24 Analysis of areal tornado data

A large number of spatial data sets have been analysed in Volumes I and II. All spatial data sets have one thing in common; the locations of the sampling sites were in a specific study area and the spatial location of each site was indexed with latitude and longitude. This type of data is also called geostatistical data.

In some studies data are aggregated per area (e.g. country, county, city), and these areas may have different sizes. Such data are called areal or lattice data. In ecology most data sets are geostatistical data sets, but occasionally we stumble upon areal data sets. In this chapter we will discuss how to analyse areal and lattice data. We will use a tornado data set that was analysed in Jagger et al. (2015). We will apply generalised additive models with spatial dependency using the conditional autoregressive (CAR) correlation. With 83% of the observations equal to 0, we are very quickly in the world of zero-inflated models again. Because there is also a temporal element, we extend the models to spatial-temporal correlation within the context of areal data.

Prerequisite for this chapter: Knowledge of multiple linear regression, generalised linear models, generalised additive models, zero-inflated Poisson (ZIP) models and how to fit these models in INLA (Chapters 7–10 of Volume I and Chapters 19 and 20 of Volume II) is required.

24.1 Areal versus geostatistical data

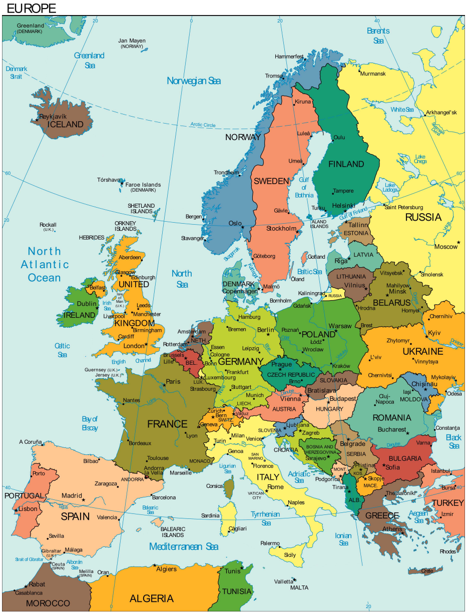
Spatial data is defined as realisations $y(\boldsymbol{s})$ of a stochastic process $Y(\boldsymbol{s})$, were $\boldsymbol{s}$ is part of a study area D. The study area is typically two dimensional.

Suppose that we sample the number of tuberculosis cases in cattle in each county in England, or the total air pollution for each European country, or the total number of fish catches per European country, or the number of deaths per health clinic in a large city, or the number of babies born per country, or the average pH per river catchment, or the crime rates per city block. In all cases $\boldsymbol{s}$ represents an areal unit (country, county, catchment for a river or health clinic) and $y(\boldsymbol{s})$ is the aggregated or averaged value for that areal unit. In all these examples D is a countable collection of spatial units. If the areas are regularly placed then we call it lattice data; otherwise we call it areal data.

Areal data are common in medical science because quite often only aggregated data (per hospital or county) are available due to patient confidentiality. In ecology, one has to search much harder to find areal or lattice data. Two examples are the aggregation of numbers of species per area (e.g. fish landings per area of the sea; see www.ices.dk for examples) and average air pollution levels per country (see for examples www.eea.europa.eu/).

In geostatistical data analysis we do not work with a value $y(\boldsymbol{s})$ for a specific areal unit $\boldsymbol{s}$. Instead $\boldsymbol{s}$ represents a spatial index like latitude and longitude. An example is the Irish pH data that was discussed in Chapters 2 and 4 of Volume I. There we have 253 observations $pH(\boldsymbol{s}_1)$, $pH(\boldsymbol{s}_2)$ to $pH(\boldsymbol{s}_{253})$ in the study area D, and the $\boldsymbol{s}_i$ contain the spatial coordinates. Geostatistical examples from Volume II are the skate data (Chapter 20), the sandeel data (Chapter 21), the Canadian seabird data (Chapter 22) and the coral reef data (Chapter 23)

The statistical techniques required for the analysis of areal data and geostatistical data are different. If we have areal data consisting of numbers per country, then it doesn't always make sense to quantify correlation via a variogram that depends on Euclidean distances. As an example of this, the map below shows Europe. Suppose we quantify a bird species diversity index per country and model this index as a function of average rainfall per country. At one point during the analysis we will need to take the residuals and assess these for spatial dependency. For this we will need the latitude and longitude for each country, and these allow us to make a sample variogram of the residuals. And here is where the first problem lies: what latitude and longitude values do we use? We could use the spatial coordinates of the point that lies exactly in the middle of a country. But this leads to a new problem. The Netherlands (which is one of the three small countries between Germany and the UK) is a small country, whereas its neighbour Germany is a very large country. If we use the latitude and longitude coordinates of the centre point of each country, then it looks as if these countries are far away from each other.



Instead of using Euclidean distances (based on the Pythagorian theorem) to quantify how close two countries are, we can use other tools. For example, we can state that two countries are close to each other if they share a common border or are within a certain distance of each other. We can quantify the first option with a 1 (sharing a border) and 0 (not sharing a border).

For geostatistical data analysis we use Euclidean distances, variograms and other tools to quantify dependency. For areal or lattice data we typically use information whether spatial units share a common border.

In certain situations (e.g. for areal or lattice data where the units are of the same size) we can use tools from geostatistical data analysis. And if we have geostatistical data and aggregate the data, then we can apply tools from areal and lattice data analysis.

24.2 Who are the neighbours?

Figure 24.1 shows a map of the state of Illinois in the United States; it has been divided into a large number of non-overlapping, irregularly shaped areas. These are the 102 counties of Illinois.

Let us sketch a couple of examples of how areal data may arise. Suppose that in a medical study we are interested in whether the number of deaths from a certain disease differ spatially. Due to patient confidentiality it can be quite difficult to obtain data at postal code level, and such data may only be available as an aggregated number per county.

In a waste management study you might be interested to see how different counties perform with respect to household rubbish collection. Suppose you record the total amount of collected rubbish per county and how much of that is recycled. It may be interesting to compare the performances per county. Other examples are crime rates per county, unemployment rates per county, income per county, etc.

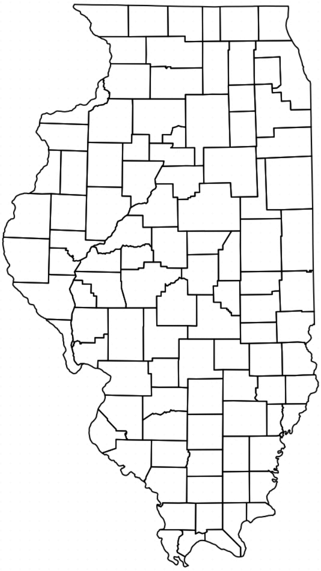


Figure 24.1. Counties (102) within the state of Illinois (US).