1

Geostatistical mapping

1.1 Basic concepts

Any measurement we take in Earth and environmental sciences, although this is often ignored, has a spatio-temporal reference. A spatio-temporal 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);

(3.) *time of measurement* (year, month, day, hour, minute etc.);

(4.) *spatio-temporal support* (size of the blocks of material associated with measurements; time interval of measurement);

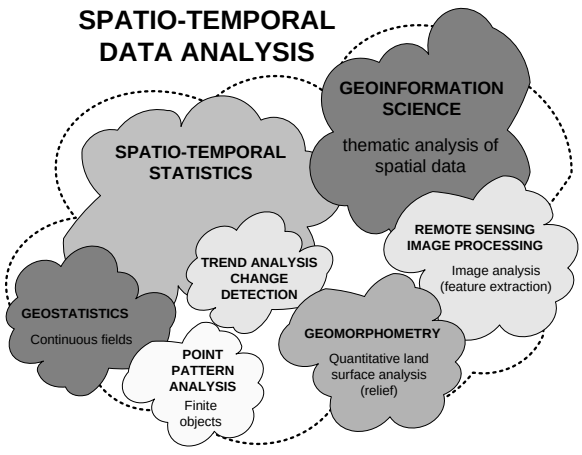

Fig. 1.1: Spatio-temporal Data Analysis is a group of research fields and sub-fields.

If at least geographical coordinates are assigned to the measurements, then we can analyze and visualize them 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 **Spatio-temporal Data Analysis** (Fig. 1.1). 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 spatio-temporal data analysis (**STDA**) is a combination of two major sciences: *geoinformation science* and *spatio-temporal statistics*; or in mathematical terms: STDA = GIS + statistics. This book focuses mainly on some parts of STDA, although many of the principles we will touch in this guide are common for any type of STDA.

As mentioned previously, **geostatistics** is a subset of statistics specialized in analysis and interpretation of geographically referenced data (Goovaerts, 1997). Cressie (1993) considers geostatistics to be only one of the three scientific fields specialized in the analysis of spatial data — the other two being *point pattern* analysis (focused on point objects; so called *"point-processes"*) and *lattice*[1] statistics (polygon objects) (Fig. 1.2).

[1] The term *lattice* here refers to discrete spatial objects.

For Ripley (2004), spatial statistics is a process of extracting data summaries from spatial data and comparing these to theoretical models that explain how spatial patterns originate and develop. 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)**. The analysis of spatio-temporally referenced data is certainly different from what you have studied so far within other fields of statistics, but there are also many direct links as we will see later in §2.1.

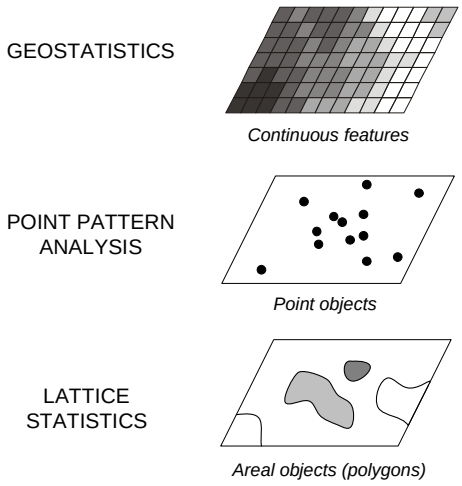

Fig. 1.2: Spatial statistics and its three major subfields after Cressie (1993).

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 (Bolstad, 2008). Today, geostatistics is not only used to analyze point data, but also increasingly in combination with various GIS data sources: e.g. to explore spatial variation in remotely sensed data, to quantify noise in the images and for their filtering (e.g. filling of the voids/missing pixels), to improve DEM generation and for simulations (Kyriakidis et al., 1999; Hengl et al., 2008), to optimize spatial sampling (Brus and Heuvelink, 2007), selection of spatial resolution for image data and selection of support size for ground data (Atkinson and Quattrochi, 2000).

According to the bibliographic research of Zhou et al. (2007) and Hengl et al. (2009a), the top 10 application fields of geostatistics are: (1) geosciences, (2) water resources, (3) environmental sciences, (4) agriculture and/or soil sciences, (5/6) mathematics and statistics, (7) ecology, (8) civil engineering, (9) petroleum engineering and (10) meteorology. The most influential (highest citation rate) books in the field are: Cressie (1993), Isaaks and Srivastava (1989), Deutsch and Journel (1998), Goovaerts (1997), and more recently Banerjee et al. (2004). These lists could be extended and they differ from country to country of course. The evolution of applications of geostatistics can also be followed through the activities of the following research groups: International Association of Mathematical Geosciences[2] (IAMG), geoENVia[3], pedometrics[4], R-sig-geo[5], spatial accuracy[6] and similar. The largest international conference that gathers geostatisticians is the GEOSTATS conference, and is held every four years; other meetings dominantly focused on the field of geostatistics are GEOENV, STATGIS, and ACCURACY.

[2] `http://www.iamg.org`
[3] `http://geoenvia.org`
[4] `http://pedometrics.org`
[5] `http://cran.r-project.org/web/views/Spatial.html`
[6] `http://spatial-accuracy.org`

For Diggle and Ribeiro Jr (2007), there are three scientific objectives of geostatistics:

(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**;

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. 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. It is often worth investing extra time to enhance the analysis and get a reliable estimate of probability associated with some important hypothesis, especially if the result affects long-term decision making. The end result of hypothesis testing is commonly a single number (probability) or a binary decision (Accept/Reject). Spatial prediction, on the other hand, is usually computationally intensive, so that sometimes, for pragmatic reasons, naïve 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; 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. This will become much clearer when we jump from the demo exercise in chapter 5 to a real case study in chapter 6.

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. We will more commonly use the term *spatial prediction* in this handbook, even though the term *spatial interpolation* has been more widely accepted (Lam, 1983; Mitas and Mitasova, 1999; Dubois and Galmarini, 2004). In geostatistics, e.g. in the case of ordinary kriging, interpolation corresponds to cases where the location being estimated is surrounded by the sampling locations and is within the spatial auto-correlation range. Prediction outside of the practical range (prediction error exceeds the global variance) is then referred to as **extrapolation**. In other words, extrapolation is prediction at locations where we do not have enough statistical evidence to make significant predictions.

An important distinction between geostatistical and conventional mapping of environmental variables is that geostatistical prediction is based on application of quantitative, statistical techniques. Until recently, maps of environmental variables have been primarily been generated by using mental models (expert systems). Unlike the traditional approaches to mapping, which rely on the use of empirical knowledge, in the case of **geostatistical mapping** we completely rely on the actual measurements and semi-automated algorithms. Although this sounds as if the spatial prediction is done purely by a computer program, the analysts have many options to choose whether to use linear or non-linear models, whether to consider spatial position or not, whether to transform or use the original data, whether to consider multicolinearity effects or not. So it is also an expert-based system in a way.

In summary, geostatistical mapping can be defined as **analytical production of maps by using field observations, explanatory information, and a computer program that calculates values at locations of interest** (a study area). It typically comprises:

(1.) design of sampling plans and computational workflow,

(2.) field data collection and laboratory analysis,

(3.) model estimation using the sampled point data (calibration),

(4.) model implementation (prediction),

(5.) model (cross-)evaluation using validation data,

(6.) final production and distribution of the output maps[7].

[7] By this I mainly think of on-line databases, i.e. data distribution portals or Web Map Services and similar.

Today, increasingly, the natural resource inventories need to be regularly updated or improved in detail, which means that after step (6), we often need to consider collection of new samples or additional samples that are then used to update an existing GIS layer. In that sense, it is probably more valid to speak about **geostatistical monitoring**.

1.1.1 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), to hydrology (water quantities and conditions) and similar (Table 1.1). They are commonly collected through field sampling (supported by remote sensing); field samples 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 and/or land degradation assessment. For example, according to Sanchez et al. (2009), the main challenges of our time that require high quality environmental information are: food security, climate change, environmental degradation, water scarcity and threatened biodiversity.

Because field data collection is often the most expensive part of a survey, 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 are usually too subjective. Field sampling is gradually being replaced with **remote sensing systems** and **sensor networks**. For example, elevations marked on topographic maps are commonly collected through land survey i.e. by using geodetic instruments. Today, airborne technologies such as LiDAR are used to map large areas with $\gg$1000 times denser sampling densities. Sensor networks consist of distributed sensors that automatically collect and send measurements to a central service (via GSM, WLAN or radio frequency). Examples of such networks are climatological stations, fire monitoring stations, radiological measurement networks and similar.

From a meta-physical perspective, what we are most often mapping in geostatistics are, in fact, **quantities of molecules of a certain kind or quantities of energy**[8]. 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. It is often important to understand the meaning of an environmental variable: for example, in the case of pH, we should know that the quantities are already on a log-scale so that no further transformation of the variable is anticipated (see further §5.4.1). By mapping pH over the whole area of interest, we will produce a continuous map of values of concentration (continuous fields) of H^+ ions.

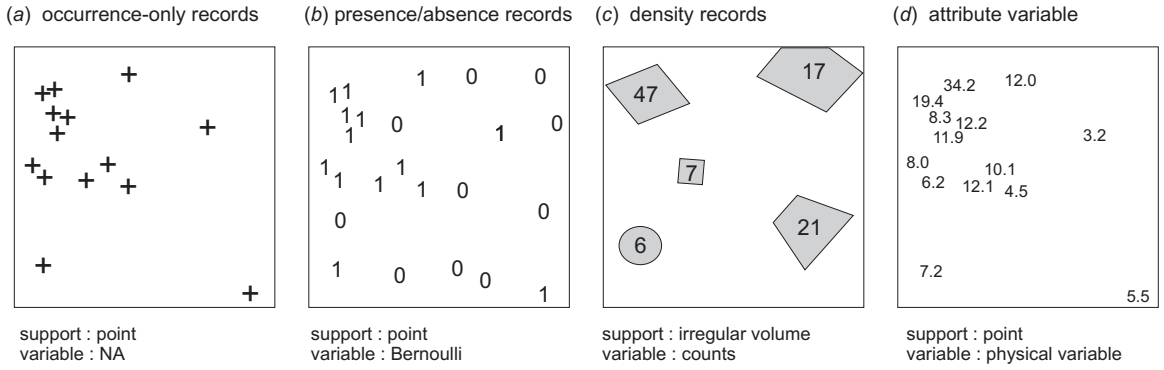

Fig. 1.3: Types of field records in ecology.

In the case of plants and animals inventories, geostatistical mapping is somewhat more complicated. Plants or animals are distinct physical **objects** (individuals), often immeasurable in quantity. In addition, animal

[8]There are few exceptions of course: elevation of land surface, wind speed (kinetic energy) etc.

species change their location dynamically, frequently 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 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 (see e.g. Latimer et al. (2004) and/or Pebesma et al. (2005)).

In vegetation mapping, most commonly field observations of the plant occurrence are recorded in terms of area coverage (from 0 to 100%). In addition to mapping of temporary distribution of species, biologists aim at developing statistical models to define optimal ecological conditions for certain species. This is often referred to as **habitat mapping** or niche modeling (Latimer et al., 2004). Densities, occurrence probability and/or abundance of species or habitat conditions can also be presented as continuous fields, i.e. using raster maps. Field records of plants and animals are more commonly analyzed using point pattern analysis and factor analysis, than by using geostatistics. The type of statistical technique that is applicable to a certain observations data set is mainly controlled by the nature of observations (Fig. 1.3). As we will show later on in §8, with some adjustments, standard geostatistical techniques can also be used to produce maps even from occurrence-only records.

1.1.2 Aspects and sources of spatial variability

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 (Table 1.1). 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 $\ll$1%.

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. 1.4[9]. 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 some parts of the area, the values seem to be in general higher i.e. there is a discrete jump in values.

From the information theory perspective, an environmental variable can be viewed as a *signal process* consisting of three components:

$$Z(\mathbf{s}) = Z^*(\mathbf{s}) + \varepsilon'(\mathbf{s}) + \varepsilon'' \qquad (1.1.1)$$

where $Z^*(\mathbf{s})$ is the deterministic component, $\varepsilon'(\mathbf{s})$ is the spatially correlated random component and ε'' is the pure noise — partially micro-scale variation, partially the measurement error. This model is, in the literature, often referred to as the **universal model of variation** (see further §2.1). Note that we use a capital letter Z because we assume that the model is probabilistic, i.e. there is a range of equiprobable realizations of the same model $\{Z(\mathbf{s}), \mathbf{s} \in \mathbb{A}\}$; $Z(\mathbf{s})$ indicates that the variable is dependent on the location $\mathbf{s}$.

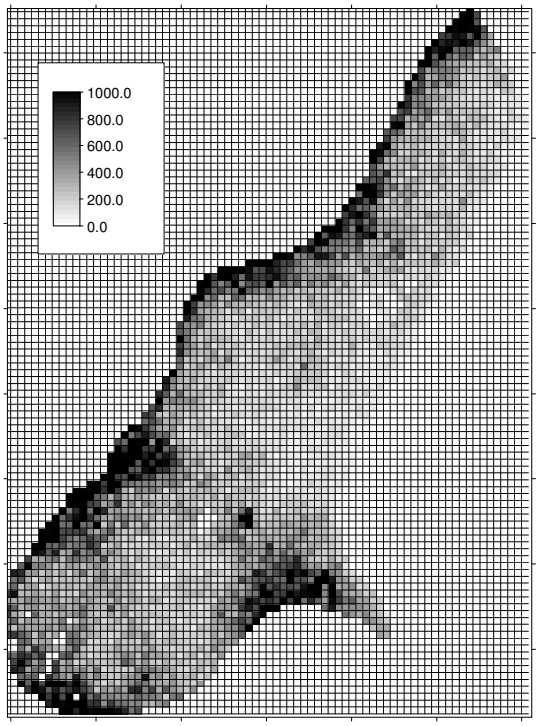

Fig. 1.4: If we were able to sample a variable (e.g. zinc concentration in soil) regularly over the whole area of interest (each grid node), we would probably get an image such as this.

[9] This image was, in fact, produced using geostatistical simulations with a regression-kriging model (see further Fig. 2.1 and Fig. 5.12; §5.5.1).

Table 1.1: 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; RSD — remote-sensing detectability. ★ — high, ⋆ — medium, − — low or non-existent. Levels approximated by the author.

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;	⋆	⋆	−	★	★
Health quality data	number of infections; hospital discharge; disease rates per 10,000; mortality rates; health risks;	−	⋆	−	★	−
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; GPS trajectories (speed); 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;	⋆	★	⋆	⋆	⋆

In theory, we could decompose a map of a target environmental variable into two grids: (1) the deterministic part (also know as the *trend surface*), and (2) the *error surface*; in practice, we are not able to distinguish the deterministic from the error part of the signal, because both can show similar patterns. In fact, even if we sample every possible part of the study area, we can never be able to reproduce the original signal exactly because of the measurement error. 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 spatial variation. Then we can try to answer how much of the variation is due to the measurement error, how much has been accounted for by the environmental factors, and how much is due to the spatial proximity. Such systematic assessment of the error budget allows us to make realistic interpretations of the results and correctly reason about the variability of the feature of interest.

The first step towards successful geostatistical mapping of environmental variables is to understand the sources of variability in the data. As we have seen previously, 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** (ε''), mainly due to the measurement errors (Burrough and McDonnell, 1998). Measurement errors typically occur during positioning in the field, during sampling or 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.

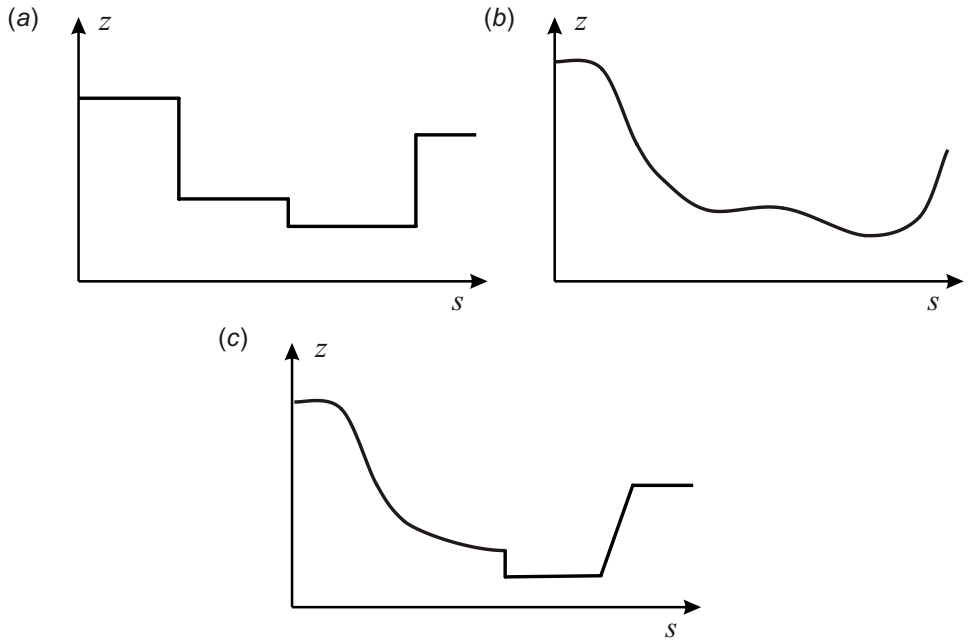

Fig. 1.5: Schematic examples of models of spatial variation: abrupt changes of values can be modeled using a discrete model of spatial variation (a), smooth changes can be modeled using a continuous model of spatial variation (b). In reality, we often need to work with a mixed (or hybrid) model of spatial variation (c).

Physical processes that dominantly control environmental variables differ depending of the type of feature of interest (Table 1.1). In the most general terms, we can say that there are five major factors shaping the status of environment on Earth:

abiotic (global) factors — these include various natural forces that broadly shape the planet. For example, Earth's gravity, rotation cycle, geological composition and tectonic processes etc. Because abiotic factors are relatively constant/systematic and cannot really be controlled, they can be regarded as global fixed conditions.

biotic factors — these include various types of living organism, from microbiological to animal and plant species. Sometimes living organisms can be the major factor shaping environmental conditions, even for wide areas.

anthropogenic factors — these include industrial and agricultural activities, food, water and material consumption, construction of dams, roads and similar. Unfortunately, the human race has irreversibly changed the environment in a short period of time. Extreme examples are the rise in global temperature, loss of biodiversity and deforestation.

transport and diffusion processes — these work upon other abiotic and biotic factors and shape the landscape locally. Unlike global factors, they are often non-linear and highly stochastic.

extra-terrestrial factors — including factors that control climate (e.g. incoming solar radiation, factors that control ice ages etc.), tectonic processes (meteors) and similar.

To illustrate how various factors shape an environmental feature, we can look at land surface (topography) as an example. Land surface is formed, first, as the result of tectonic and volcanic processes. Erosional processes further produce hydrological patterns (river networks, terraces, plains etc.). Living organisms produce soil material and form specific landscapes etc. In some cases extreme events happen such as fall of meteorites, that can suddenly completely change the initial conditions. Again, all these factor work in combination and often with chaotic behavior, so that no simple simulation model of land surface evolution can be constructed. Hence the only way to get an accurate estimate of land surface is to sample.

The second step towards reliable modeling of environmental variables is to 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: the *vertical*, *temporal* and *scale* aspects. Below is an overview of the main concepts and problems associated with each of these (see also Table 1.1):

Geographical variation (2D) The results of spatial prediction are either visualised 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, carbon monoxide concentrations etc. can be measured at various altitudes, even below Earth's surface. Geographical part of variation can be modeled using either a **continuous**, **discrete** or **mixed model of spatial variation** (Fig. 1.5).

Vertical variation (3D) Many environmental variables also vary with depth or altitude above the ground surface. In many cases, the measured difference between the values is higher at a depth differing by a few centimeters than at geographical distance of few meters. Consider variables such as temperature or bird density — to explain their vertical distribution can often be more difficult than for the horizontal space. 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. Some authors suggest the use of cumulative values on volume (areal) basis to simplify mapping of the 3D variables. For example, McKenzie and Ryan (1999) produced maps of total phosphorus and carbon estimated in the upper 1 m of soil and expressed in tons per hectare, which then simplifies production and retrieval. See also further section 7.6.

Temporal variation As mentioned previously, environmental variables connected with distribution of animal and plant species vary not only within a season but also within few moments. Even 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 (Heuvelink and Webster, 2001). Temporal variability makes geostatistical mapping especially complex and expensive. Maps of environmental variables produced for two different times can differ significantly. Changes can happen abruptly in time. This means that most of the 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 we do not necessarily require very dense sampling in time domain (see further §2.5).

Support size Support size is the size or volume associated with measurements, but is also connected with properties such as shape and orientation of areas associated with measurements. Changing the support of a variable creates a different variable which is related to the original, but has different spatial properties (Gotway and Young, 2002). The concept of spatial support should not be confused with the various discretization level of measurements. In the case of spatial predictions, there are two spatial discretization levels: the size of the blocks of land sampled (support size), and grid resolution of the auxiliary maps. Both concepts are closely related with cartographic scale (Hengl, 2006). 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. explanatory 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. 1.6). There are two solutions to this problem: (a) to up-scale the auxiliary maps or work with high resolution satellite images, 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.

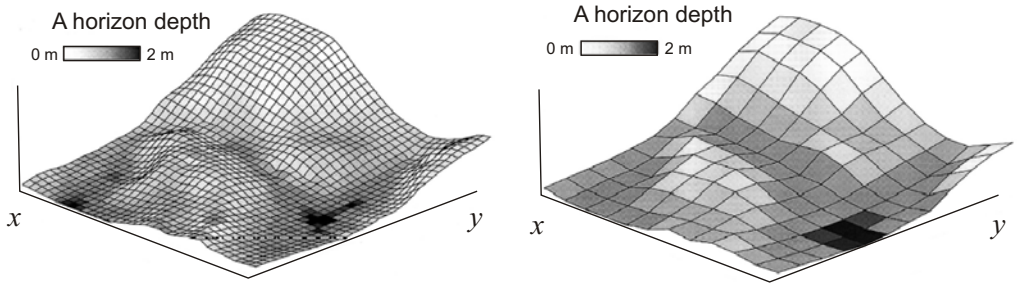

Fig. 1.6: Influence of the support (grid cell) size: predictions of the same variable at coarse grid will often show much less contrast, i.e. it will miss many local hot-spots. Example from Thompson et al. (2001).

In practice, given the space-time domain and feature of interest, one makes measurements by fixing either 2D space, elevation/depth or time. 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 modeling is solely the geographical component (2D), then the samples need to be taken under fixed conditions: same season, same depths, same blocks of land. Likewise, if the focus of analysis is generation of spatio-temporal patterns, some minimum of point samples in both space and time domains is needed. Analysts that produce 2D maps often ignore concepts such as temporal variability and support size. To avoid possible misinterpretation, each 2D map of environmental variables generated using geostatistics **should always indicate a time reference (interval), applicable vertical dimension[10], sampling locations, borders of the area of interest, and the size of sampling blocks (support size)**.

1.1.3 Spatial prediction models

In an ideal situation, variability of environmental variables is determined by a finite set of inputs and they exactly follow some known physical law. If the algorithm (formula) is known, the values of the target variables can be predicted exactly. In reality, the relationship between the feature of interest and physical environment is so complex[11] that it cannot be modeled exactly (Heuvelink and Webster, 2001). This is because we either do not exactly know: (a) the final list of inputs into the model, (b) the rules (formulas) required to derive the output from the inputs and (c) the significance of the random component in the system. So the only possibility is that we try to estimate a model by using the actual field measurements of the target variable. This can be referred to as the *indirect* or *non-deterministic* estimation.

Let us first define the problem using mathematical notation. Let a set of observations of a **target variable** (also known as *response* variable) Z be denoted as $z(\mathbf{s}_1)$, $z(\mathbf{s}_2)$,..., $z(\mathbf{s}_n)$, where $\mathbf{s}_i = (x_i, y_i)$ is a location and x_i and y_i are the coordinates (primary locations) in geographical space and n is the number of observations (Fig. 1.7). The geographical domain of interest (area, land surface, object) can be denoted as $\mathbb{A}$. We deal with only one reality (samples $z(\mathbf{s}_n)$), which is a realization of a process ($\mathbf{Z} = \{Z(\mathbf{s}), \forall \mathbf{s} \in \mathbb{A}\}$) that could have produced many realities.

[10] Orthogonal distance from the ground surface.

[11] Because either the factors are unknown, or they are too difficult to measure, or the model itself would be too complex for realistic computations.

Assuming that the samples are *representative*, *non-preferential* and *consistent*, values of the target variable at some new location $\mathbf{s}_0$ can be derived using a **spatial prediction model**. In statistical terms, a spatial prediction model draws realizations — either the most probable or a set of equiprobable realizations — of the feature of interest given a list of inputs:

$$\hat{z}(\mathbf{s}_0) = E\left\{Z|z(\mathbf{s}_i),\, q_k(\mathbf{s}_0),\, \gamma(\mathbf{h}),\, \mathbf{s} \in \mathbb{A}\right\} \tag{1.1.2}$$

where $z(\mathbf{s}_i)$ is the input point data set, $\gamma(\mathbf{h})$ is the covariance model defining the spatial autocorrelation structure (see further Fig. 2.1), and $q_k(\mathbf{s}_0)$ is the list of deterministic predictors, also known as *covariates* or explanatory variables, which need to be available at any location within $\mathbb{A}$. In other words, a spatial prediction model comprises list of procedures to generate predictions of value of interest given the calibration data and spatial domain of interest.

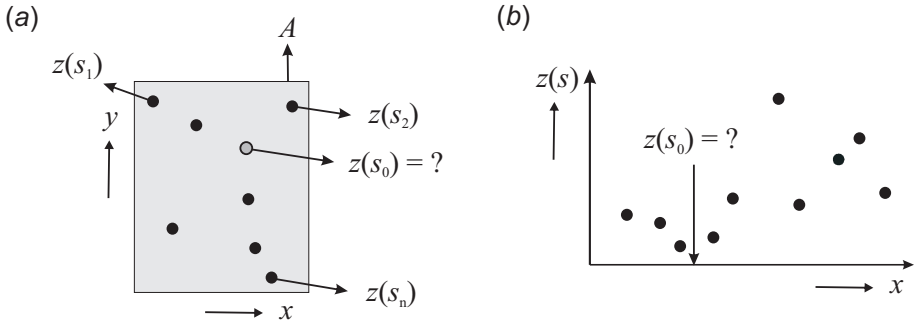

Fig. 1.7: Spatial prediction is a process of estimating the value of (quantitative) properties at unvisited site within the area covered by existing observations: (a) a scheme in horizontal space, (b) values of some target variable in a one-dimensional space.

In raster GIS terms, the geographical domain of interest is a rectangular matrix, i.e. an array with rows×columns number of grid nodes over the domain of interest (Fig. 1.8):

$$\mathbf{z} = \left\{z(\mathbf{s}_j), j = 1, \ldots, m\right\}; \quad \mathbf{s}_j \in \mathbb{A} \tag{1.1.3}$$

where $\mathbf{z}$ is the data array, $z(\mathbf{s}_j)$ is the value at the grid node $\mathbf{s}_j$, and m is the total number of grid nodes. Note that there is a difference between predicting values at grid node (punctual) and prediction values of the whole grid cell (block), which has a full topology[12].

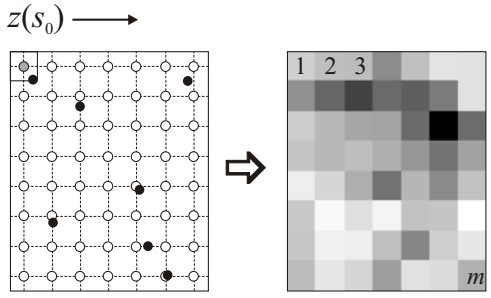

Fig. 1.8: Spatial prediction implies application of a prediction algorithm to an array of grid nodes (*point á point* spatial prediction). The results are then displayed using a raster map.

There seem to be many possibilities to interpolate point samples. At the Spatial Interpolation Comparison 2004 exercise, for example, 31 algorithms competed in predicting values of gamma dose rates at 1008 new locations by using 200 training points (Dubois and Galmarini, 2004; Dubois, 2005). The competitors ranged from splines, to neural networks, to various kriging algorithms. Similarly, the software package Surfer[13] offers dozens of interpolation techniques: Inverse Distance, Kriging, Minimum Curvature, Polynomial Regression, Triangulation, Nearest Neighbor, Shepard's Method, Radial Basis Functions, Natural Neighbor, Moving Average, Local Polynomial, etc. The list of interpolators available in R via its interpolation packages (akima, loess, spatial, gstat, geoR etc.) is even longer.

[12] The sp package in R, for example, makes a distinction between the Spatial Pixel data frame (grid nodes) and a Spatial Grid data frame (grid cells) to distinguish between regular grid with point support and block support.

[13] http://www.ssg-surfer.com

An inexperienced user will often be challenged by the amount of techniques to run spatial interpolation. Li and Heap (2008), for example, list over 40 unique techniques in their extensive review of the spatial prediction methods. Most spatial prediction models are in fact somehow connected. As we will see later on, many standard linear models are in fact just a special case of a more general prediction model. This makes things much less complicated for the non-geostatisticians[14]. It is thus more important to have a clear idea about the connection or hierarchy of predictors, than to be able to list all possible predictors and their variants.

Spatial prediction models (algorithms) can be classified according to the amount of statistical analysis i.e. amount of expert knowledge included in the analysis:

(1.) **MECHANICAL (DETERMINISTIC) MODELS** — These are models where arbitrary or empirical model parameters are used. No estimate of the model error is available and usually no strict assumptions about the variability of a feature exist. The most common techniques that belong to this group are:

- *Thiessen polygons*;
- *Inverse distance interpolation*;
- *Regression on coordinates*;
- *Natural neighbors*;
- *Splines*;
- ...

(2.) **LINEAR STATISTICAL (PROBABILITY) MODELS** — In the case of statistical models, the model parameters are commonly estimated in an objective way, following probability theory. The predictions are accompanied with an estimate of the prediction error. A drawback is that the input data set usually need to satisfy strict statistical assumptions. There are at least four groups of linear statistical models:

- *kriging* (plain geostatistics);
- *environmental correlation* (e.g. regression-based);
- *Bayesian-based models* (e.g. Bayesian Maximum Entropy);
- *hybrid models* (e.g. regression-kriging);
- ...

(3.) **EXPERT-BASED SYSTEMS** — These models can be completely subjective (*ergo* irreproducible) or completely based on data; predictions are typically different for each run. Expert systems can also largely be based on probability theory (especially Bayesian statistics), however, it is good to put them in a different group because they are conceptually different from standard linear statistical techniques. There are at least three groups of expert based systems:

- *mainly knowledge-driven expert system* (e.g. hand-drawn maps);
- *mainly data-driven expert system* (e.g. based on neural networks);
- *machine learning* algorithms (purely data-driven);

Spatial prediction models can also be classified based on the:

Smoothing effect — whether the model smooths predictions at sampling locations or not:

- *Exact* (measured and estimated values coincide);
- *Approximate* (measured and estimated values do not have to coincide);

Transformation of a target variable — whether the target variable is used in its original scale or transformed:

- *Untransformed or Gaussian* (the variable already follows close to a normal distribution);

[14] As we will see later on in §2.1.2, spatial prediction can even be fully automated so that a user needs only to provide quality inputs and the system will select the most suitable technique.

- *Trans-Gaussian* (variable transformed using some link function);

Localization of analysis — whether the model uses all sampling locations or only locations in local proximity:

- *Local* or *moving window* analysis (a local sub-sample; local models applicable);
- *Global* (all samples; the same model for the whole area);

Convexity effect — whether the model makes predictions outside the range of the data:

- *Convex* (all predictions are within the range);
- *Non-convex* (some predictions might be outside the range);

Support size — whether the model predicts at points or for blocks of land:

- *Point-based* or punctual prediction models;
- *Area-based* or block prediction models;

Regularity of support — whether the output data structure is a grid or a polygon map:

- *Regular* (gridded outputs);
- *Irregular* (polygon maps);

Quantity of target variables — whether there is one or multiple variables of interest:

- *Univariate* (model is estimated for one target variable at a time);
- *Multivariate* (model is estimated for multiple variables at the same time);

Another way to look at spatial prediction models is to consider their ability to represent models of spatial variation. Ideally, we wish to use a mixed model of spatial variation (Fig. 1.5c) because it is a generalization of the two models and can be more universally applied. In practice, many spatial prediction models are limited to one of the two models of spatial variation: predicting using polygon maps (§1.3.3) will show discrete changes (Fig. 1.5a) in values; ordinary kriging (§1.3.1) will typically lead to smooth maps (Fig. 1.5b).

1.2 Mechanical spatial prediction models

As mentioned previously, mechanical spatial prediction models can be very flexible and easy to use. They can be considered to be subjective or empirical, because the user him/her-self selects the parameters of the model, often without any deeper analysis, often based only on a visual evaluation — the '*look good*' assessment. Most commonly, a user typically accepts the default parameters suggested by some software, hence the name *mechanical* models. The most widely used mechanical spatial prediction models are Thiessen polygons, inverse distance interpolation, regression on coordinates and various types of splines (Lam, 1983; Myers, 1994; Mitas and Mitasova, 1999). In general, mechanical prediction models are more primitive than statistical models and are often sub-optimal. However, in some situations they can perform as well as statistical models (or better).

1.2.1 Inverse distance interpolation

Probably one of the oldest spatial prediction techniques is **inverse distance interpolation** (Shepard, 1968). As with many other spatial predictors, in the case of inverse distance interpolation, a value of target variable at some new location can be derived as a weighted average:

$$\hat{z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i(\mathbf{s}_0) \cdot z(\mathbf{s}_i) \tag{1.2.1}$$

where λ_i is the weight for neighbor i. The sum of weights needs to equal one to ensure an unbiased interpolator. Eq.(1.2.1) in matrix form is:

$$\hat{z}(\mathbf{s}_0) = \lambda_0^{\mathrm{T}} \cdot \mathbf{z} \tag{1.2.2}$$

The simplest approach for determining the weights is to use the **inverse distances** from all points to the new point:

$$\lambda_i(\mathbf{s}_0) = \frac{\frac{1}{d^\beta(\mathbf{s}_0,\mathbf{s}_i)}}{\sum_{i=0}^{n}\frac{1}{d^\beta(\mathbf{s}_0,\mathbf{s}_i)}}; \qquad \beta > 1 \tag{1.2.3}$$

where $d(\mathbf{s}_0, \mathbf{s}_i)$ is the distance from the new point to a known sampled point and β is a coefficient that is used to adjust the weights. The principle of using inverse distances is largely a reflection of Waldo Tobler's first law in geography which states that "*Everything is related to everything else, but near things are more related than distant things.*" (Tobler, 1970, p.236); hence, points which are close to an output pixel will obtain large weights and that points which are farther away from an output pixel will obtain small weights. The β parameter is used to *emphasize* spatial similarity. If we increase β less importance will be put on distant points. The remaining problem is how to estimate β objectively so that it reflects the true strength of auto-correlation.

Inverse distance interpolation is an exact, convex interpolation method that fits only the continuous model of spatial variation. For large data sets ($\gg 10^3$ points) it can be time-consuming so it is often a good idea to set a threshold distance (search radius) to speed up the calculations.

1.2.2 Regression on coordinates

Assuming that the values of a target variable at some location are a function of coordinates, we can determine its values by finding a function which passes through (or close to) the given set of discrete points. This group of techniques can be termed *regression on coordinates*, although it is primarily known in literature by names **trend surfaces** and/or **moving surface interpolation**, depending on whether the function is fitted for the whole point data set (trend) or for a local (moving) neighbourhood (Hardy, 1971). Regression on coordinates is based on the following model (Webster and Oliver, 2001, p.40–42):

$$Z(\mathbf{s}) = f(x, y) + \varepsilon \tag{1.2.4}$$

and the predictions are made by:

$$\hat{z}(\mathbf{s}_0) = \sum_{r,s\in n} a_{rs} \cdot x^r y^s = \mathbf{a^T} \cdot \mathbf{s_0} \tag{1.2.5}$$

where $r + s < p$ is the number of transformations of coordinates, p is the order of the surface. The model coefficients ($\mathbf{a}$) are determined by maximizing the local fit:

$$\sum_{i=1}^{n} (\hat{z}_i - z_i)^2 \rightarrow \min \tag{1.2.6}$$

which can be achieved by the **Ordinary Least Squares** solution (Kutner et al., 2004):

$$\mathbf{a} = \left(\mathbf{s^T} \cdot \mathbf{s}\right)^{-1} \cdot \left(\mathbf{s^T} \cdot \mathbf{z}\right) \tag{1.2.7}$$

In practice, local fitting of the moving surface is more widely used to generate maps than trend surface interpolation. In the case of a moving surface, for each output grid node, a polynomial surface is fitted to a larger[15] number of points selected by a moving window (circle). The main problem of this technique is that, by introducing higher order polynomials, we can generate many artifacts and cause serious overshooting of the values locally (see further Fig. 1.13). A moving surface will also completely fail to represent discrete changes in space.

[15] The number of points needs to be at least larger than the number of parameters.

Regression on coordinates can be criticized for not relying on empirical knowledge about the variation of a variable (Diggle and Ribeiro Jr, 2007, p.57). As we will see later on in §1.3.2, it is probably advisable to avoid using x, y coordinates and their transforms and instead use **geographic predictors** such as the distance from a coast line, latitude, longitude, distance from water bodies and similar. Similar recommendation also applies to Universal kriging (see p.36) where coordinates are used to explain the deterministic part of variation.

1.2.3 Splines

A special group of interpolation techniques is based on **splines**. A spline is a type of piecewise polynomial, which is preferable to a simple polynomial interpolation because more parameters can be defined including the amount of smoothing. The smoothing spline function also assumes that there is a (measurement) error in the data that needs to be smoothed locally. There are many versions and modifications of spline interpolators. The most widely used techniques are **thin-plate splines** (Hutchinson, 1995) and **regularized spline with tension and smoothing** (Mitasova and Mitas, 1993).

In the case of regularized spline with tension and smoothing (implemented e.g. in GRASS GIS), the predictions are obtained by (Mitasova et al., 2005):

$$\hat{z}(\mathbf{s}_0) = a_1 + \sum_{i=1}^{n} w_i \cdot R(\upsilon_i) \tag{1.2.8}$$

where the a_1 is a constant and $R(\upsilon_i)$ is the radial basis function determined using (Mitasova and Mitas, 1993):

$$R(\upsilon_i) = -\left[E_1(\upsilon_i) + \ln(\upsilon_i) + C_E\right] \tag{1.2.9}$$

$$\upsilon_i = \left[\varphi \cdot \frac{\mathbf{h_0}}{2}\right]^2 \tag{1.2.10}$$

where $E_1(\upsilon_i)$ is the exponential integral function, C_E=0.577215 is the Euler constant, φ is the generalized tension parameter and $\mathbf{h_0}$ is the distance between the new and interpolation point. The coefficients a_1 and w_i are obtained by solving the system:

$$\sum_{i=1}^{n} w_i = 0 \tag{1.2.11}$$

$$a_1 + \sum_{i=1}^{n} w_i \cdot \left[R(\upsilon_i) + \delta_{ij} \cdot \frac{\varpi_0}{\varpi_i}\right] = z(\mathbf{s}_i); \qquad j = 1, .., n \tag{1.2.12}$$

where ϖ_0/ϖ_i are positive weighting factors representing a smoothing parameter at each given point $\mathbf{s}_i$. The tension parameter φ controls the distance over which the given points influence the resulting surface, while the smoothing parameter controls the vertical deviation of the surface from the points. By using an appropriate combination of tension and smoothing, one can produce a surface which accurately fits the empirical knowledge about the expected variation (Mitasova et al., 2005). Regularized spline with tension and smoothing are, in a way, equivalent to universal kriging (see further §2.1.4) where coordinates are used to explain the deterministic part of variation, and would yield very similar results.

Splines have been widely regarded as highly suitable for interpolation of densely sampled heights and climatic variables (Hutchinson, 1995; Mitas and Mitasova, 1999). However, their biggest criticism is their inability to incorporate larger amounts of auxiliary maps to model the deterministic part of variation. In addition, the smoothing and tension parameters are commonly determined subjectively.

1.3 Statistical spatial prediction models

As mentioned previously, in the case of statistical models, model parameters (coefficients) used to derive outputs are estimated in an objective way following the theory of probability. Unlike mechanical models, in

the case of statistical models, we need to follow several statistical data analysis steps before we can generate maps. This makes the whole mapping process more complicated but it eventually helps us: (a) produce more reliable/objective maps, (b) understand the sources of errors in the data and (c) depict problematic areas/points that need to be revisited.

1.3.1 Kriging

Kriging has for many decades been used as a synonym for geostatistical interpolation. It originated in the mining industry in the early 1950's as a means of improving ore reserve estimation. The original idea came from the mining engineers D. G. Krige and the statistician H. S. Sichel. The technique was first[16] published in Krige (1951), but it took almost a decade until a French mathematician G. Matheron derived the formulas and basically established the whole field of linear geostatistics[17] (Cressie, 1990; Webster and Oliver, 2001). Since then, the same technique has been independently discovered many times, and implemented using various approaches (Venables and Ripley, 2002, pp.425–430).

A standard version of kriging is called **ordinary kriging** (**OK**). Here the predictions are based on the model:

$$Z(\mathbf{s}) = \mu + \varepsilon'(\mathbf{s}) \tag{1.3.1}$$

where μ is the constant *stationary* function (global mean) and $\varepsilon'(\mathbf{s})$ is the spatially correlated stochastic part of variation. The predictions are made as in Eq.(1.2.1):

$$\hat{z}_{\mathrm{OK}}(\mathbf{s}_0) = \sum_{i=1}^{n} w_i(\mathbf{s}_0) \cdot z(\mathbf{s}_i) = \lambda_{\mathbf{0}}^{\mathrm{T}} \cdot \mathbf{z} \tag{1.3.2}$$

where $\lambda_{\mathbf{0}}$ is the vector of kriging weights (w_i), $\mathbf{z}$ is the vector of n observations at primary locations. In a way, kriging can be seen as a sophistication of the inverse distance interpolation. Recall from §1.2.1 that the key problem of inverse distance interpolation is to determine how much importance should be given to each neighbor. Intuitively thinking, there should be a way to estimate the weights in an objective way, so the weights reflect the true spatial autocorrelation structure. The novelty that Matheron (1962) and colleagues introduced to the analysis of point data is the derivation and plotting of the so-called **semivariances** — differences between the neighboring values:

$$\gamma(\mathbf{h}) = \frac{1}{2} E\left[\left(z(\mathbf{s}_i) - z(\mathbf{s}_i + \mathbf{h})\right)^2\right] \tag{1.3.3}$$

where $z(\mathbf{s}_i)$ is the value of a target variable at some sampled location and $z(\mathbf{s}_i + \mathbf{h})$ is the value of the neighbor at distance $\mathbf{s}_i + \mathbf{h}$. Suppose that there are n point observations, this yields $n \cdot (n-1)/2$ pairs for which a semivariance can be calculated. We can then plot all semivariances versus their separation distances, which will produce a variogram cloud as shown in Fig. 1.9b. Such clouds are not easy to describe visually, so the values are commonly averaged for a standard distance called the "*lag*". If we display such averaged data, then we get a standard **experimental or sample variogram** as shown in Fig. 1.9c. What we usually expect to see is that semivariances are smaller at shorter distance and then they stabilize at some distance within the extent of a study area. This can be interpreted as follows: the values of a target variable are more similar at shorter distance, up to a certain distance where the differences between the pairs are more less equal to the global variance[18].

From a meta-physical perspective, spatial auto-correlation in the data can be considered as a result of **diffusion** — a random motion causing a system to decay towards uniform conditions. One can argue that, if there is a physical process behind a feature, one should model it using a deterministic function rather than

[16] A somewhat similar theory was promoted by Gandin (1963) at about the same time.

[17] Matheron (1962) named his theoretical framework the *Theory of Regionalized Variables*. It was basically a theory for modeling stochastic surfaces using spatially sampled variables.

[18] For this reason, many geostatistical packages (e.g. `lsatis`) automatically plot the global variance (horizontal line) directly in a variogram plot.

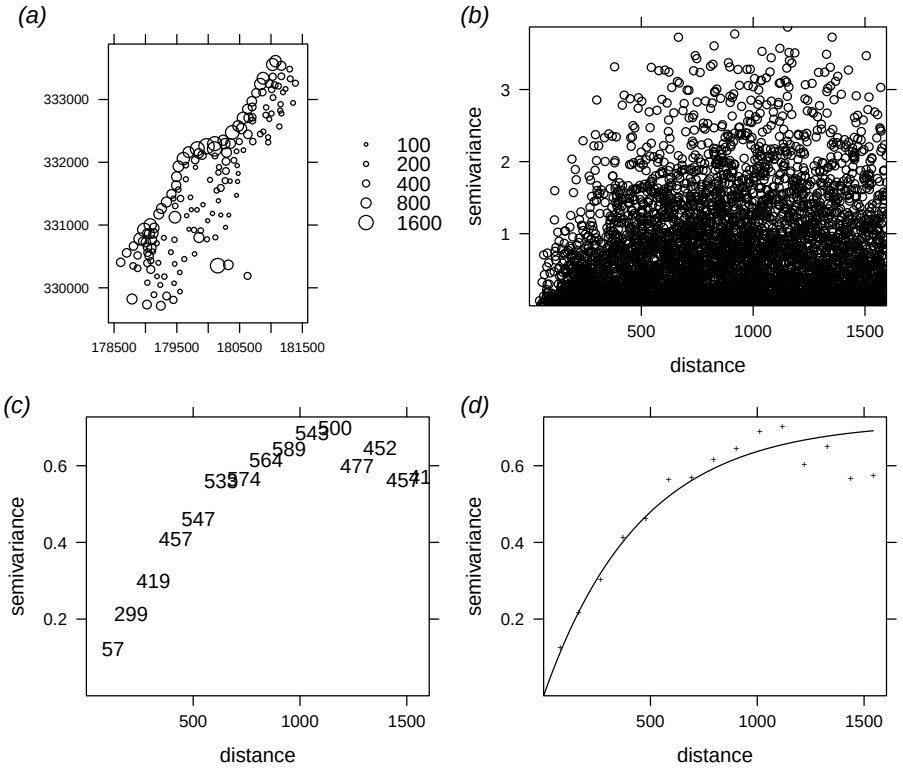

Fig. 1.9: Steps of variogram modeling: (a) sampling locations (155) and measured values of the target variable, (b) variogram cloud showing semivariances for all pairs (log-transformed variable), (c) semivariances aggregated to lags of about 100 m, and (d) the final variogram model fitted using the default settings in `gstat`. See further p.130.

treating it as a stochastic component. Recall from section 1.1.2, diffusion is a random motion so that there is a meta-statistical argument to treat it as a stochastic component.

Once we calculate an experimental variogram, we can fit it using some of the **authorized variogram models**, such as *linear*, *spherical*, *exponential*, *circular*, *Gaussian*, *Bessel*, *power* and similar (Isaaks and Srivastava, 1989; Goovaerts, 1997). The variograms are commonly fitted by iterative reweighted least squares estimation, where the weights are determined based on the number of point pairs or based on the distance. Most commonly, the weights are determined using $N_j/\mathbf{h}_j^2$, where N_j is the number of pairs at a certain lag, and $\mathbf{h}_j$ is the distance (Fig. 1.9d). This means that the algorithm will give much more importance to semivariances with a large number of point pairs and to shorter distances. Fig. 1.9d shows the result of automated variogram fitting given an experimental variogram (Fig. 1.9c) and using the $N_j/\mathbf{h}_j^2$–weights: in this case, we obtained an exponential model with the nugget parameter = 0, sill parameter = 0.714, and the range parameter = 449 m. Note that this is only a sample variogram — if we would go and collect several point samples, each would lead to a somewhat different variogram plot. It is also important to note that there is a difference between the range factor and the range of spatial dependence, also known as the **practical range**. A practical range is the lag $\mathbf{h}$ for which e.g. $\gamma(\mathbf{h}) \cong 0.95\ \gamma(\infty)$, i.e. that is distance at which the semivariance is close to 95% of the sill (Fig. 1.10b).

The target variable is said to be *stationary* if several sample variograms are '*similar*' (if they do not differ statistically), which is referred to as the **covariance stationarity** or second order stationarity. In summary, three important requirements for ordinary kriging are: (1) the trend function is constant (μ =constant); (2) the variogram is constant in the whole area of interest; (3) the target variable follows (approximately) a normal distribution. In practice, these requirements are often not met, which is a serious limitation of ordinary kriging.

Once we have estimated[19] the variogram model, we can use it to derive semivariances at all locations and

[19]We need to determine the parameters of the variogram model: e.g. the nugget (C_0), sill (C_1) and the range (R) parameter. By knowing these parameters, we can estimate semivariances at any location in the area of interest.

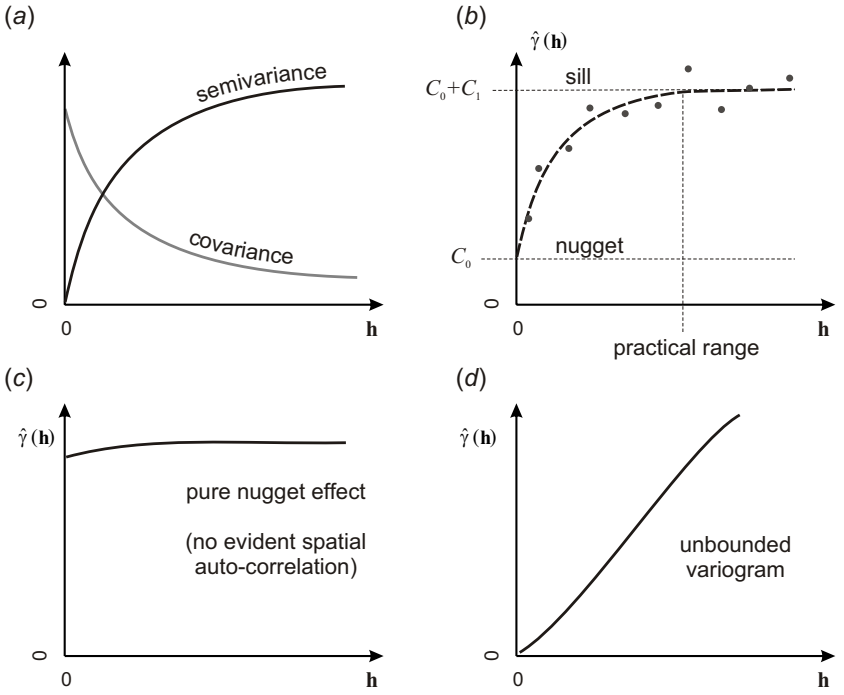

Fig. 1.10: Some basic concepts about variograms: (a) the difference between semivariance and covariance; (b) it is often important in geostatistics to distinguish between the sill variation ($C_0 + C_1$) and the sill parameter (C_1) and between the range parameter (R) and the practical range; (c) a variogram that shows no spatial correlation can be defined by a single parameter (C_0); (d) an unbounded variogram.

solve the kriging weights. The kriging OK weights are solved by multiplying the covariances:

$$\lambda_{\mathbf{0}} = \mathbf{C}^{-1} \cdot \mathbf{c_0}; \qquad C(|\mathbf{h}| = 0) = C_0 + C_1 \tag{1.3.4}$$

where $\mathbf{C}$ is the covariance matrix derived for $n \times n$ observations and $\mathbf{c_0}$ is the vector of covariances at a new location. Note that the $\mathbf{C}$ is in fact $(n+1) \times (n+1)$ matrix if it is used to derive kriging weights. One extra row and column are used to ensure that the sum of weights is equal to one:

$$\begin{bmatrix} C(\mathbf{s}_1,\mathbf{s}_1) & \cdots & C(\mathbf{s}_1,\mathbf{s}_n) & 1 \\ \vdots & & \vdots & \vdots \\ C(\mathbf{s}_n,\mathbf{s}_1) & \cdots & C(\mathbf{s}_n,\mathbf{s}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} C(\mathbf{s}_0,\mathbf{s}_1) \\ \vdots \\ C(\mathbf{s}_0,\mathbf{s}_n) \\ 1 \end{bmatrix} = \begin{bmatrix} w_1(\mathbf{s}_0) \\ \vdots \\ w_n(\mathbf{s}_0) \\ \varphi \end{bmatrix} \tag{1.3.5}$$

where φ is the so-called *Lagrange multiplier*.

In addition to estimation of values at new locations, a statistical spatial prediction technique produces a measure of associated uncertainty of making predictions by using a given model. In geostatistics, this is often referred to as the **prediction variance**, i.e. the estimated variance of the prediction error. OK variance is defined as the weighted average of covariances from the new point ($\mathbf{s}_0$) to all calibration points ($\mathbf{s}_1, .., \mathbf{s}_n$), plus the Lagrange multiplier (Webster and Oliver, 2001, p.183):

$$\hat{\sigma}^2_{\mathrm{OK}}(\mathbf{s}_0) = (C_0 + C_1) - \mathbf{c_0^T} \cdot \lambda_{\mathbf{0}} = C_0 + C_1 - \sum_{i=1}^{n} w_i(\mathbf{s}_0) \cdot C(\mathbf{s}_0, \mathbf{s}_i) + \varphi \tag{1.3.6}$$

where $C(\mathbf{s}_0, \mathbf{s}_i)$ is the covariance between the new location and the sampled point pair, and φ is the Lagrange multiplier, as shown in Eq.(1.3.5).

Outputs from any statistical prediction model are commonly two maps: (1) predictions and (2) prediction variance. The mean of the prediction variance at all locations can be termed the **overall prediction variance**, and can be used as a measure of the overall precision of the final map: if the overall prediction variance gets close to the global variance, then the map is 100% imprecise; if the overall prediction variance tends to zero, then the map is 100% precise[20] (see further Fig. 5.19).

Note that a common practice in geostatistics is to model the variogram using a semivariance function and then, for reasons of computational efficiency, use the **covariances**. In the case of solving the kriging weights, both the matrix of semivariances and covariances give the same results, so you should not really make a difference between the two. The relation between the covariances and semivariances is (Isaaks and Srivastava, 1989, p.289):

$$C(\mathbf{h}) = C_0 + C_1 - \gamma(\mathbf{h}) \tag{1.3.7}$$

where $C(\mathbf{h})$ is the covariance, and $\gamma(\mathbf{h})$ is the semivariance function (Fig. 1.10a). So for example, an exponential model can be written in two ways:

$$\gamma(\mathbf{h}) = \begin{cases} 0 & \text{if} \quad |\mathbf{h}| = 0 \\ C_0 + C_1 \cdot \left[1 - e^{-\left(\frac{\mathbf{h}}{R}\right)}\right] & \text{if} \quad |\mathbf{h}| > 0 \end{cases} \tag{1.3.8}$$

$$C(\mathbf{h}) = \begin{cases} C_0 + C_1 & \text{if} \quad |\mathbf{h}| = 0 \\ C_1 \cdot \left[e^{-\left(\frac{\mathbf{h}}{R}\right)}\right] & \text{if} \quad |\mathbf{h}| > 0 \end{cases} \tag{1.3.9}$$

The covariance at zero distance ($C(0)$) is by definition equal to the mean residual error (Cressie, 1993) — $C(\mathbf{h}_{11})$ also written as $C(\mathbf{s}_1, \mathbf{s}_1)$, and which is equal to $C(0) = C_0 + C_1 = Var\{z(s)\}$.

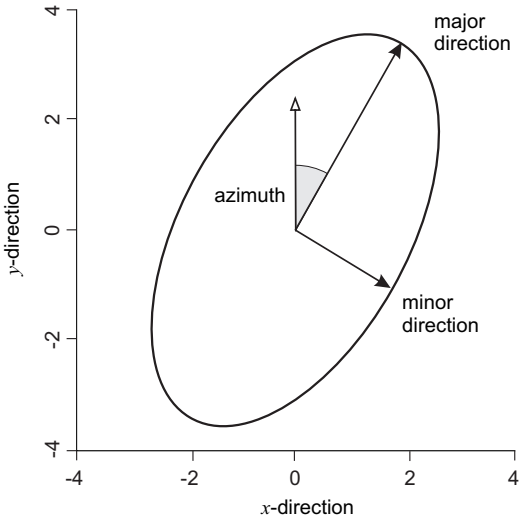

Fig. 1.11: Range ellipse for anisotropic model. After gstat User's manual.

[20] As we will see later on, the precision of mapping is only a measure of how well did we fit the point values. The true quality of map can only be accessed by using validation points, preferably independent from the point data set used to make predictions.

The variogram models can be extended to even larger number of parameters if either (a) **anisotropy** or (b) smoothness are considered in addition to modeling of nugget and sill variation. The 2D geometric anisotropy in gstat[21], for example, is modeled by replacing the range parameter with three parameters — range in the major direction (direction of the strongest correlation), angle of the principal direction and the anisotropy ratio, e.g. (Fig. 1.11):

```
> vgm(nugget=1, model="Sph", sill=10, range=2, anis=c(30,0.5))
```

where the value of the angle of major direction is `30` (azimuthal direction measured in degrees clockwise), and the value of the anisotropy ratio is `0.5` (range in minor direction is two times shorter). There is no universal rule whether to use always anisotropic models or to use them only if the variogram shows significant anisotropy. As a rule of thumb, we can say that, if the variogram confidence bands (see further Fig. 5.15) in the two orthogonal directions (major and minor direction) show <50% overlap, than one needs to consider using anisotropic models.

Another sophistication of the standard 3–parameter variograms is the Matérn variogram model, which has an additional parameter to describe the smoothness (Stein, 1999; Minasny and McBratney, 2005):

$$\gamma(\mathbf{h}) = C_0 \cdot \delta(\mathbf{h}) + C_1 \cdot \left[\frac{1}{2^{\nu-1} \cdot \Gamma(\nu)} \cdot \left(\frac{\mathbf{h}}{R} \right)^{\nu} \cdot K_{\nu} \cdot \left(\frac{\mathbf{h}}{R} \right) \right] \tag{1.3.10}$$

where $\delta(\mathbf{h})$ is the Kronecker delta, K_ν is the modified Bessel function, Γ is the gamma function and ν is the smoothness parameter. The advantage of this model is that it can be used universally to model both short and long distance variation (see further section 10.3.2). In reality, variogram models with more parameters are more difficult to fit automatically because the iterative algorithms might get stuck in local minima (Minasny and McBratney, 2005).

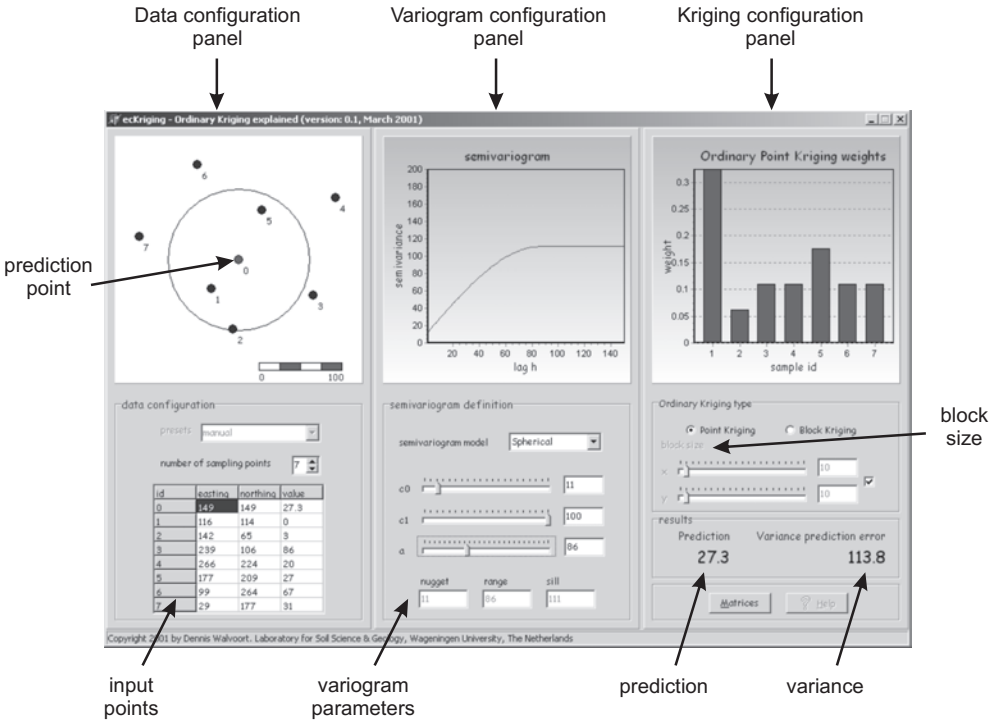

Fig. 1.12: Ordinary kriging explained: EZ-Kriging. Courtesy of Dennis J.J. Walvoort, Wageningen University.

The fastest intuitive way to understand the principles of kriging is to use an educational program called **EZ-Kriging**, kindly provided by Dennis J.J. Walvoort from the Alterra Research institute. The GUI of EZ-

[21] http://www.gstat.org/manual/node20.html

Kriging consists of three panels: (1) data configuration panel, (2) variogram panel, and (3) kriging panel (Fig. 1.12). This allows you to zoom into ordinary kriging and explore its main characterizes and behavior: how do weights change for different variogram models, how do data values affect the weights, how does block size affect the kriging results etc. For example, if you study how model shape, nugget, sill and range affect the kriging results, you will notice that, assuming some standard variogram model (zero nugget, sill at global variance and practical range at 10% of the largest distance), the weights will decrease exponentially[22]. This is an important characteristic of kriging because it allows us to limit the search window to speed up the calculation and put more emphasize on fitting the semivariances at shorter distances. Note also that, although it commonly leads to smoothing of the values, kriging is an exact and non-convex interpolator. It is exact because the kriging estimates are equal to input values at sampling locations, and it is non-convex because its predictions can be outside the data range, e.g. we can produce negative concentrations.

Another important aspect of using kriging is the issue of the support size. In geostatistics, one can control the support size of the outputs by averaging multiple (randomized) point predictions over regular blocks of land. This is known as **block prediction** (Heuvelink and Pebesma, 1999). A problem is that we can sample elevations at point locations, and then interpolate them for blocks of e.g. 10×10 m, but we could also take composite samples and interpolate them at point locations. This often confuses GIS users because as well as using point measurements to interpolate values at regular point locations (e.g. by point kriging), and then display them using a raster map (see Fig. 1.8), we can also make spatial predictions for blocks of land (block kriging) and display them using the same raster model (Bishop and McBratney, 2001). For simplicity, in the case of block-kriging, one should always try to use a cell size that corresponds to the support size.

1.3.2 Environmental correlation

If some exhaustively-sampled explanatory variables or **covariates** are available in the area of interest and if they are significantly correlated with our target variable (spatial cross-correlation), and assuming that the point-values are not spatially auto-correlated, predictions can be obtained by focusing only on the deterministic part of variation:

$$Z(\mathbf{s}) = f\left\{q_k(\mathbf{s})\right\} + \varepsilon \tag{1.3.11}$$

where q_k are the auxiliary predictors. This approach to spatial prediction has a strong physical interpretation. Consider Rowe and Barnes (1994) observation that earth surface energy-moisture regimes at all scales/sizes are the dynamic driving variables of functional ecosystems at all scales/sizes. The concept of vegetation/soil-environment relationships has frequently been presented in terms of an equation with six key **environmental factors** as:

$$V \times S[x,y,\tilde{t}] = f\left\{\begin{array}{l} s[x,y,\tilde{t}]\; c[x,y,\tilde{t}]\; o[x,y,\tilde{t}] \\ r[x,y,\tilde{t}]\; p[x,y,\tilde{t}]\; a[x,y,\tilde{t}] \end{array}\right. \tag{1.3.12}$$

where V stands for vegetation, S for soil, c stands for climate, o for organisms (including humans), r is relief, p is parent material or geology, a is age of the system, x, y are the coordinates and t is time dimension. This means that the predictors which are available over entire areas of interest can be used to predict the value of an environmental variable at unvisited locations — first by modeling the relationship between the target and explanatory environmental predictors at sample locations, and then by applying it to unvisited locations using the known value of the explanatory variables at those locations. Common explanatory environmental predictors used to map environmental variables are land surface parameters, remotely sensed images, and geological, soil and land-use maps (McKenzie and Ryan, 1999). Because many auxiliary predictors (see further section 4) are now also available at low or no cost, this approach to spatial prediction is ever more important (Pebesma, 2006; Hengl et al., 2007a).

Functional relations between environmental variables and factors are in general unknown and the correlation coefficients can differ for different study areas, different seasons and different scales. However, in

[22] In practice, often >95% of weights will be explained by the nearest 30–50 points. Only if the variogram is close to the pure nugget model, the more distant points will receive more importance, but then the technique will produce poor predictions anyhow.

many cases, relations with environmental predictors often reflect causal linkage: deeper and more developed soils occur at places of higher potential accumulation and lower slope; different type of forests can be found at different slope expositions and elevations; soils with more organic matter can be found where the climate is cooler and wetter etc. This makes this technique especially suitable for natural resource inventory teams because it allows them to validate their empirical knowledge about the variation of the target features in the area of interest.

There are (at least) four groups of statistical models that have been used to make spatial predictions with the help of environmental factors (Chambers and Hastie, 1992; McBratney et al., 2003; Bishop and Minasny, 2005):

Classification-based models — Classification models are primarily developed and used when we are dealing with discrete target variables (e.g. land cover or soil types). There is also a difference whether **Boolean** (crisp) or **Fuzzy** (continuous) classification rules are used to create outputs. Outputs from the model fitting process are class boundaries (class centres and standard deviations) or classification rules.

Tree-based models — Tree-based models (classification or regression trees) are often easier to interpret when a mix of continuous and discrete variables are used as predictors (Chambers and Hastie, 1992). They are fitted by successively splitting a data set into increasingly homogeneous groupings. Output from the model fitting process is a **decision tree**, which can then be applied to make predictions of either individual property values or class types for an entire area of interest.

Regression models — Regression analysis employs a family of functions called **Generalized Linear Models** (GLMs), which all assume a linear relationship between the inputs and outputs (Neter et al., 1996). Output from the model fitting process is a set of regression coefficients. Regression models can be also used to represent non-linear relationships with the use of **General Additive Models** (GAMs). The relationship between the predictors and targets can be solved using one-step data-fitting or by using iterative data fitting techniques (neural networks and similar).

Each of the models listed above can be equally applicable for mapping of environmental variables and each can exhibit advantages and disadvantages. For example, some advantages of using tree-based regression are that they: (1) can handle missing values; (2) can use continuous and categorical predictors; (3) are robust to predictor specification; and (4) make very limited assumptions about the form of the regression model (Henderson et al., 2004). Some disadvantages of regression trees, on the other hand, are that they require large data sets and completely ignore spatial position of the input points.

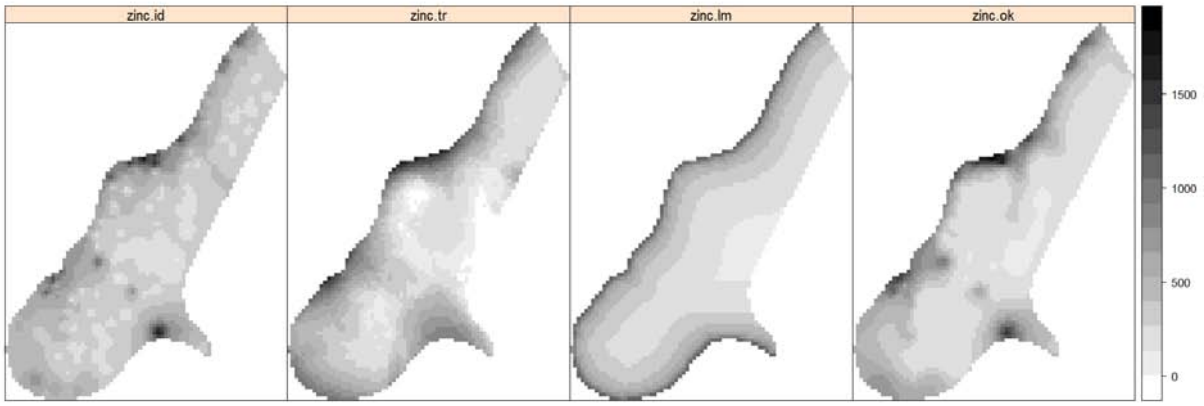

Fig. 1.13: Comparison of spatial prediction techniques for mapping `Zinc` (sampling locations are shown in Fig. 1.9). Note that inverse distance interpolation (`.id`) and kriging (`.ok`) are often quite similar; the moving trend surface (`.tr`; 2nd order polynomial) can lead to artifacts (negative values) — locally where the density of points is poor. The regression-based (`.lm`) predictions were produced using distance from the river as explanatory variable (see further §5).

A common regression-based approach to spatial prediction is **multiple linear regression** (Draper and Smith, 1998; Kutner et al., 2004). Here, the predictions are again obtained by weighted averaging (compare

with Eq.(1.3.2)), this time by averaging the predictors:

$$\hat{z}_{\text{OLS}}(\mathbf{s}_0) = \hat{b}_0 + \hat{b}_1 \cdot q_1(\mathbf{s}_0) + \ldots + \hat{b}_p \cdot q_p(\mathbf{s}_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(\mathbf{s}_0); \qquad q_0(\mathbf{s}_0) \equiv 1 \tag{1.3.13}$$

or in matrix algebra:

$$\hat{z}_{\text{OLS}}(\mathbf{s}_0) = \hat{\beta}^{\mathbf{T}} \cdot \mathbf{q} \tag{1.3.14}$$

where $q_k(\mathbf{s}_0)$ are the values of the explanatory variables at the target location, p is the number of predictors or explanatory variables[23], and $\hat{\beta}_k$ are the regression coefficients solved using the **Ordinary Least Squares**:

$$\hat{\beta} = \left(\mathbf{q}^{\mathbf{T}} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathbf{T}} \cdot \mathbf{z} \tag{1.3.15}$$

where $\mathbf{q}$ is the matrix of predictors ($n \times p + 1$) and $\mathbf{z}$ is the vector of sampled observations. The prediction error of a multiple linear regression model is (Neter et al., 1996, p.210):

$$\hat{\sigma}^2_{\text{OLS}}(\mathbf{s}_0) = MSE \cdot \left[1 + \mathbf{q}_0^{\mathbf{T}} \cdot \left(\mathbf{q}^{\mathbf{T}} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}_0\right] \tag{1.3.16}$$

where *MSE* is the mean square (residual) error around the regression line:

$$MSE = \frac{\sum_{i=1}^{n} \left[z(\mathbf{s}_i) - \hat{z}(\mathbf{s}_i)\right]^2}{n-2} \tag{1.3.17}$$

and $\mathbf{q}_0$ is the vector of predictors at new, unvisited location. In the univariate case, the variance of the prediction error can also be derived using:

$$\hat{\sigma}^2(\mathbf{s}_0) = MSE \cdot \left[1 + \frac{1}{n} + \frac{\left[q(\mathbf{s}_0) - \bar{q}\right]^2}{\sum_{i=1}^{n} \left[q(\mathbf{s}_i) - \bar{q}\right]^2}\right] = MSE \cdot \left[1 + v(\mathbf{s}_0)\right] \tag{1.3.18}$$

where v is the curvature of the confidence band around the regression line. This reflects the amount of extrapolation in the feature space (Ott and Longnecker, 2001, p.570). It can be seen from Eq. (1.3.18) that the prediction error, for a given sampling intensity ($n/\mathbb{A}$), depends on three factors:

(1.) Mean square residual error (*MSE*);

(2.) Spreading of points in the feature space $\sum \left[q(\mathbf{s}_i) - \bar{q}\right]^2$;

(3.) '*Distance*' of the new observation from the centre of the feature space $\left[q(\mathbf{s}_0) - \bar{q}\right]$.

So in general, if the model is linear, we can decrease the prediction variance if we increase the spreading of the points in feature space. Understanding this principles allows us to prepare sampling plans that will achieve higher mapping precision and minimize extrapolation in feature space (see further §2.8).

[23] To avoid confusion with geographical coordinates, we use the symbol q, instead of the more common x, to denote a predictor.

The sum of squares of residuals (*SSE*) can be used to determine the **adjusted coefficient of multiple determination** (R^2_a), which describes the goodness of fit:

$$\begin{aligned} R^2_a &= 1 - \left(\frac{n-1}{n-p}\right) \cdot \frac{SSE}{SSTO} \\ &= 1 - \left(\frac{n-1}{n-p}\right) \cdot \left(1 - R^2\right) \end{aligned} \tag{1.3.19}$$

where *SSTO* is the total sum of squares (Neter et al., 1996), R^2 indicates amount of variance explained by the model, whereas R^2_a adjusts for the number of variables (p) used. For many environmental mapping projects, a $R^2_a \geq 0.85$ is already a very satisfactory solution and higher values will typically only mean over-fitting of the data (Park and Vlek, 2002).

The principle of predicting environmental variables using factors of climate, relief, geology and similar, is often referred to as **environmental correlation**. The *environmental correlation approach* to mapping is a true alternative to ordinary kriging (compare differences in generated patterns in Fig. 1.13). This is because both approaches deal with different aspects of spatial variation: regression deals with the deterministic and kriging with the spatially-correlated stochastic part of variation.

The biggest criticism of the pure regression approach to spatial prediction is that the position of points in geographical space is completely ignored, both during model fitting and prediction. Imagine if we are dealing with two point data sets where one data set is heavily clustered, while the other is well-spread over the area of interest — a sophistication of simple non-spatial regression is needed to account for the clustering of the points so that the model derived using the clustered points takes this property into account.

One way to account for this problem is to take the distance between the points into account during the estimation of the regression coefficients. This can be achieved by using the **geographically weighted regression** (Fotheringham et al., 2002). So instead of using the OLS estimation (Eq.1.3.15) we use:

$$\hat{\beta}_{\mathtt{WLS}} = \left(\mathbf{q}^{\mathrm{T}} \cdot \mathbf{W} \cdot \mathbf{q}\right)^{-1} \cdot \mathbf{q}^{\mathrm{T}} \cdot \mathbf{W} \cdot \mathbf{z} \tag{1.3.20}$$

where **W** is a matrix of weights, determined using some distance decay function e.g.:

$$w_i(\mathbf{s}_i, \mathbf{s}_j) = \sigma^2_E \cdot \exp\left[-3 \cdot \frac{d^2(\mathbf{s}_i, \mathbf{s}_j)}{\urcorner^2}\right] \tag{1.3.21}$$

where σ^2_E is the level of variation of the error terms, $d(\mathbf{s}_i, \mathbf{s}_j)$ is the Euclidian distance between a sampled point pair and $\urcorner$ is known as the bandwidth, which determines the degree of *locality* — small values of $\urcorner$ suggest that correlation only occurs between very close point pairs and large values suggest that such effects exist even on a larger spatial scale. Compare further with Eq.(2.1.3). The problem remains to select a search radius (Eq.1.3.21) using objective criteria. As we will see further (§2.2), geographically weighted regression can be compared with regression-kriging with a moving window where variograms and regression models are estimated locally.

The main benefit of geographically weighted regression (**GWR**) is that this method enables researchers to study local differences in responses to input variables. It therefore more often focuses on coefficients' explanation than on interpolation of the endogenous variables. By setting up the search radius (bandwidth) one can investigate the impact of spatial proximity between the samples on the regression parameters. By fitting the regression models using a moving window algorithm, one can also produce maps of regression coefficients and analyze how much the regression model is dependent on the location. However, the coefficient maps generated by GWR are usually too smooth to be true. According to Wheeler and Tiefelsdorf (2005) and Griffith (2008), the two main problems with GWR are: (1) strong multicollinearity effects among coefficients make the results even totally wrong, and (2) lose degrees of freedom in the regression model. Hence, spatial hierarchical model under bayesian framework (Gelfand et al., 2003) and spatial filtering model (Griffith, 2008) may be better structures for such analyzes than GWR.

1.3.3 Predicting from polygon maps

A special case of environmental correlation is prediction from polygon maps i.e. stratified areas (different land use/cover types, geological units etc). Assuming that the residuals show no spatial auto-correlation, a value at a new location can be predicted by:

$$\hat{z}(\mathbf{s}_0) = \sum_{i=1}^{n} w_i \cdot z(s); \qquad w_i = \begin{cases} 1/n_k & \text{for } x_i \in k \\ 0 & \text{otherwise} \end{cases} \tag{1.3.22}$$

where k is the unit identifier. This means that the weights within some unit will be equal so that the predictions are made by simple averaging per unit (Webster and Oliver, 2001):

$$\hat{z}(\mathbf{s}_0) = \bar{\mu}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} z(\mathbf{s}_i) \tag{1.3.23}$$

Consequently, the output map will show only abrupt changes in the values between the units. The prediction variance of this prediction model is simply the within-unit variance:

$$\hat{\sigma}^2(\mathbf{s}_0) = \frac{\sigma_k^2}{n_k} \tag{1.3.24}$$

From Eq.(1.3.24) it is obvious that the precision of the technique will be maximized if the within-unit variation is infinitely small. Likewise, if the within-unit variation is as high as the global variability, the predictions will be as bad as predicting by taking any value from the normal distribution.

Another approach to make predictions from polygon maps is to use multiple regression. In this case, the predictors (mapping units) are used as indicators:

$$\hat{z}(\mathbf{s}_0) = \hat{b}_1 \cdot MU_1(\mathbf{s}_0) + \ldots + \hat{b}_k \cdot MU_k(\mathbf{s}_0); \qquad MU_k \in [0|1] \tag{1.3.25}$$

and it can be shown that the OLS fitted regression coefficients will equal the mean values within each strata ($b_k = \bar{\mu}(MU_k)$), so that the Eqs.(1.3.25) and (1.3.23) are in fact equivalent.

If, on the other hand, the residuals do show spatial auto-correlation, the predictions can be obtained by **stratified kriging**. This is basically ordinary kriging done separately for each strata and can often be impractical because we need to estimate a variogram for each of the k strata (Boucneau et al., 1998). Note that the strata or sub-areas need to be known *a priori* and they should never be derived from the data used to generate spatial predictions.

1.3.4 Hybrid models

Hybrid spatial prediction models comprise of a combination of the techniques listed previously. For example, a hybrid geostatistical model employs both correlation with auxiliary predictors and spatial autocorrelation simultaneously. There are two main sub-groups of hybrid geostatistical models (McBratney et al., 2003): (a) **co-kriging**-based and (b) **regression-kriging**-based techniques, but the list could be extended.

Note also that, in the case of environmental correlation by linear regression, we assume some basic (additive) model, although the relationship can be much more complex. To account for this, a linear regression model can be extended to a diversity of statistical models ranging from regression trees, to General Additive Models and similar. Consequently, the hybrid models are more generic than pure kriging-based or regression-based techniques and can be used to represent both discrete and continuous changes in the space, both deterministic and stochastic processes.

One can also combine deterministic, statistical and expert-based estimation models. For example, one can use a deterministic model to estimate a value of the variable, then use actual measurements to fit a calibration model, analyze the residuals for spatial correlation and eventually combine the statistical fitting

and deterministic modeling (Hengl et al., 2007a). Most often, expert-based models are supplemented with the actual measurements, which are then used to refine the rules, e.g. using neural networks (Kanevski et al., 1997).

1.4 Validation of spatial prediction models

OK or OLS variance (Eqs.1.3.6; and 1.3.18) is the statistical estimate of the model uncertainty. Note that the '*true*' prediction power can only be assessed by using an independent (control) data set. The prediction error is therefore often referred to as the *precision of prediction*. The true quality of a map can be best assessed by comparing estimated values ($\hat{z}(\mathbf{s}_j)$) with actual observations at validation points ($z^*(\mathbf{s}_j)$). Commonly, two measures are most relevant here — (1) the mean prediction error (*ME*):

$$ME = \frac{1}{l} \cdot \sum_{j=1}^{l} \left[\hat{z}(\mathbf{s}_j) - z^*(\mathbf{s}_j)\right]; \qquad E\{ME\} = 0 \tag{1.4.1}$$

and (2) the root mean square prediction error (*RMSE*):

$$RMSE = \sqrt{\frac{1}{l} \cdot \sum_{j=1}^{l} \left[\hat{z}(\mathbf{s}_j) - z^*(\mathbf{s}_j)\right]^2}; \qquad E\{RMSE\} = \sigma(\mathbf{h} = 0) \tag{1.4.2}$$

where l is the number of validation points. We can also standardize the errors based on the prediction variance estimated by the spatial prediction model:

$$RMNSE = \sqrt{\frac{1}{l} \cdot \sum_{j=1}^{l} \left[\frac{\hat{z}(\mathbf{s}_j) - z^*(\mathbf{s}_j)}{\hat{\sigma}_j}\right]^2}; \qquad E\{RMNSE\} = 1 \tag{1.4.3}$$

In order to compare accuracy of prediction between variables of different types, the *RMSE* can also be normalized by the total variation:

$$RMSE_r = \frac{RMSE}{s_z} \tag{1.4.4}$$

which will show how much of the global variation budget has been explained by the model. As a rule of thumb, a value of $RMSE_r$ that is close to 40% means a fairly satisfactory accuracy of prediction (R-square=85%). Otherwise, if $RMSE_r$ >71%, this means that the model accounted for less than 50% of variability at the validation points. Note also that *ME*, *RMSE* and *RMNSE* estimated at validation points are also only a sample from a population of values — if the validation points are poorly sampled, our estimate of the map quality may be equally poor.

Because collecting additional (independent) samples is often impractical and expensive, validation of prediction models is most commonly done by using **cross-validation** i.e. by subsetting the original point set in two data set — calibration and validation — and then repeating the analysis. There are several types of cross-validation methods (Bivand et al., 2008, pp.221–226):

- the k–fold cross-validation — the original sample is split into k equal parts and then each is used for cross-validation;
- *leave-one-out* cross-validation (LOO) — each sampling point is used for cross-validation;
- *Jackknifing* — similar to LOO, but aims at estimating the bias of statistical analysis and not of predictions;

Both k–fold and the *leave-one-out* cross validation are implemented in the `krige.cv` method of `gstat` package. The LOO algorithm works as follows: it visits a data point, predicts the value at that location by kriging *without* using the observed value, and proceeds with the next data point. This way each individual

point is assessed versus the whole data set. The results of cross-validation can be visualised to pinpoint the most problematic points, e.g. exceeding three standard deviations of the normalized prediction error, and to derive a summary estimate of the map accuracy. In the case of many outliers and blunders in the input data, the LOO cross-validation might produce strange outputs. Hence many authors recommend 10–fold cross-validation as the most robust approach. Note also that cross-validation is not necessarily independent — points used for cross-validation are subset of the original sampling design, hence if the original design is biased and/or non-representative, then also the cross-validation might not reveal the true accuracy of a technique. However, if the sampling design has been generated using e.g. random sampling, it can be shown that also randomly taken subsets will be unbiased estimators of the true accuracy.

To assess the accuracy of predicting categorical variables we can use the **kappa statistics**, which is a common measure of classification accuracy (Congalton and Green, 1999; Foody, 2004). Kappa statistics measures the difference between the actual agreement between the predictions and ground truth and the agreement that could be expected by chance (see further p.135). In most remote sensing-based mapping projects, a kappa larger than 85% is considered to be a satisfactory result (Foody, 2004). The kappa is only a measure of the overall mapping accuracy. Specific classes can be analyzed by examining the percentage of correctly classified pixels per each class:

$$P_c = \frac{\sum_{j=1}^{m} \left(\hat{C}(s_j) = C(s_j)\right)}{m} \tag{1.4.5}$$

where P_c is the percentage of correctly classified pixels, $\hat{C}(s_j)$ is the estimated class at validation locations (s_j) and m is total number of observations of class c at validation points.

Further reading:

★ Cressie, N.A.C., 1993. **Statistics for Spatial Data**, revised edition. John Wiley & Sons, New York, 416 p.

★ Goovaerts, P., 1997. **Geostatistics for Natural Resources Evaluation** (Applied Geostatistics). Oxford University Press, New York, 496 p.

★ Isaaks, E.H. and Srivastava, R.M. 1989. **An Introduction to Applied Geostatistics**. Oxford University Press, New York, 542 p.

★ Webster, R. and Oliver, M.A., 2001. **Geostatistics for Environmental Scientists**. Statistics in Practice. John Wiley & Sons, Chichester, 265 p.

★ `http://www.wiley.co.uk/eoenv/` — The Encyclopedia of Environmetrics.

★ `http://geoenvia.org` — A research association that promotes use of geostatistical methods for environmental applications.

★ `http://www.iamg.org` — International Association of Mathematical Geosciences.