EGS 2405: GEOSTATISTICS

GIS CLASS 2023

Jomo Kenyatta University of Agriculture and Technology

Lecture notes 1

References:

Isaaks, E. H., & Srivastava, R. M. (1989). Applied geostatistics (Vol. 561). New York: Oxford University Press.

Oliver, M. A., & Webster, R. (2015). Basic steps in geostatistics: the variogram and kriging (No. 11599). Cham,

Switzerland: Springer International Publishing.

McKillup, S., & Dyar, M. D. (2010). Geostatistics explained: an introductory guide for earth scientists. Cambridge University Press.

A Practical Guide to Geostatistical Mapping

Other Resources:

Stanford Geostatistical Modeling Software (SGeMS)

1 Introduction

1.1 What is Geostatistics?

The term *geostatistique* was coined by the French engineer Georges Matheron. He was inspired by the clear meaning and success of the older terms, geochemistry and geophysics, in which the prefix geo- was added to the name of some classical body of knowledge to denote an application of such knowledge to the modeling and understanding of processes of interest in earth sciences and technology (Matheron, 1962; Journel and Huijbregts, 1978). Geostatistics is, therefore, an invaluable tool that can be used to characterize spatial or temporal phenomena.

Geostatistics originated from the mining and petroleum industries, starting with the work by Daniel Krige in the 1950's and was further developed by Georges Matheron in the 1960's. In both industries, geostatistics is successfully applied to solve cases where decisions concerning expensive operations are based on interpretations from sparse data located in space. Geostatistics has since been extended to many other fields in or related to the earth sciences, e.g., hydrogeology, hydrology, meteorology, oceanography, geochemistry, geography, soil sciences, forestry, landscape ecology, and agriculture. In this class, both the fundamental development of geostatistics and the simple, practical applications in the earth sciences will be explored, with exercises illustrating the applications of geostatistics in the characterization, estimation, prediction, and modeling of spatial datasets.

Definitions of Geostatistics

- In its broadest sense, geostatistics can be defined as the branch of statistical sciences that studies spatial/temporal phenomena and capitalizes on spatial relationships to model possible values of variable(s) at unobserved, unsampled locations (Caers, 2005).
- Geostatistics is a set of models and methods that are designed to study variables which are distributed in space (or possibly space-time). Such variables possess both a structured and a random aspect and cannot be simply described by a regular function of the coordinates.
- A class of statistics used to analyze and predict the values associated with spatial or spatiotemporal phenomena. Geostatistics provides a means of exploring spatial data and generating continuous surfaces from selected sampled data points
- Geostatistics can be regarded as a collection of numerical techniques that deal with the characterization of spatial attributes, employing primarily random models in a manner similar to the way in which time series analysis characterizes temporal data (Olea, 1999).
- Geostatistics offers a way of describing the spatial continuity of natural phenomena and provides adaptations of classical regression techniques to take advantage of this continuity. (Isaaks and Srivastava, 1989).

For a given sample of measurements {z1,z2,...,zn} at locations {x1,x2,...,xn} the main goal is to estimate the value z at some new point x.

1.2 Why Geostatistics

Classic statistics is generally devoted to the analysis and interpretation of uncertainties caused by limited sampling of a property under study. Geostatistics, however deviates from classic statistics in that Geostatistics is not tied to a population distribution model that assumes, for example, all samples of a population are normally distributed and independent from one another. Most of the earth science data (e.g., rock properties, contaminant concentrations) often do not satisfy these assumptions as they can be highly skewed and/or possess spatial correlation (i.e., data values from locations that are closer together

tend to be more similar than data values from locations that are further apart). To most geologists, the fact that closely spaced samples tend to be similar is not surprising since such samples have been influenced by similar physical and chemical depositional/transport processes.
Compared to the classic statistics which examine the statistical distribution of a set of sampled data, geostatistics incorporates both the statistical distribution of the sample data and the spatial correlation among the sample data. Because of this difference, many earth science problems are more effectively addressed using geostatistical methods. As stated by Marc Cromer (in Geostatistics for environmental and geotechnical applications, 1996, ASTM International, edited by Rouhani et al.):
Geostatistics is useful in Geosciences because of the following reasons:

1. The environment is continuous, but we can afford to measure properties at only a finite number of places. Thus, the best we can do is to estimate or predict in a spatial sense, see Figure 1.1.
2. It models variability better, i.e.
    - Controllable degree of spatial variability.
    - Estimates are more reliable
3. It provides a framework to integrate data for example, direct measurements (hard data) and secondary variables (soft data). The data can also represent different supports.
4. Geostatistical methodologies are repeatable i.e. can be audited.

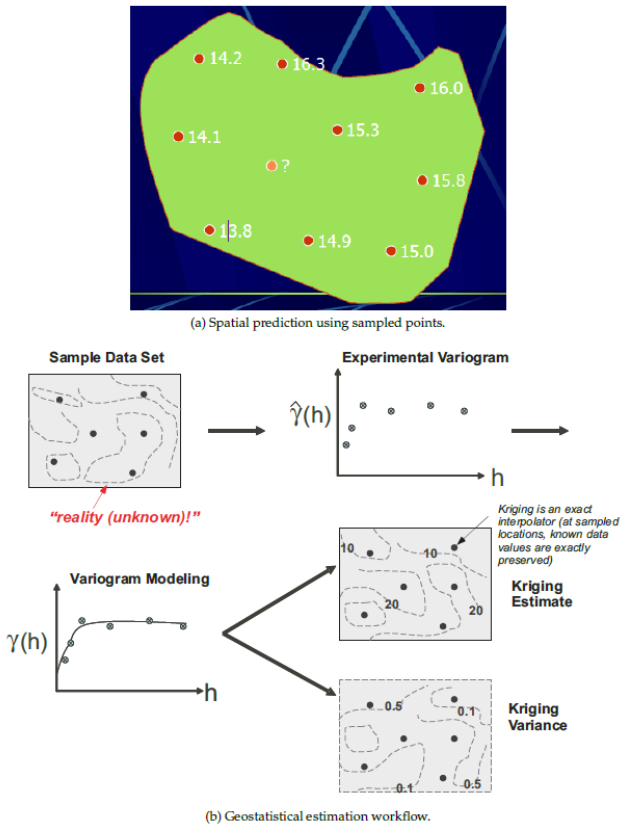


Figure 1.1.1 Illustration of application of Geostatistics for prediction.

Classical statistics are based on random sampling, linear sum of data, all of whom carry the same weight. If there is spatial correlations, then by stratifying, we can estimate more precisely or sample more effectively. If the strata are of different sizes, then we can vary the weights attributable to their data in proportion. Geostatistics rely on spatial models, while classical statistics don't. Classical statistics are based on sampling design, which implies unbiasedness and provides estimates of error if the choice of sampling design is suitable. It requires no assumptions about the nature of the variable

itself. Geostatistics assumes that the variable is random and the outcome of one or more random processes. The models on which predictions are based are of these random processes.
In summary, **geostatistical mapping** can be defined as the analytical production of maps by using field observations, auxiliary information and a computer program that calculates values at locations of interest (a study area). It typically comprises the following five steps:

(1.) Design the sampling and data processing,
(2.) Collect field data and do laboratory analysis,
(3.) Analyse the point's data and estimate the model,
(4.) Implement the model and evaluate its performance,
(5.) Produce and distribute the output geoformation

1.3 Basic Concepts

Any measurement we take in Earth and Environmental Sciences has a spatiotemporal reference. A spatiotemporal reference is determined by (at least) four parameters:
(1) Geographic location (longitude and latitude or projected X, Y coordinates);
(2) Height above the ground surface (elevation, Z);
(3) Time of measurement (year, month, day, hour, minute etc.);
(4) Spatiotemporal support (size of the blocks of material associated with measurements; time interval of measurement);

Example of Fish density

(i) The bottom depth at 2D point, (ii) fish density in 2D, or (iii) concentration in 3D (number or weight of fish per unit 2D area or 3D volume). Occasionally, x can represent a point in 1D, e.g. the transect biomass obtained by summing fish densities along transects with a given direction. A time aspect can be introduced, for example the concentration of fish over a period of 10 years. This yields the fourth dimension (4D)
Then, the abundance Q over a region V is the sum of the fish density over this region:

$$Q = \int_V z(x)\mathrm{d}x \qquad \text{Equation 1.3.1}$$

and the mean fish density over this region is:

$$z(V) = \frac{Q}{V} = \frac{1}{V}\int_V z(x)\mathrm{d}x \qquad \text{Equation 1.3.2}$$

Such variables are usually not known everywhere in space. Data may be available at isolated data points (e.g. sampling stations for trawl surveys), along transects (e.g. acoustic or video surveys), or, for example, over a gridded map (satellite data).
New variables obtained by transforming the original ones can also be considered. For example, the indicator of the presence of fish: regionalized variable equal to 0 where the fish density is 0 and equal to 1 otherwise. Or the logarithm of a non-zero concentration to better describe a histogram and reduce the influence of the largest values (care should be taken, however, when using such non-linear transformations: back-transforming statistics are not straightforward, for example, the antilog of the mean of the logarithm is not the mean of the variable).

Measurements with assigned geographical coordinates can be analyzed and visualized using a set of specialized techniques. A general name for a group of sciences that provide methodological solutions for the analysis of spatially (and temporally) referenced measurements is Spatiotemporal Data Analysis

(STDA) Figure 1.2. Image processing techniques are used to analyze remotely sensed data; point pattern analysis is used to analyze discrete point and/or line objects; geostatistics is used to analyze continuous spatial features (fields); geomorphometry is a field of science specialized in the quantitative analysis of topography. We can roughly say that STDA is a combination of two major sciences: geoinformation science and spatiotemporal statistics, or in mathematical terms: STDA =
GIS + statistics.

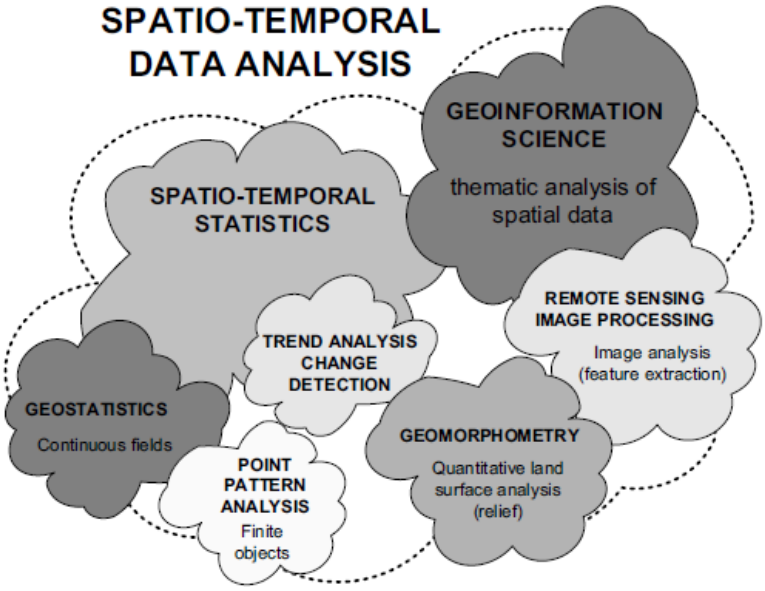


Figure 1.1.2 Spatio-temporal Data Analysis is a group of research fields and sub-fields.

The temporal dimension is starting to play an increasingly important role, so that many principles of spatial statistics (hence geostatistics also) will need to be adjusted. Because geostatistics evolved in the mining industry, for a long time it meant statistics applied to geology. Since then, geostatistical techniques have successfully found application in numerous fields ranging from soil mapping, meteorology, ecology, oceanography, geochemistry, epidemiology, human geography, geomorphometry and similar. Contemporary geostatistics can, therefore best be defined as a branch of statistics that specializes in the analysis and interpretation of any spatially (and temporally) referenced data, but with a focus on inherently continuous features (spatial fields). Typical questions of interest to a geostatistician are:

- How does a variable vary in space-time?
- What controls its variation in space-time?
- Where to locate samples to describe its spatial variability?
- How many samples are needed to represent its spatial variability?
- What is a value of a variable at some new location/time?
- What is the uncertainty of the estimated values?

In the most pragmatic terms, geostatistics is an analytical tool for statistical analysis of sampled field data that has three scientific objectives:
(1) **Model estimation**, i.e. inference about the model parameters;
(2) **Prediction**, i.e. inference about the unobserved values of the target variable;
(3) **Hypothesis testing** (Bolstad, 2008; Diggle and Ribeiro Jr, 2007).
Model estimation is the basic analysis step, after which one can focus on prediction and/or hypothesis testing. In most cases, all three objectives are interconnected and depend on each other i.e. they are not mutually exclusive. The difference between hypothesis testing and prediction is that, in the case of

hypothesis testing, we typically look for the most reliable statistical technique that provides both a good estimate of the model, and a sound estimate of the associated uncertainty. Spatial prediction, on the other hand, is usually computationally intensive, so that sometimes, for pragmatic reasons, naïve (simplistic) approaches are more frequently used to generate outputs; uncertainty associated with spatial predictions is often ignored or overlooked. In other words, in the case of hypothesis testing we are often more interested in the uncertainty associated with some decision or claim while in the case of spatial prediction we are more interested in generating maps (within some feasible time-frame) i.e. exploring spatio-temporal patterns in data. Spatial prediction or spatial interpolation aims at predicting values of the target variable over the whole area of interest, which typically results in images or maps. Note that there is a small difference between the two because prediction can imply both interpolation and extrapolation.

As stated by Marc Cromer (in Geostatistics for environmental and geotechnical applications, 1996, ASTM International, edited by Rouhani et al.): Geostatistical methods provide the tools to capture, through rigorous examination, the descriptive information on a phenomenon from sparse, often biased, and often-expensive sample data. The continued examination and quantitative rigor of the procedure provide a means for integrating qualitative and quantitative understanding by allowing the data to "speak for themselves". In effect, the process produces the most plausible interpretation by continued examination of the data in response to conflicting interpretations. ... The application of geostatistics to environmental problems (e.g., groundwater contaminant clean up) has also proven to be a powerful integration tool, allowing the coordination of activities from field data acquisition to design analysis. For example, data collection is often incomplete, resulting in uncertainties in understanding the problem and increasing the risk of regulatory failure. While these uncertainties can often be reduced with additional sampling, the benefits must be balanced with increasing costs. Thus, geostatistics offers a means to quantify uncertainty while leveraging existing data to support sampling optimization.

Justification of Geostatistics via an Example:

Imagine a situation where a farmer has asked you to survey the soil of his farm. In particular, the farmer wants you to determine the phosphorus content; but they will not be satisfied with the mean value for each field as would have been the case a few years ago. The farmer now wants more details so that they can add fertilizer only where the soil is deficient, not everywhere. The survey involves taking numerous samples of soil, which you must transport to the laboratory for analysis. You dry the samples, crush them, sieve them, extract the phosphorus with some reagent, and finally measure it in the extracts. The entire process is both time-consuming and costly. Nevertheless, at the end, you have data from all the points from which you took the soil—just what the farmer wants, you might think! The farmer's disappointment is evident, however. 'Oh', he says, 'this information is for a set of points, but I have to farm continuous tracts of land. I really want to know how much phosphorus the soil contains everywhere. I realize that that is impossible; nevertheless, I should really like some information at places between your sampling points. What can you tell me about those, and how do your small cores of soil relate to the blocks of land over which my machinery can spread fertilizer, that is, in bands e.g. 24 m wide?' This raises further issues that you must now consider. Can you say what values to expect at intervening places between the sample points and over blocks the width of the farmer's fertilizer spreader? Moreover, how densely should you sample for such information to be reliable? At all times, you must consider the balance between the cost of providing the information and the financial gains that will accrue to the farmer by differential fertilizing. In the wider context, there may be an additional gain if you can help to avoid over-fertilizing and thereby protect the environment from pollution by excess phosphorus. Your task, as a surveyor/ geoscientist, is to be able to use sparse affordable data to estimate or predict, the average values of phosphorus in the soil over blocks of land. Can you provide the farmer with spatially referenced values that he can use in an automated fertilizer spreader? Precision farming technology can position machines accurately to 2m in the field, measure and record the yields of crops continuously at harvest, can modulate the amount of fertilizer added to match demand etc., but providing the information on the nutrient status of the soil at an affordable price remains a major challenge in

modern precision farming (Lake et al., 1997). So, how can you achieve this? The answer is to use geostatistics and we can change the context from precision farming to soil salinity, pollution of soil and water bodies by heavy metals, arsenic in ground water, rainfall, barometric pressure, to mention just a few of the many variables and materials that have been and are of interest to geoscientists. What is common to them all is that the environment is continuous, but in general, we can afford to measure properties at only a finite number of places. Elsewhere the best we can do is to estimate, or predict, in a spatial sense. This is the principal reason for geostatistics -- it enables us to do so without bias and with minimum error. It allows us to deal with properties that vary in ways that are far from systematic i.e. random and at all spatial scales. As mentioned earlier, estimation and prediction yields uncertainties and are subject to error. However, it is important to note that, geostatistics can never provide complete information but, given the data, it can enable you to estimate the probabilities that true values exceed specified thresholds. Taking the example of precision farming, this means that you can assess the farmer's risks of losing yield by doing nothing where the true values are less than the threshold or of wasting money by fertilizing where they exceed it. In some situations, the conditional probabilities of exceeding thresholds are as important as the estimates themselves because there are matters of law involved. Examples include limits on the arsenic content of drinking water (what is the probability that a limit is exceeded at an unsampled well?) and heavy metals in soil (what is the probability that there is more cadmium in the soil than the statutory maximum?)

The challenges of and opportunities for geostatistics can, therefore be summarized as follows:

- We have the tools that allow the integration of statistical methods and spatiotemporal data i.e.
- GIS + statistics integration
- There is more and more data: e.g. MODIS (global coverage, 250 m, daily, 36 bands), Meteorological images (e.g. Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS), 5 km, Daily) and Spinning Enhanced Visible and InfraRed Imager (SEVIRI) 1 km, every 15 minutes., 12 bands), Terrain data (e.g. SRTM DEM (90 m and 30 m))
- We can automate data analysis using programming e.g. R and Python, and cloud computing platforms e.g. GEE (Get results sooner, with more accuracy)
- With the increasing diversity of available data, computation, and modelling techniques, and the requirement to integrate or incorporate them in modeling and simulation, the issue of scaling different data arises. The proposed solutions are particular to the available datasets and are not intended for use as general recipes. The hallmark of a good geostatistical study is the customization of the approach to the problem at hand. This requires a thorough understanding of the data and the underlying logic of geostatistics techniques.

In short, geoscientists collect (spatio-temporal) data and need tools that can help them answer field-specific questions i.e. that can help produce outputs of interest maps, predictions, and statistical measures). Since geoscientists are often forced to work with samples, they need to know how to sample, and they need to know how confident they can be about making generalizations from these samples.

1.3.1 Sampling, Data Collection and Sample Design

All items in any field of inquiry constitute a ' universe' or 'population'. From a practical point of view, it is not possible to collect all items in a population for analysis, unless the population itself is a very small one. Therefore, we resort to forming what is known as a 'sample'. Mathematically, if the population size is N and a part of it, say n ($n<N$), is selected according to some rule for studying some

characteristic(s) of the population, this set of n units/elements is known as a sample. The individual items/elements in a sample should as much as possible, be representative of the population. The selection process is called a sampling technique/sampling procedure.

Geostatistical analysis is all about modelling spatially indexed random fields or variables. It differs from point or event analyses (recall GIS spatial analysis) since the random fields are considered to come from a continuous smooth surface. The challenge is that, for a continuous surface, there is no way of measuring it at all points. Instead, we have a finite number of sample points, and from this sample points we try to infer the model or process that generates the surface. A process can be formally defined as set of random fields/ variables (Z) indexed by a spatial index (s), which is a subset of a larger continuum (D), and is continuous in 2D or 3D.

$$\{\boldsymbol{Z(s)}\}: \boldsymbol{s} \in \boldsymbol{D}$$

We therefore have a finite set (N) of spatial locations i.e. $\{s_1, s_2 \dots \dots \dots . . s_n\}$ and at these sample locations we measure random variables i.e. $Zs_1, Zs_2 \dots \dots \dots . . Zs_n$, and using these random variables, we attempt to infer what is driving the variability of the magnitude of Z in space.

A researcher must prepare a sample design for his study and its size i.e. the number of sample points/elements. Data collection and sampling is a key aspect of any geostatistical study and includes an orderly collection of various types of data e.g. choosing locations for collecting soil samples in a given region, selecting locations for taking gravity readings etc., for exploration. The sampling method could be random sampling, stratified sampling, systematic sampling, cluster sampling etc., depending on the need/ application. In the process of developing a sampling design, some essential considerations include:

a) Type of universe: The first step in developing any sample design is to define the set of objects - typically called the "universe" or "population" to be studied. The population can be finite or infinite. In a finite population, the number of items/ elements is limited, while in an infinite population, the number of items is infinite. Examples of finite population include the number of outcrops in a geological terrain, the number of boreholes in a region, the number of persons in a city and so on. Examples of infinite population include the number of stars in the sky; and the number of ore samples that can be taken from a gold-bearing lode etc.

b) Sampling unit: A sampling unit may be a geographical one such as a state, district, village etc. From a geological point of view, it could be a geological zone, a rock specimen etc.

c) Source list: It is also known as the 'sampling frame' from which a sample is to be drawn.

Such a list should be comprehensive and reliable. It is important for the source list to be representative of the population.

d) Type of sampling: A researcher must decide on the type of sample to be collected and the techniques to be employed.

e) Sample size: This refers to the number of items to be selected from the universe/population to constitute a sample. An optimum sample size is one that fulfils the requirements of efficiency, representation, reliability, and flexibility. While deciding on the size of the sample, a researcher must determine the desired precision as well as an acceptable confidence level of the estimate. The size of the sample variance needs to be considered in relation to population variance. If the variance of the sample is large, then a larger sample size may be needed. The size of the population and the parameters of interest in a research study must also be kept in view while deciding on the size of the sample.

f) Parameters of interest: A researcher must address the question of specific population parameters that are of interest. For example, we may be interested in estimating the population mean of mine samples when the distribution is lognormal or some other characteristics of the population. In addition, a researcher must select a sample design which gives lesser sampling error for a given sample size and cost.

When selecting a procedure for drawing a sample, a researcher must ensure that it causes relatively small sampling error for a given sample size, it minimizes the costs associated with data collection, and helps in controlling systematic bias. A systematic bias is the result of one or more of the following factors:
(i) Inappropriate sampling frame. If the sampling frame is inappropriate, a biased representation of the population and hence a systematic bias occurs. For example, in Figure 1.3(a) below, an example of "under coverage" is depicted, wherein bias results from the sampling frame not including major parts of the population. However, in Figure 1.3.1 (b), the sampling frame is close to the population hence selection bias is minimized or avoided.
(ii) Defective measuring device. If the measuring device is constantly in error, it will result in a systematic bias in the data collected by using that device.

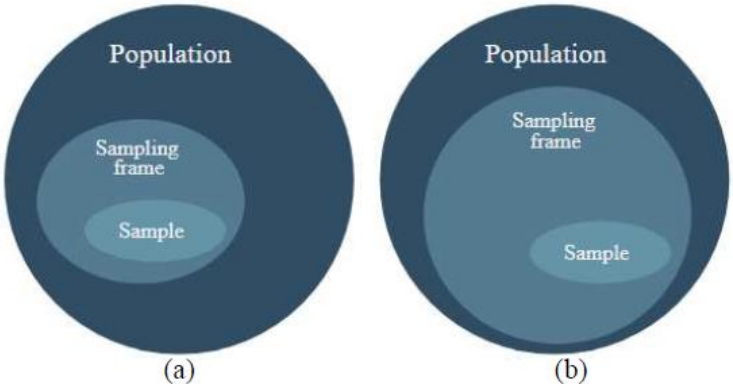


Figure 1.3.1 Illustration of sample selection within a sampling frame of population

In summary, a good sample design must:
(i) Result in a truly representative sample
(ii) Lead to only a small sampling error
(iii) Be cost-effective
(iv) Be one that controls the systematic bias
(v) Be one such that the results of the sample study can be applied for the population with a reasonable degree of confidence.

When considering the types of sample design, there are two main factors which are the basis for the existence of different sample designs. These factors are:
a) Representation basis: In this case, the samples may be drawn based on probability sampling or non-probability sampling. Probability sampling is based on the concept of random sampling, while non-probability sampling is based on the concept of non-random sampling. Probability sampling is also known as ' random sampling' or' chance sampling', where every item of the universe has an equal chance of inclusion in the sample. Random sampling ensures statistical regularity, which means that if, on average, the sample chosen is a random one, the sample will have the same composition and characteristics as the universe/ population. Non-probability sampling is also known by different names such as, deliberate sampling, purposive sampling, and judgement sampling. In this type of sampling, elements for the sample are selected deliberately by the researcher as per his choice (subjective). Suppose a region is surveyed for exploration activity for a gold-bearing lode. The area is divided into blocks. Some blocks may not be geologically favourable. Out of the favourable ones, a geologist may like to collect gold-bearing rock specimens as per his judgment or choice. In non-probability sampling design, personal element plays a great role. Quota sampling is also an example of non-probability sampling. Under this scheme, an interviewer is simply given quotas to be filled from the different strata with some restrictions on how they are to be filled. In other words, the actual selection of items for the sample is left to the interviewer's discretion. These samples so selected do not possess the characteristics of random samples. However, if the enumerator initially chooses units at random, rejecting those that are not needed, this method is equivalent to stratified random sampling.

b) Element selection technique: In this approach, the sample may be either unrestricted or restricted. When each sample element is drawn individually from the population, then the sample so drawn is known as the unrestricted sample, whereas all other forms of sampling are covered under the term 'restricted sampling'.

Table 1.3.1 Criteria for sample design selection and suitability for sampling methods

Representative Basis		Element selection technique
Probability sampling	Non-probability sampling	
Simple random sampling	Haphazard sampling Convenience sampling	Unrestricted sampling
Stratified random sampling Systematic sampling Cluster sampling etc.	Purposive sampling (such as quota sampling, judgment sampling etc.)	Restricted sampling

1.3.2 Environmental variables

Environmental variables are quantitative or descriptive measures of different environmental features. Environmental variables can belong to different domains, ranging from biology (distribution of species and biodiversity measures), soil science (soil properties and types), vegetation science (plant species and communities, land cover types), climatology (climatic variables at surface and beneath/above), hydrology (water quantities and conditions) and similar. They are commonly collected through field sampling (supported by remote sensing), which are then used to produce maps showing their distribution in an area. Such accurate and up-to-date maps of environmental features represent a crucial input to spatial planning, decision making, land evaluation or land degradation assessment.

From a meta-physical perspective, what we are most often mapping in geostatistics are, in fact, quantities of molecules of a certain kind. For example, a measure of soil or water acidity is the pH factor. By definition, pH is a negative exponent of the concentration of the H^+ ions. By mapping it over the whole area of interest, we produce a map of continuous values of concentration (continuous fields) of H^+ ions.

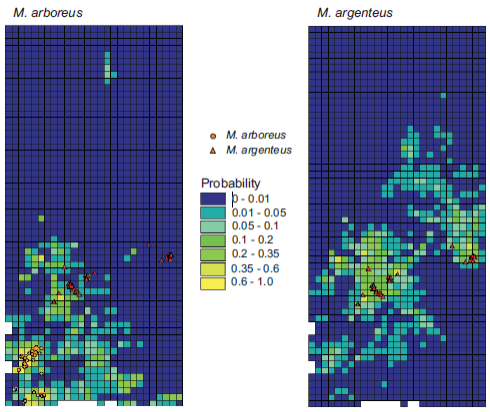


Figure 1.3.2 Example of species distribution estimation using Geostatistical mapping

In the case of plants and animals, geostatistical mapping becomes much more complicated. Here, we deal with distinct physical objects (individuals), often immeasurable in quantity. In addition, animal species change their location dynamically, often in unpredictable directions and with unpredictable spatial patterns (non-linear trajectories), which asks for high sampling density in both space and time domains.

To account for these problems, spatial modelers rarely aim at mapping the distribution of individuals (e.g. represented as points), but instead use compound measures that are suitable for management and decision making purposes. For example, animal species can be represented using density or biomass

measures. In vegetation mapping, most commonly, field observations of the plant occurrence (ranging from 0 to 100%) are recorded (Fig. 1.3.2). In addition to the mapping of the temporary distribution of species, biologists aim at developing statistical models to define optimal ecological conditions for a certain species. This is often referred to as habitat mapping and can also be dealt with in geostatistics. The occurrence of species or habitat conditions can also be presented as continuous fields, i.e. using raster maps.

Table 1.3.2 Some common environmental variables of interest to decision-making and their properties SRV — short-range variability; TV — temporal variability; VV — vertical variability; SSD — Standard sampling density; DRS — remote-sensing detectability. F — high — medium, − — low or non-existent.

Environmental features/topics	**Common variables of interest to decision making**	**SRV**	**TV**	**VV**	**SSD**	**RSD**
Mineral exploration: oil, gas, mineral resources	mineral occurrence and concentrations of minerals; reserves of oil and natural gas; magnetic anomalies;	⋆	−	★	⋆	⋆
Freshwater resources and water quality	O_2, ammonium and phosphorus concentrations in water; concentration of herbicides; trends in concentrations of pollutants; temperature change;	⋆	⋆	⋆	⋆	−
Socio-economic parameters	population density; population growth; GDP per km^2; life expectancy rates; human development index; noise intensity;	⋆	⋆	−	★	★
Land degradation: erosion, landslides, surface runoff	soil loss; erosion risk; quantities of runoff; dissolution rates of various chemicals; landslide susceptibility;	⋆	⋆	−	−	★
Natural hazards: fires, floods, earthquakes, oil spills	burnt areas; fire frequency; water level; earthquake hazard; financial losses; human casualties; wildlife casualties;	★	★	−	⋆	★
Human-induced radioactive contamination	gama doze rates; concentrations of isotopes; PCB levels found in human blood; cancer rates;	⋆	★	−	⋆	★
Soil fertility and productivity	organic matter, nitrogen, phosphorus and potassium in soil; biomass production; (grain) yields; number of cattle per ha; leaf area index;	★	⋆	⋆	⋆	⋆
Soil pollution	concentrations of heavy metals especially: arsenic, cadmium, chromium, copper, mercury, nickel, lead and hexachlorobenzene; soil acidity;	★	⋆	−	★	−
Distribution of animal species (wildlife)	occurrence of species; biomass; animal species density; biodiversity indices; habitat conditions;	★	★	−	⋆	−
Distribution of natural vegetation	land cover type; vegetation communities; occurrence of species; biomass; density measures; vegetation indices; species richness; habitat conditions;	⋆	⋆	−	★	★
Meteorological conditions	temperature; rainfall; albedo; cloud fraction; snow cover; radiation fluxes; net radiation; evapotranspiration;	⋆	★	⋆	⋆	★
Climatic conditions and changes	mean, minimum and maximum temperature; monthly rainfall; wind speed and direction; number of clear days; total incoming radiation; trends of changes of climatic variables;	−	★	⋆	⋆	⋆
Global atmospheric conditions	aerosol size; cirrus reflectance; carbon monoxide; total ozone; UV exposure;	⋆	★	★	−	★
Air quality in urban areas	NO_x, SO_2 concentrations; emission of greenhouse gasses; emission of primary and secondary particles; ozone concentrations; Air Quality Index;	★	★	★	★	−
Global and local sea conditions	chlorophyll concentrations; biomass; sea surface temperature; emissions to sea;	⋆	★	⋆	⋆	⋆

1.3.3 Aspects of spatial variability

Relevant and detailed geoformation is a prerequisite for the successful management of natural resources in many applied environmental and geosciences. Until recently, maps of environmental variables have primarily generated by using mental models (expert systems). Because field data collection is often the most expensive part of a survey teams typically visit only a limited number of sampling locations and then, based on the sampled data and statistical and/or mental models, infer conditions for the whole area of interest. As a consequence, maps of environmental variables have often been of limited and inconsistent quality and usually too subjective.

Spatial variability of environmental variables is commonly a result of complex processes working at the same time and over long periods of time, rather than an effect of a single realization of a single factor. To explain variation of environmental variables has never been an easy task. Many environmental variables vary not only horizontally but also with depth, not only continuously but also abruptly. Field observations are, on the other hand, usually very expensive and we are often forced to build 100% complete maps by using a sample of <<1%.

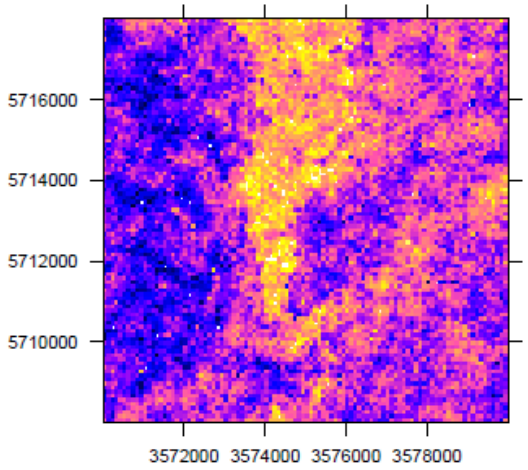


Fig. 1.3.3 If we were able to sample a soil variable over the whole area of interest, we would probably get an image such as t. This image was, in fact, produced using the geostatistical simulations with a regression-kriging model.

Imagine if we had enough funds to inventory each grid node in a study area, then we would be able to produce a map which would probably look as the map shown in Fig. 1.23. By carefully looking at this map, you can notice several things: (1) there seems to be a spatial pattern of how the values change; (2) values that are closer together are more similar; (3) locally, the values can differ without any systematic rule (randomly); (4) in the middle of the area, the values seem to be in general higher (a discrete change) etc. From the statistical perspective, an environmental variable can be viewed as an information signal consisting of three components:

$$Z(\mathbf{s}) = Z^*(\mathbf{s}) + \varepsilon'(\mathbf{s}) + \varepsilon''$$

Where Z*(s) is the deterministic component, ε'(s) is the spatially correlated random component and ε'' is the pure noise, usually the result of the measurement error. This model is in literature often referred to as the **universal model of variation.**

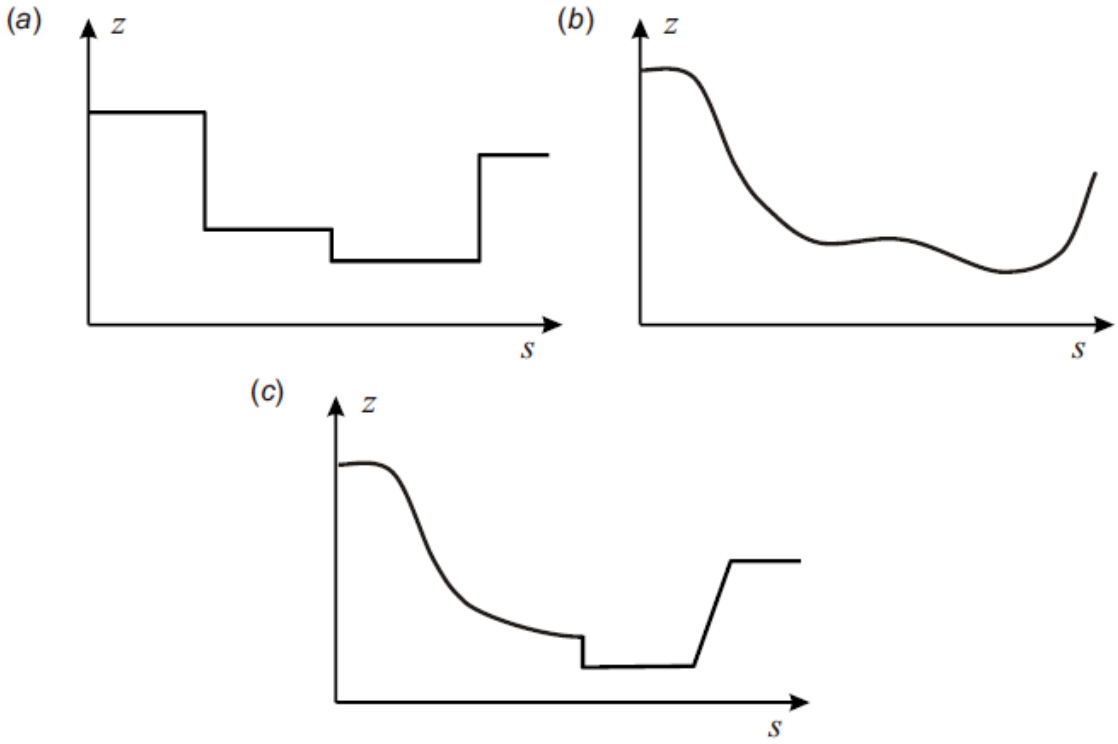
In theory, we could decompose a map of an environmental variable into two grids: (1) the deterministic and (2) the error surface; in practice, we are not able to do distinguish the

deterministic from the error part of the signal because both can show similar patterns. By collecting field measurements at different locations and with different sampling densities, we might be able to infer about the source of variability and estimate probabilistic models of variation (error budget assessment).This way we can try to answer questions such;
How much of the variation is due to the measurement error?
How much has been accounted for by the environmental factors?
How much is due to the spatial similarity of the values?
How much is uncorrelated noise?
Such systematic assessment of the error budget allows us to make realistic interpretations and utilize models which reflect our knowledge about the variability of target variables.

1.3.4 Step towards successful geostatistical mapping of environmental variables

- **Understand the sources of variability in the data**

As pointed earlier, the variability is a result of deterministic and stochastic processes plus the pure noise. In other words, the variability in data is a sum of two components:
(a) The natural spatial variation and
(b) The inherent noise, mainly do the measurement errors (Burrough and McDonnell, 1998). Measurement errors typically occur during the positioning in the field, during sampling or the laboratory analysis. These errors should ideally be minimized, because they are not of primary concern for a mapper. What the mappers are interested in is the **natural spatial variation**, which is mainly due to the physical processes that can be explained (up to a certain level) by a mathematical model.



Schematic examples of models of spatial variation: abrupt changes of values can be modelled using a discrete model of spatial variation (a), smooth changes can be modelled using a continuous model of spatial variation (b). In reality, we often need to work with a mixed (or hybrid) model of spatial variation (c).

- **Consider all aspects of natural variation.**

Although spatial prediction of environmental variables is primarily concerned with geographical variability, there are also other aspects of natural soil variation that are often overlooked by mappers: vertical, temporal and scale aspect. Overview of each aspect is described below

Geographical variation (2D) - The results of spatial prediction are either visualized as 2D maps or cross-sections. Some environmental variables, such as thickness of soil horizons, the occurrence of vegetation species or soil types, do not have a third dimension, i.e. they refer to the Earth's surface only. Others, such as temperature, population densities etc. can be measured at various altitudes, even below Earth's surface. Geographical part of variation can be modelled using either a continuous, discrete or mixed model of spatial variation.

Vertical variation (3D) many environmental variables also vary with depth or altitude. In many cases, the measured difference between the values is higher at a depth differing by a few centimetres than at geographical distance of few meters.

Transition between different soil layers, for example, can also be both gradual and abrupt, which requires a double-mixed model of soil variation for 3D spatial prediction.

Temporal variation Environmental variables connected with animal and plant species vary not only within season but often within few moments. Even the soil variables such as pH, nutrients, water-saturation levels and water content, can vary over a few years, within a single season or even over a few days. Temporal variability makes geostatistical mapping especially complex and expensive. Maps of environmental variables produced for two different time references can differ significantly. This means that most of maps are valid for a certain period (or moment) of time only. In many cases the seasonal periodicity of environmental variables is regular, so that prediction does not necessarily require new samples.

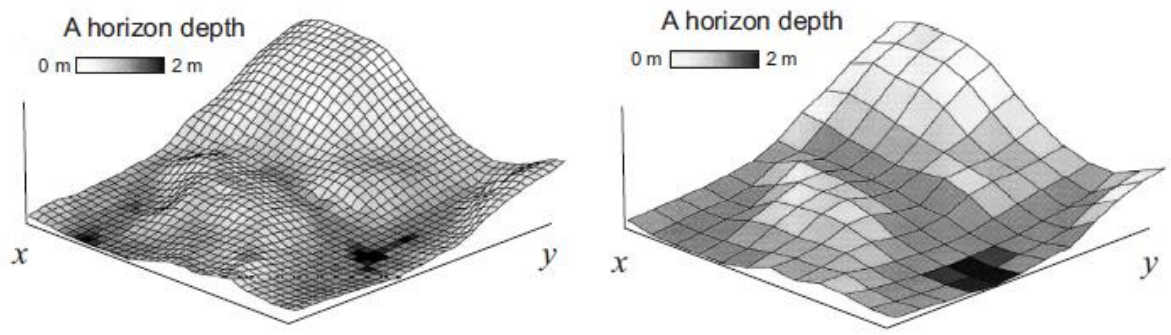


Fig. 1.4: Influence of the support (grid cell) size: predictions of the same variable at coarse grid will often show much less contrast. Example from Thompson et al. (2001).

Support size Support size is the discretisation level of a geographical surface and is related to the concept of scale. In the case of spatial predictions, there are two support sizes: the size of the blocks of land sampled, and grid resolution of the auxiliary maps. Field observations are typically collected as point samples.

The support size of the auxiliary maps is commonly much larger than the actual blocks of land sampled, e.g. auxiliary variables are in general averaged (smoothed), while the environmental variables can describe local (micro) features. As a result, the correlation between the auxiliary maps and measured environmental variables is often low or insignificant (Fig. 1.4). There are two solutions to this problem: (a) To up-scale the auxiliary maps or work with super-high resolution/detail data (e.g. IKONOS images of 1 m resolution), or (b) to average bulk or composite samples within the regular blocks of land (Patil, 2002).

The first approach is more attractive for the efficiency of prediction, but at the cost of more processing power and storage. The second solution will only result in a better fit, whereas the efficiency of

prediction, validated using point observations, may not change significantly. This means that mixing of lab data from different seasons, depths and with different support sizes in general means lower predictive power and problems in fully interpreting the results. If the focus of prediction modelling is solely the geographical component (2D), then the samples need to be taken under fixed conditions: same season, same depths, and same blocks of land. This also means that each 2D map of an environmental variable should always indicate a time reference (interval), applicable vertical dimension and the sample (support) size i.e. the effective scale.