1

From classical statistics to geostatistics

1.1 Not all spatial data are geostatistical data

As pointed out in Schabenberger and Gotway (2005, p. 6), because spatial data arise in a myriad of fields and applications, there is also a myriad of spatial data types, structures and scenarios. Thus, an exhaustive classification of spatial data would be a very difficult challenge and this is why we have opted for embracing the general, simple and useful classification of spatial data provided by Cressie (1993, pp. 8–13). Cressie's classification of spatial data is based on the nature of the spatial domain under study. Depending on this, we can have: geostatistical data, lattice data and point patterns.

Following Cressie (1993), let $\mathbf{s} \in \mathbb{R}^d$ be a generic location in a d-dimensional Euclidean space and $\{Z(\mathbf{s})\colon \mathbf{s} \in \mathbb{R}^d\}$ be a spatial random function, Z denoting the attribute we are interested in.

Geostatistical data arise when the domain under study is a fixed set D that is continuous. That is: (i) $Z(\mathbf{s})$ can be observed at any point of the domain (continuous); and (ii) the points in D are non-stochastic (fixed, D is the same for all the realizations of the spatial random function). From (i) it can be easily seen that geostatistical data are identified with spatial data with a continuous variation (the spatial process is indexed over a continuous space).

Some examples of geostatistical data are the level of a pollutant in a city, the precipitation or air temperature values in a country, the concentrations of heavy metals in the top soil of a region, etc. It is obvious that, at least theoretically, the level of a specific pollutant could be measured at any location of the city; the same can be said for measurements of precipitations or air temperatures across a

Spatial and Spatio-Temporal Geostatistical Modeling and Kriging, First Edition.
José-María Montero, Gema Fernández-Avilés, and Jorge Mateu.

Companion Website: www.wiley.com/go/montero/spatial

country or concentrations of a heavy metal across a region. However, in practice, an exhaustive observation of the spatial process is not possible. Usually, the spatial process is observed at a set of locations (for example, the level of a specific pollutant in a city is observed at the points where the monitoring stations are located) and, based on such observed values, geostatistical analysis reproduces the behavior of the spatial process across the entire domain of interest. Sometimes the goal is not so ambitious and the aim is the prediction at one or some few non-observed points or the estimation of an average value over small areas, or over the whole area under study. In geostatistical analysis the most important thing is to quantify the spatial correlation between observations (through the basic tool in geostatistics, the semivariogram) and use this information to achieve the above goals.

Figure 1.1 depicts the locations where the main pollutants are measured in Madrid, Spain (the location of the monitoring stations), along with the mapping of the level of nitrogen oxide (NOx) for the whole city (average of the NOx levels at 10 pm in the week days of the 50th week of 2008).

The fact that the attribute of interest is continuous or discrete has nothing to do with the data being geostatistical or not. Also, how observation points are selected (according to our convenience, using a monitoring network, using a probabilistic sampling scheme ...) has nothing to do with the data being geostatistical or not.

Lattice data arise when: (i) the domain under study D is discrete, that is, $Z(\mathbf{s})$ can be observed in a number of fixed locations that can be enumerated. These locations can be points or regions, but they are usually ZIP codes, census tracks, neighborhoods, provinces, etc., and the data in most of cases are spatially aggregated data over these areal regions. Although these regions can be regularly shaped, usually the shape they exhibit is irregular, and this, together with the spatially aggregated character of the data, is why lattice data are also called regional data. And (ii) the locations in D are

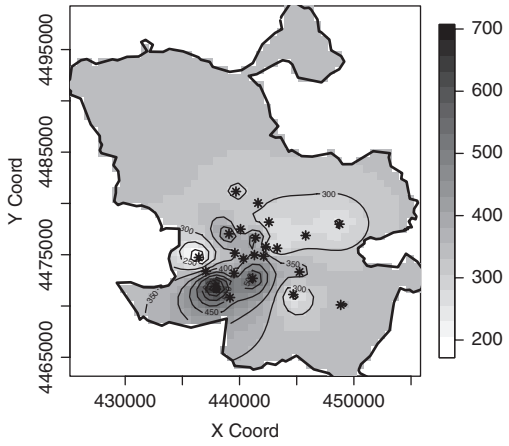

Figure 1.1 Location of the pollution monitoring stations in Madrid and map of predicted NOx levels (10 pm; average of the week days; 50*th week of* 2008*) using geostatistical techniques.*

non-stochastic. Of course, a core concept in lattice data analysis is the neighborhood. Some examples of lattice data include the unemployment rate by states, crime data by counties, agricultural yields in plots, average housing prices by provinces, etc. Unlike geostatistical data, lattice data can be exhaustively observed and in this case prediction makes no sense. However, smoothing and clustering acquire special importance when dealing with this type of spatial data. Similar to geostatistical data, the response measured can be discrete or continuous, and this has nothing to do with the data being lattice data or not.

Figure 1.2 depicts the percentage of households with problems of pollution and odors in each census tract of Madrid, Spain, in 2001. As can be observed, the attribute under study is aggregated over each census tract, the domain is the set of the 128 census tracts (discrete) and sites in the domain (the census tracts) are fixed (non-stochastic).

While in both geostatistical and lattice data the domain D is fixed, in point pattern data it is discrete or continuous, but random. Point patterns arise when the attribute under study is the location of events (observations). That is, the interest lies in where events of interest occur. Some examples of point patterns are the location of fires in Castilla-La Mancha, a Spanish region (see Figure 1.3), the location of trees in a forest or the location of nests in a breeding colony of birds, among many others. In these cases, it is obvious that D is random and the observation points do not depend on the

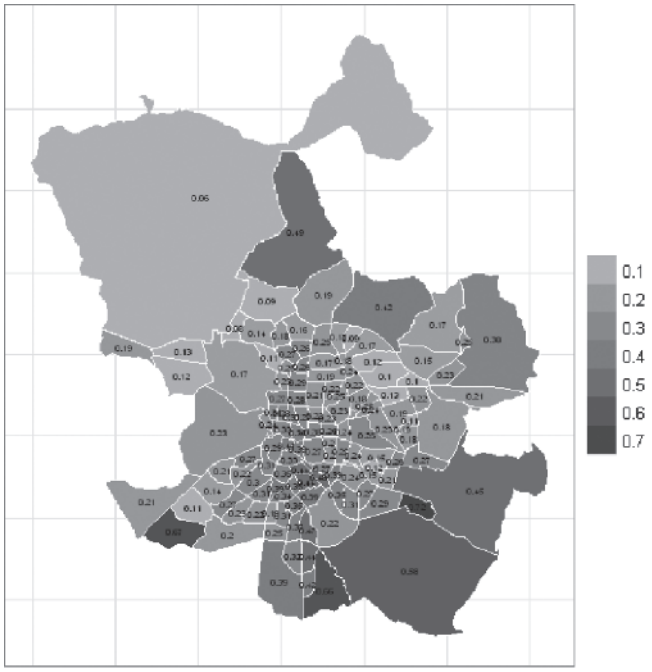

Figure 1.2 Percentage of households with problems of pollution and odors in Madrid, Spain, 2001 *(census tracts).*

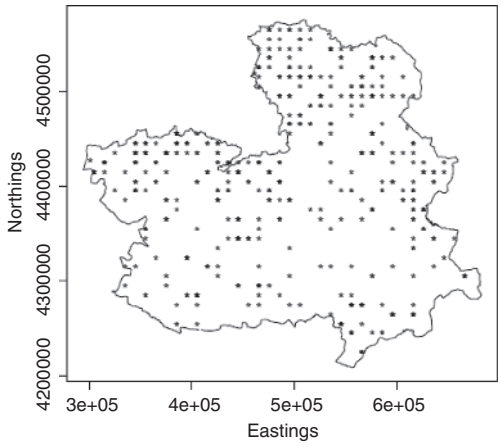

Figure 1.3 Fires in Castilla-La Mancha, Spain, 1998.

researcher. The realizations of spatial point processes are arrangements or patterns of points and we can observe all the points of such patterns or a sample. The main goal of point pattern analysis is to determine if the location of events tends to exhibit a systematic pattern over the area under study or, on the contrary, they are randomly distributed. More specifically, we are interested in analyzing if the location of events is completely random spatially (the location where events occur is not affected by the location of other events), uniform or regular (every point is as far from all of its neighbors as possible) or clustered or aggregated (the location of events is concentrated in clusters). Some other interesting questions in point pattern analysis include: How does the intensity of a point pattern vary over an area? Over what spatial scales do patterns exist? If along with the location of events a stochastic attribute is observed, the pattern is called a marked pattern; otherwise it will be named an unmarked pattern. Obviously, marked patterns extend the possibilities of spatial analysis.

The above refers to merely spatial data, but in recent years the spatio-temporal data analysis has become a core research area in a large variety of scientific disciplines. In the spatio-temporal context, the observed data are viewed as partial realizations of a spatio-temporal random function which spreads out in space and evolves in time. Thus, spatio-temporal data simultaneously capture spatial and temporal aspects of data.

Recalling some of the examples used to illustrate the types of spatial data, if we observe every hour the level of a specific pollutant in a city at the points where the monitoring stations are located, we have a spatio-temporal geostatistical dataset. Now, based on the spatio-temporal observations, we aim to reproduce the behavior of the spatio-temporal pollution process, or simply predict its value at a space-time point. Geostatistics takes advantage of the spatio-temporal correlations existing in the spatio-temporal data (the interaction of space-time is crucial) to make predictions at unobserved space-time locations. If we annually record the percentage of households with problems of pollution and odors in the census tracts of Madrid,

we have a collection of spatio-temporal lattice data. Now we can study how the spatial percentage pattern evolves in time. If we observe the location of fires in Castilla-La Mancha for the last ten years, we have a spatio-temporal point pattern database. Now we can express the relationship of points not only by distance but by time lag, and can study whether there is complete spatio-temporal randomness in the disposition of the space-time events, or they exhibit a spatio-temporal aggregation, or their spatio-temporal disposition is regular or uniform.

1.2 The limits of classical statistics

As is well known, classic statistics is based on the independence of the observed values. These observed values are considered as independent realizations of the same random variable. However, when the observed values are anchored in space, the hypothesis of independence is no longer acceptable. As stated in the First Law of Geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler 1970).

In this section we illustrate some of the consequences of ignoring the spatial dependencies in the data and using classical statistical methods with spatial data. For the sake of simplicity, we will focus on the spatial case but the consequences of using classical statistics with spatio-temporal data are the same.

Suppose that $\{Z(\mathbf{s}_1), Z(\mathbf{s}_2), \ldots, Z(\mathbf{s}_n)\}$ are n identically distributed observations recorded at spatial points $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. More specifically, suppose that they follow a Gaussian distribution with unknown mean, μ, and known variance, σ_0^2, and that covariances between the observed points are positive and diminish with the distance between them, h, according to the expression: $C(h) = \sigma_0^2 \left(1 - \left(\frac{3}{2}\frac{h}{110} - \frac{1}{2}\frac{h^3}{110^3}\right)\right)$.

If our goal is the estimation of the unknown mean and we ignore the existing spatial correlations, the sample mean, $\bar{Z}$, would undoubtedly be the estimator proposed for μ. Ignoring the spatial correlations, it is well known that $E(\bar{Z}) = \mu$ and $V(\bar{Z}) = \sigma_0^2/n$. But, if we consider such correlations, the sample mean continues to be an unbiased estimator of μ but now its variance is:

$$\begin{aligned} V(\bar{Z}) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} C(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \\ &= \frac{1}{n^2}\left(n\sigma_0^2 + 2\sum_{i<j}^{n} C\left(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\right)\right) \\ &= \frac{1}{n^2}\left(n\sigma_0^2 + 2\sum_{i<j}^{n} \sigma_0^2\left(1 - \left(\frac{3}{2}\frac{h}{110} - \frac{1}{2}\frac{h^3}{110^3}\right)\right)\right) \\ &= \frac{\sigma_0^2}{n}\left(1 + \frac{2}{n}\sum_{i<j}^{n}\left(1 - \left(\frac{3}{2}\frac{h}{110} - \frac{1}{2}\frac{h^3}{110^3}\right)\right)\right), \end{aligned} \tag{1.1}$$

that is, the variance of the sample mean is larger than in the random sample case.

If we ignore the existing correlations in the data and use $\bar{Z}$ as an estimator of μ, the under-estimation of $V(\bar{Z})$ brings unfortunate consequences for inference about μ. The classical confidence intervals for a specific confidence level will be narrower than they really are, or, in other words, the confidence level of classical intervals is larger than it really is. When testing, if spatial correlations are not taken into account, the p-values will be larger than they really are and this will lead to undesirable rejections of the null hypothesis. In addition, the power of the tests will be overstated.

If the number of observations is 16 and they are recorded over a regularly spaced 4×4 grid ($75m \times 75m$) we have, ignoring the spatial correlations between observations, $V(\bar{Z}) = \sigma_0^2$, so that the 95% confidence interval for μ is $\left(\bar{Z} - 1.96\frac{\sigma_0}{\sqrt{16}}; \bar{Z} + 1.96\frac{\sigma_0}{\sqrt{16}}\right)$ and the test-statistic for $\mu = 0$ is $t = \frac{\bar{Z}}{\sigma_0\sqrt{1/16}}$.

When we take into account the spatial correlations between the observations, we obtain:

(i) $V(\bar{Z}) = \frac{\sigma_0^2}{n}\left(1 + \frac{2}{n}\sum_{i<j}^{n}\left(1 - \left(\frac{3}{2}\frac{h}{110} - \frac{1}{2}\frac{h^3}{110^3}\right)\right)\right) = 0.3969\sigma_0^2$.

(ii) 95% confidence interval:

$$\left(\bar{Z} - 1.96 \times 0.630\sigma_0; \bar{Z} + 1.96 \times 0.630\sigma_0\right) = \left(\bar{Z} - 1.235\sigma_0; \bar{Z} + 1.235\sigma_0\right).$$

(iii) Test-statistic for testing $\mu = 0 : t = \frac{\bar{Z}}{0.630\sigma_0}$.

As can be seen, when we take into account the spatial correlations, the variance of $\bar{Z}$ lengthens by more than six, the width of the 95% confidence interval increases by 2.52 and the value of the test-statistic decreases by 2.52.

When estimating the mean, the classical estimator is unbiased in the presence of spatial correlations. However, if we are interested in the estimation of σ^2, the quasi-variance S^{*2} is not. In effect,

$$E(S^{*2}) = \sigma^2 - \left(\frac{2}{n(n-1)}\sum_{i<j}^{n} C\left(Z(\mathbf{s}_i), Z(\mathbf{s}_j)\right)\right). \tag{1.2}$$

The lesson taken from the above example is that ignoring spatial correlations and continuing to rely on the best estimators for the case of independent observations is not a good idea. These are not the appropriate estimators when data are correlated and using them usually leads to wrong decisions. A better idea would be to obtain the best estimators for the case of spatial correlations.

The perverse effects of not taking into account the spatial correlation and using classic estimators also appear in the field of prediction. Some examples can be found in Cressie (1993, pp. 15–17) and Schabenberger and Gotway (2005, pp. 32–4), among others.

1.3 A real geostatistical dataset: data on carbon monoxide in Madrid, Spain

In order to illustrate the main concepts presented in the book, we use toy examples, theoretical examples, classical examples, and examples based on simulated and real data. Pretend, theoretical and classical examples as well as simulated data-based examples are useful when illustrating and helping to better understand the core geostatistical concepts. However, the use of real data reveals some difficulties that the researcher may encounter during the process of applying geostatistical procedures, which are not usually encountered when data are simulated or simply any set of numbers. This is why we find it extremely exciting to deal with real data.

In this book real data has been used to illustrate specific geostatistical questions. For example, in Section 1.1 data on the percentage of households with problems of pollution and odors in each census tract of Madrid, Spain, 2001, and on the number of fires in Castilla-La Mancha, Spain, 1998, were used to illustrate lattice data and point pattern data, respectively; in Section 4.8 real data on coal-ash for the Robena Mine Property in Greene County, Pennsylvania, are used to illustrate the median-polish kriging procedure; in Section 9.1 daily temperature data for 35 Canadian monitoring stations are used to convert raw data into functional data.

But the real database we use in this book to illustrate the spatial, spatio-temporal, and functional kriging procedures focuses on air pollution in Madrid, Spain, and more specifically on carbon monoxide (CO). The reasons for choosing a database on air pollution (and specifically on CO) for this purpose are the following:

(i) The health effects of air pollution. Air pollution is one of the most important pollution problems in the world. Many health problems (e.g., respiratory and cardiovascular) can be caused or worsened by exposure to air pollution on a daily basis. Therefore, it is not surprising that the World Health Organisation has ranked urban air pollution the 13th contributor to global deaths in its 2012 World Health Report.

(ii) The economic costs (explicit or accounting costs and implicit or social costs) of pollution at the local, provincial, regional, national, or global scale are considerable, and the cost of ignoring them would be even higher.

(iii) The urban environment is presently an issue of the utmost concern, not only for the citizens themselves, increasingly more aware of the conditions of the environment they live in, but also for health authorities and political leaders, who have also become aware of these problems.

The reason for choosing the city of Madrid as the study area is that it is one of the European largest cities where air pollution remains a serious problem. It is the third largest city in the European Union in terms of population and is undergoing

Figure 1.4 Location of the monitoring stations in the city of Madrid.

a significant suburbanization process that is resulting in both people and workplaces being concentrated not only in the heart of the city, but also in surrounding areas. This process is resulting in an increase in Madrid dwellers being much more dependent on cars than in the past, as many of the jobs that have been created are located in areas up to 20km from the center of the city, which means more vehicle exhaust emissions than desired (the main source of CO in the city of Madrid). Although Madrid does have a remarkable urban transport system, the truth is that in autumn, winter, and the month of July, there is a massive tendency to use cars, as well as the urban trips that are so necessary for everyday economic activity.

In the city of Madrid, as in all large cities in the world, there is a monitoring network that measures the level of the pollutants particularly harmful to human health (CO, NO_2, NOx, O_3, PM_{10}, and SO_2, among others) on an hourly basis. The Atmosphere Pollution Monitoring Network (APMN) of the city of Madrid (http://www.mambiente.munimadrid.es/opencms/opencms/calaire) is made up of 27 fixed monitoring stations (see Figure 1.4), although only 23 have been continuously operative (monitoring stations 2, 17, 26, and 27 have experienced in the last few years a number of interruptions in their operation that led us to remove them from the database).

As a consequence of the fact that the main source of CO emissions in the city of Madrid is vehicle exhaust emissions, it is no surprise that (i) levels of CO are different on weekdays and weekends and holidays, and controlling such levels is only of interest in the week (or work) days; and that (ii) there are no significant differences in the levels of CO for the five weekdays. This is the reason why a "typical working day" was created for each week (that is, weekends and bank holidays were eliminated) and the hourly average for working days was computed.

Thus, taking as a starting point the hourly data (in mg/cm^3) provided by APMN in 2008, we constructed a new database as follows:

1. We removed the weekdays and holidays from the initial database and considered only the weekdays as a single entity: "typical working day."

2. For each of the 52 weeks of the year 2008 we averaged the log CO data registered on the five weekdays at each of the 23 monitoring stations operating in Madrid in 2008 on an hourly basis. The reason for making a logarithmic transformation of the raw data and working with the log of CO is that CO hourly measurements do not follow a Gaussian distribution. However, after the logarithmic transformation of the data, a significant departure from a Gaussian distribution in any of the stations cannot be found.

3. As a consequence, for each of the 23 monitoring stations, we have $52 \times 24 = 1248$ spatio-temporal data $\bar{x}_{i,j}^{k}$, where $\bar{x}$ represents the averaged log of CO, $i = 1, \ldots, 24$ indicates the hour, $j = 1, \ldots, 52$, the week, and $k = 1, \ldots, 23$, the monitoring station. For example, $\bar{x}_{13,40}^{9}$ indicates the average of log CO measurements registered in station 9 at 13 hours on the five weekdays of week 40. Notwithstanding, in order to directly refer to the pollutant under study, in what follows we will refer to such a mean as $logCO^{*}$, the asterisk indicating the arithmetic mean.

4. Therefore, the final database contains a total of $23 \times 1248 = 28704$ $logCO^{*}$ data available from the book's website: www.wiley.com/go/montero/spatial.

Although the monitoring stations that make up the APMN are located in accordance with their ultimate purpose, it is also true that on numerous occasions the size of such stations and/or local government regulations in force at the time are significant restrictions when it comes to locating them in the optimal places. For this reason it is of vital importance to predict the level of the main pollutants in places that are especially affected by traffic congestion and, particularly, at the main intersections in the center of the city, because this is where the health of a large number of Madrid residents and people visiting the city suffers. We will use spatial, spatio-temporal, and functional kriging procedures to, among other things, predict the level of CO at such non-observed sites.