplines with linear algebra and numerical analysis of particular importance (For a comprehensive overview over the methods see Chilès and Delfiner, 1999).

These days we not only estimate, but also simulate spatial distributions. Moreover, we do so using personal computers and very sophisticated software. While mining is still one of the main areas in which geostatistics is used, the breadth of applications is breath-taking. They range from pollution studies through to the modelling of fishery data to lion populations, the spread of bushfires, all the way to health-related data.

Because of the nature of the data, dealing with them is typically messy and the user cannot simply rely on a set framework of algorithms to be applied each time. There is definitely a need to get one's fingers dirty. Throughout modelling decisions need to be taken and because of the potential impact, carefully justified. These decision concern the algorithm to be used and in particular the checking of the validity of model assumptions. In this paper I use a synthetic example to give an overview of the basic methodology and then discuss some applications.

A typical problem

In Figure 1 the histogram and spatial distribution of a synthetic sample are shown. The spatial map is colour-coded with warm colours representing high values and cold colours low values. The data are positively skewed with a long tail of high values, typical for mineral distributions, such as gold. From an inspection of the spatial map we see that pairs of samples that are separated by a short distance are more likely to have similar values than pairs of samples far apart. Moreover, there do not appear to be any features favouring particular directions in space, such as banding.

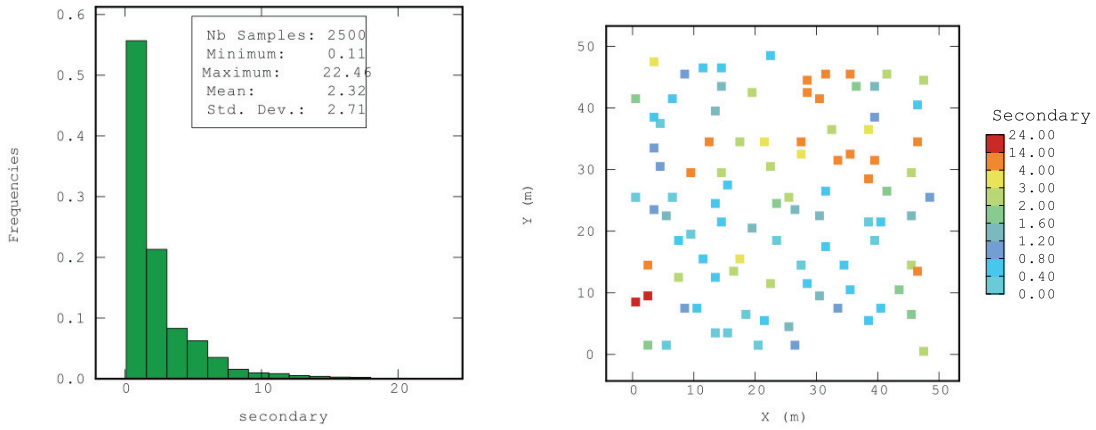

Figure 1. A gold sample: histogram (left) and spatial distribution with key statistics (right).

If we assume that the data do indeed represent a gold mineralisation, then there are several questions that would need to be answered:

- What is the exhaustive distribution of the mineral within the study region?
- What is the overall tonnage within the region and the average grade to be extracted?
- Do we have a deposit here that is worth mining, given the above results, and if so what are the optimal boundaries for the open pit that needs to be built and what mining schedule ought to be used?

The first of these questions requires “filling in the blanks.” Exhaustive drilling is evidently not the way to go because of the cost involved. An algorithm is needed to calculate estimates for the un-sampled locations. One requirement on the estimator is unbiasedness: the mean of the estimates is equal to the population mean. There are many different ways in which this filling in of the blanks can be done, including allocating the grade of the nearest neighbour, fitting a polynomial in the space coordinates to obtain an estimate and calculating a weighted average within a search window. The simplest method is to allocate the mean grade within the search window, giving equal weight to all samples it contains or else to take account of separation and possibly value. The latter is the approach taken in *ordinary kriging*, the most prevalent geostatistical estimation algorithm. This algorithm is named after one of the pioneers of geostatistics, Danie Krige who was one of the first to use windowed multilinear regression to calculate estimates.

The results for several “filling ins” are shown in Figure 2. Each one of these methods honours the data, in that the values at the sample locations are the actual sample values.

However, not all of the maps appear equally realistic. The spatial distributions obtained from nearest neighbour interpolation or moving window averaging appear patchy, while that from polynomial interpolation appears too smooth when compared with reality. Moreover, of the above approaches to the estimation problem, kriging is the only method that allows one to obtain a measure of the uncertainty in the form of an estimation error. Scrutiny of the histograms of the estimates and the exhaustive data (called "reality"[5]) in Figure 3 shows that only moving windows averages and ordinary kriging estimates have histograms that are close to the sample histogram. Moreover, polynomial interpolation results in negative estimates, which are not realistic in the given context.

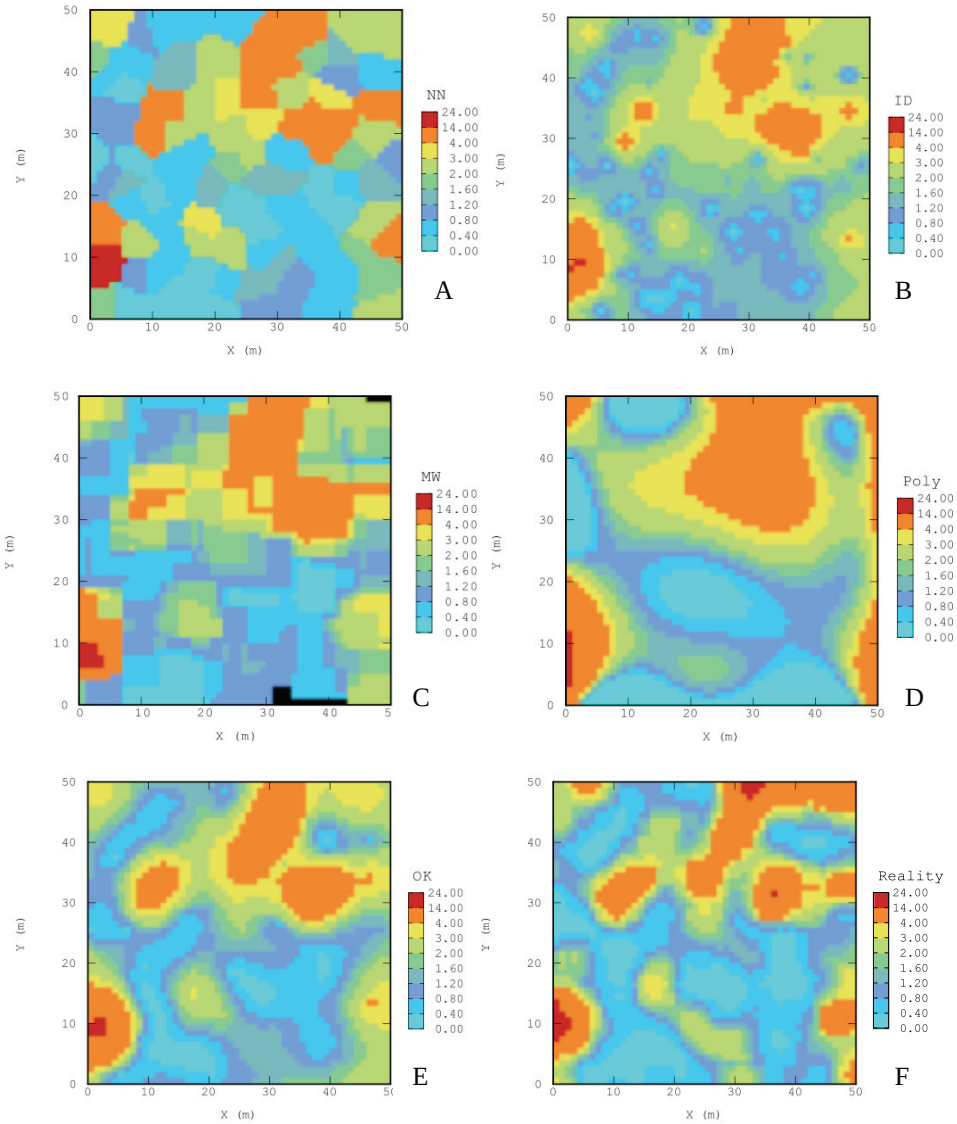

Figure 2. Estimated distributions resulting from different estimation methods: nearest neighbour estimation (A), inverse distance interpolation (B), moving window averages (C), polynomial of degree 6 in X and Y (D), ordinary kriging (E) and reality (F).

[5] While we know 'reality' in this case, of course this is not true in general.

If we accept that ordinary kriging provides a reasonable estimate for the spatial distribution of the gold mineralisation, then we can go ahead and use our estimates to tackle the second question and determine the average grade and calculate grade tonnage curves, the information required to decide if it is worthwhile to further develop the resource and maybe open a mine. Ultimately these decisions depend upon financial considerations, such as the type of gold mineralisation which in turn impacts on the milling process and the current gold price, as well as the possibility of forward sales.

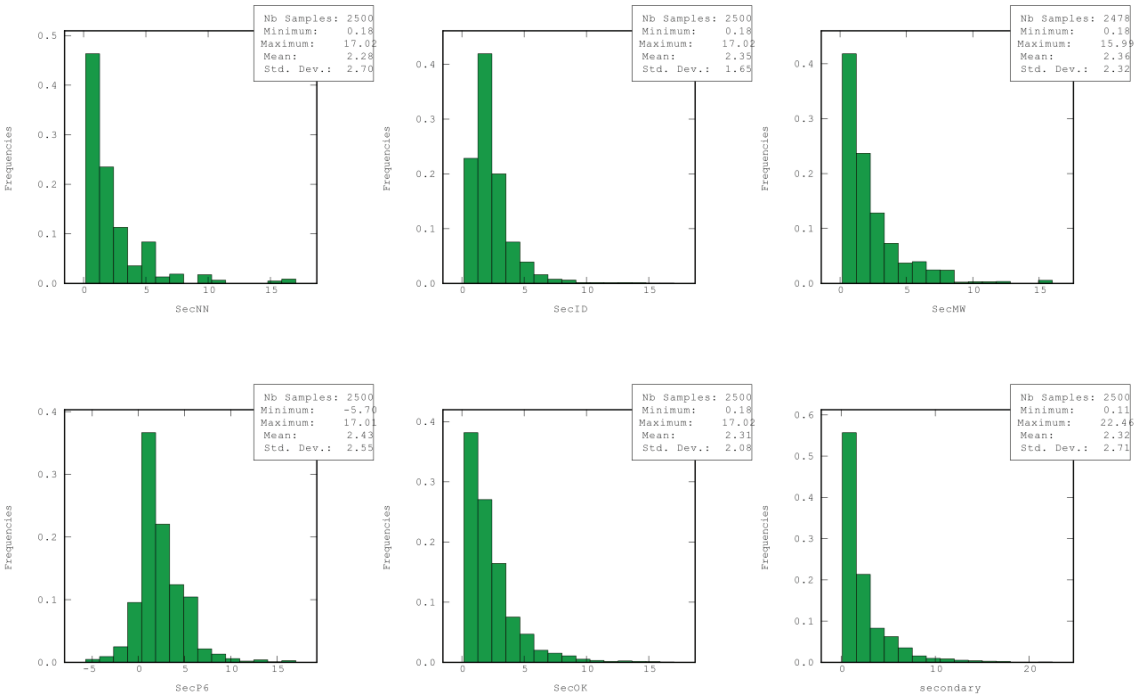

Figure 3. Histograms with key statistics for the estimates together with reality: nearest neighbour estimation (top left), inverse distance interpolation (top centre), moving window averages (top right), polynomial of degree 6 in X and Y (bottom left), ordinary kriging (bottom centre) and reality (bottom right).

To fully answer the question in relation to mine planning, potentially the last stage of the exercise, we need to use simulation. Based on strong model assumptions we generate equiprobable spatial distributions of the gold variable (see Figure 4) consistent with the sample data to assess the risk of choosing a particular pit design.

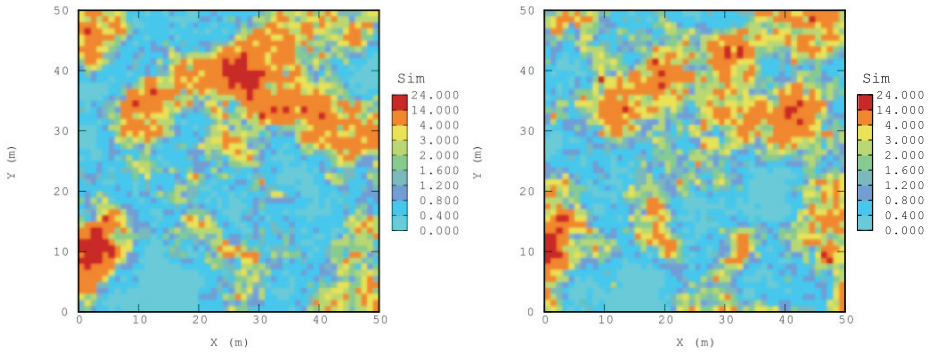

Figure 4. Two simulated spatial distributions of gold based on the sample in Figure 1.

As with the estimation, the simulations are consistent with the sample data, but each distribution represents a different possible reality, which in expected value approximates the map of ordinary kriging estimates. The simulations are used to generate histograms of the distributions at grid locations (see Figure 6) and maps that clearly indicate regions of high and low values as well as probability maps (see Figure 5) that allow one to visualise the probability of exceeding a threshold of interest, such as the minimum gold grade that will make the deposit economic.

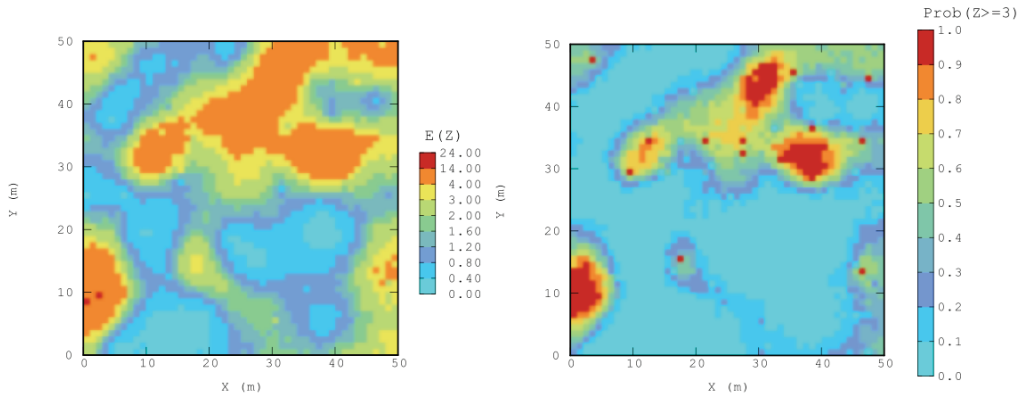

Figure 5. Average value of the simulations computed location by location (left) and spatial map of the probability of exceeding a grade of 3 ppm.

The spatial map of the mean grade calculated location by location shown in Figure 5 clearly highlights a region of high grades in the north eastern part of the region and if a cut-off grade of 3 ppm was applied, then only the north-east and the south-west would contain regions where the probability of exceeding this grade are high. From the map of the average values it is already apparent that the grade frequency distributions will differ from location to location. In fact, while the shape of the overall grade distribution of an individual simulation is not dissimilar from that of the exhaustive data (see Figure 4), the shapes of the distributions of the simulated grades at individual locations do not resemble the overall distribution of grades (see Figure 6). The grade distributions for locations (12.5,5.5) in the south-west and (25.5,36.5) in the north are both positively skewed, but they have very different ranges and kurtosis.

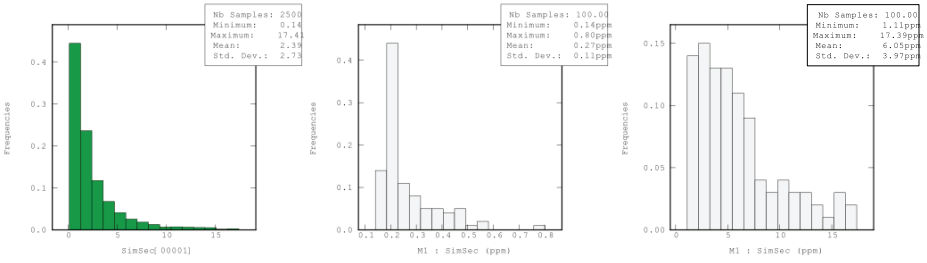

Figure 6. Histogram of simulation #1(left) and histograms for locations (12.5,5.5)(centre) and (25.5,36.5) (right).

The modelling approach

In the previous section I have given an overview over the typical workflow of a geostatistical study, moving from an exploratory data analysis to estimation and finally simulation taking into consideration the spatial continuity. I will now briefly consider the mathematical framework. The data with which geostatistics is concerned have the property that once sampled at the selected sites, there is no possibility of replication: once, say, a drill core has been pulled at a site, the action cannot be repeated. In essence, the process we are considering is deterministic. However, it is usually so complex that the construction of a deterministic model is not practical; just recall the poor job done by polynomial interpolation in Figure 1. It is for this reason that a stochastic framework is adopted.

This framework is known as the *Random Function Model*. We assume that at each location $\mathbf{u}$ in the study region there is a random variable $Z(\mathbf{u})$ and that the observed value $z(\mathbf{u}_\alpha)$ at a sample location $\mathbf{u}_\alpha$ is nothing but a realisation drawn from the random variable $Z(\mathbf{u}_\alpha)$ at the location. Because of this construct meaningful multivariate datasets may be constructed that enable the user to develop an insight into the spatial features of the attribute under study. Specifically the model allows us to calculate covariances that are functions of the separation distance. They are necessary for computing estimates and simulated values at un-sampled locations. The estimate at an un-sampled location is a weighted linear combination of the sample grades in the vicinity of the location. Determining the weights and hence the estimate ultimately comes down to solving a linear system, a procedure akin to standard linear regression.

A rigorous mathematical formulation of the estimation procedure was first given by Matheron who also coined the term *kriging*. The first formal courses in the subject were held at the Ecole des Mines de Paris in France and graduates from that school where responsible for the proliferation and dissemination of the techniques developed there worldwide. In this paper, rather than dwell on the technicalities of geostatistical techniques we will have a look at some of the applications.

Applications of geostatistics

This section includes an overview of some applications.

Mining

Mining applications are a stalwart of geostatistics. They cover all types of mining resources from metals to diamonds to coal and also petroleum. Here in Western Australia, geostatistical estimation is used regularly in the Pilbara iron ore mines for planning the mining, and in the WA goldfields for scheduling the extraction of gold. In the case of iron, it is not enough to analyse and model the distribution of iron, but in addition alumina and silica need to be modelled as their distribution impacts on the quality of the iron ore. Variables such as the grade of gold and the iron content are continuous variables, but diamonds, for which the use of geostatistics is well established, are discrete objects and so a model of the distribution of sizes is required. Their size distribution is highly positively skewed and as the interest is in large diamonds, extreme value modelling needs to be undertaken (Lantuéjoul 2008).

In the area of estimating the size of an oil reservoir hard data are scarce, as the drilling of oil wells is expensive. It is often the case that only a few oil wells are available, so denser secondary information such as seismic data are used to improve estimation.

Natural resources

One branch of natural resources modelling is concerned with fisheries. Since the early 1990s studies were conducted on survey data from the North Sea. The objective was an abundance assessment of commercially interesting fish species, such as herring and hake. An example from Western Australia concerns the catch and catch rate distribution of prawns and scallops in Shark Bay. One part of this study dealt with the ability of the annual scallop survey to adequately predict the subsequent scallop catch and an assessment as to whether or not trawling for prawns prior to the start of the scallop fishing season disturbed the settlement of the scallops. Given that scallops move very little, a distortion or shift in the distribution between the time of the survey and the distribution based on catch would have indicated an adverse effect of pre-season trawling. Our findings showed no clear evidence for a disturbance of scallop settlement (Mueller et al., 2008).

Environment

Applications in this field cover soil contamination, air quality in cities, water transport and the abundance of wildlife. The spatial distribution of whales within the Mediterranean whale sanctuary located in the waters between Spain and Italy has been studied extensively (Monestiez et al. (2006)). The raw data are the sightings of whales (by observers) over a period of ten years. The data are count data and a specifically tailored kriging algorithm was used to construct a map of the spatial distribution of whales. A complication in the modelling of wildlife data is the use of enthusiastic volunteers for data collection. They tend to frequent areas where observation of the animal of interest is more likely. An attempt to deal with this obstacle was presented in a paper on the spatial modelling of bird distributions in Croatia (Hengl et al., 2008).

Reforestation is another area of environmental application. The variable of interest is the number of plants surviving. To assess the survival rate a sampling design is necessary that locally allows the prediction of the rate with a prescribed maximum error. To be cost-effective, the sampling design needs to contain as few sites as possible and still be sufficiently accurate. In a case study from Chile (Emery et al., 2008) an initial distribution of sampling sites was available and sites for in-fill samples needed to be found to provide reliable estimates of the survival rate. In this case study two approaches to determine such an in-fill pattern for a plantation in Chile are discussed. The interest in the design of an in-fill pattern arose because of financial incentives by the Chilean government for the establishment of new plantations, but in order to qualify for the payment, there is a requirement of a 75% plant survival rate.

Health

Applications to health geography are fairly recent and are often a mixture of Geographic Information Systems and geostatistics. One of the human diseases of interest is cancer and there have been several studies concerned with the analysis of the spatial distribution of the incidence of various cancers (e.g.,, Goovarets, 2005). The objective

of such studies is to obtain good estimates of the incidence risk and factors influencing the level of risk. The availability of reliable maps of incidence risk is important for public health campaigns and eradication programs. The applications are not restricted to cancer or human diseases. Contagious diseases like cholera and dysentery have also been investigated, for example for the Matlab region of Bangladesh (Ali et al., 2006). The mapping revealed a patchier cholera risk map than dysentery risk map, and also identified higher risk in the more urban areas for both diseases. An example of an animal disease studied in this way is foot and mouth disease (Perez et al., 2006). One of the common features of these studies is the need to use imperfect data. There are usually reporting inaccuracies, and in the case of human disease the need to use census data that are only recorded in selected years, contributes to the imperfection.

Concluding remark: Geostatistics in the secondary classroom

Geostatistics is a fascinating discipline with a wide variety of applications. Its interdisciplinarity and the nature of its applications make it a good candidate for highlighting the importance of mathematics in many different disciplines and its relevance in the modern world. Consideration of spatial data can provide students with interesting applications of some of the statistical techniques they learn in high school. While the construction of a semivariogram and its evaluation for kriging or simulation may be too difficult, some of the basic ideas, such as the construction of an abundance map, are accessible at the upper secondary level and would make interesting extension exercises. The construction of a spatial sample map requires the use of a meaningful colour scale so that the user can glean relevant information from it readily. This relies on calculating descriptive statistics and possibly deciles of the sample distribution. Going on from there, an exploration of characteristics of the sample is possible and the construction of estimates at unsampled locations using either moving windows or another weighted linear combination of the data, allowing an exploration of the impact of the sampling support and of the importance of a good estimator.

References

Ali, M., et al., (2006) Application of poisson kriging to the mapping of cholera and dysentery incidence in an endemic area of Bangladesh. *International Journal of Health Geographics*, *5*, 45.

Chiles, J. P. & Delfiner, P. (1999). *Geostatistics: Modelling Spatial Uncertainty*. New York; John Wiley & Sons.

Emery, X., Hernández, J., Corvalán, P. & Montaner, D. (2008). Developing a cost-effective sampling design for forest inventory, In *Geostats 2008; Proceedings of the eighth Geostatiscal Congress.* Santiago, Chile.

Goovaerts, P. (2005). Geostatistical analysis of disease data: Estimation of cancer mortality risk from empirical frequencies using Poisson kriging. *International Journal of Health Geographics, 4:31*

Hengl, T., Radovic, A. Van Loon, E. E. (2008). Mapping bird nests densities using kernel smoothing and regression-kriging. *Proceedings of the Seventh European Conference on Environmental Applications of Geostatistics*, 137.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa,* December, 119–139

Matheron, G. (1962–63). Traité de géostatistique appliquée, Tome I, Tome II: Le krigeage I: *Mémoires du Bureau de Recherches Géologiques et Minières*, *14*(1962), Editions Technip, Paris; II: *Mémoires du Bureau de Recherches Géologiques et Minières*, No. *24*(1963), Editions B.R.G.M, Paris.

Monestiez, P., Dubroca, L. & Bonin, E. (2006). Geostatistical modelling of spatial distribution of Balaenoptera physalus in the northwestern Mediterranean Sea from sparse count data and heterogeneous observation efforts. *Ecological Modelling 193*, 615–628.

Mueller, U., Dickson, J., Kangas, M. & Caputi, N. (2008). *Geostatistical Modeling of the Scallop Density Distribution in Shark Bay, Western Australia, from Survey Data*. Edith Cowan University.

Perez, A., Thurmond, M. C. & Carpenter, T. E. (2006). Spatial distribution of foot and mouth disease in Pakistan using imperfect data. *Preventive Veterinary Medicine*. *76*, 280–289.