**Lecture 10 Types of Spatial Data and Models for Geostatistical Data**

**10.1 Types of spatial data**

Spatial data have spatial reference: they have ***coordinate values*** and a system of reference for these coordinates. If we also have the date and time of the observation, this information is called an ***attribute***: it is non-spatial in itself, but this attribute information is believed to exist for each spatial entity.

Spatial data can be classified by how the location associated with each observation is defined: by a point (e.g., latitude and longitude of the location) or by an area or region.

Point spatial data can be further classified by whether it is geostatistical or a spatial point pattern:

* ***Geostatistical data*** have point locations associated with them and usually one or more variables are measured at each location. For example, data from an air quality monitoring network may include observations of ozone, particulate matter, temperature, humidity, radiation, wind speed, and time of day for each monitoring location.
* ***Spatial point pattern data*** have point locations associated with them and the locations themselves are the variable of interest, e.g., the locations of bird nests. Often one of the main hypotheses of interest is whether the locations are random, clustered, or regular.

***Lattice data*** are observations associated with an area or region, such as the vegetation index of a pixel on a remote-sensing image or the cancer rate within each county of a state.

In the following talks we will use geostatistical data to illustrate a few tools for spatial statistics.

A few data related items are as follows:

* *Point*, a single point location, such as a GPS reading or a geocoded address;
* *Line*, a set of ordered points, connected by straight line segments;
* *Polygon*, an area, marked by one or more enclosing lines, possibly containing holes;
* *Grid*, a collection of points or rectangular cells, organised in a regular lattice.

**10.2 Visualising spatial data**

**Example 10.1** Re the benthic data (in {*EnvStats*}), Figure 10.1 shows the sampling station locations. Recall that the bubble plot of benthic indices (Figure 1.8) exhibits the locations of the observations and the sizes of the plotting symbol proportional to the values of the benthic index. Because of the placement of the sampling locations, the bubble plot is difficult to interpret, but it looks like low values of the benthic index occur at the southern stations, and also between latitudes 38.5 to 39 at longitude 76.4. Figure 10.3 displays a contour plot of the benthic index and Figure 10.4 displays a surface plot. Here it is easier to see the areas of low index values. Both of these plots are based on a ***loess smooth*** (locally weighted scatterplot smoothing) in three dimensions.

|  |
| --- |
| library(EnvStats)  attach(Benthic.df)  plot(Longitude, Latitude,xlab = "-Longitude (Degrees West)",ylab = "Latitude",main = "Figure 10.1 Sampling Station Locations")  pairs(~ Index + Salinity + Silt, data = Benthic.df, main="Figure 10.2 Pairs")  library(sp)  loess.fit <- loess(Index ~ Longitude \* Latitude,  data=Benthic.df, normalize=FALSE, span=0.25)  lat <- Benthic.df$Latitude  lon <- Benthic.df$Longitude  Latitude <- seq(min(lat), max(lat), length=50)  Longitude <- seq(min(lon), max(lon), length=50)  predict.list <- list(Longitude=Longitude,  Latitude=Latitude)  predict.grid <- expand.grid(predict.list)  predict.fit <- predict(loess.fit, predict.grid)  index.chull <- chull(lon, lat)  inside <- point.in.polygon(point.x = predict.grid$Longitude,  point.y = predict.grid$Latitude,  pol.x = lon[index.chull],  pol.y = lat[index.chull])  predict.fit[inside == 0] <- NA  contour(Longitude, Latitude, predict.fit, levels=seq(1, 5, by=0.5), labcex=0.75,xlab="-Longitude (degrees West)", ylab="Latitude (degrees North)")  title(main=paste("Figure 10.3 Contour Plot of Benthic Index", "Based on Loess Smooth", sep="\n"))  persp(Longitude, Latitude, predict.fit, xlim = c(-77.3, -75.9), ylim = c(38.1, 39.5), zlim = c(0, 6), theta = -45, phi = 30, d = 0.5, xlab="-Longitude (degrees West)",  ylab="Latitude (degrees North)",  zlab="Benthic Index", ticktype = "detailed")  title(main=paste("Figure 10.4 Surface Plot of Benthic Index", "Based on Loess Smooth", sep="\n"))  detach("Benthic.df")  rm(loess.fit, lat, lon, Latitude, Longitude, predict.list, predict.grid, predict.fit, index.chull, inside) |
|  |

█

**10.3 Models for geostatistical data**

In Part II we discussed models for time series data (data collected over time), including linear regression models (Lecture 7). Sometimes we explicitly included time as a predictor variable in the model (Examples 7.1-3) and sometimes we included variables in the model that vary with time without explicitly including time as a predictor variable (Exercise 7.2). We can extend these same ideas to geostatistical data (data collected over space).

With time series models we often use *t* as a subscript or variable to denote the time at which an observation was taken. For geostatistical data we use ***s*** to denote the location, where ***s*** *=* (*x,y*) denotes the Cartesian coordinates of the location. The observation at location ***s*** is denoted by *z*(***s***) = *z*(*x,y*).

A general form for a model of geostatistical data is given by

Z(***s***) = *f*(***s***) + ɛ(***s***)

where*f* is some parametric, or nonparametric, functionof ***s***, a vector of predictors such as the location variables and any other variables thought to be important. The error term of the model is ɛ, assumed to have a mean of 0 and a constant standard deviation σ.

The simplest parametric model for a spatial trend involves fitting a plane:

z =  + *x* + *y*.

The polynomial function is often used in geostatistical models. For instance, a model fits a quadratic surface:

z =  + *x* + *y* + + + *xy*.

In general, one can fit any kind of polynomial surface to describe changes in the response variable *Z* as a function of location. One can also add other predictor variables to the model, and in fact adding other predictor variables to the model will probably change your model for spatial trend and may even make the trend disappear altogether, obviating the need for the location variables as predictor variables in the model.

The nonparametric models for geostatistical data involve fitting a smooth surface locally within neighbourhoods of a point, for example, the *loess* algorithm.

**Example 10.2** In Example 10.1, the loess smooth with span = 0.25 was used to construct the contour and surface plots. In this example, we fit a 4th degree (quartic) polynomial surface.

|  |
| --- |
| library(EnvStats)  attach(Benthic.df)  library(sp)  poly4.fit <- lm(Index ~ poly(Longitude, Latitude, degree=4),data=Benthic.df)  lat <- Benthic.df$Latitude  lon <- Benthic.df$Longitude  Latitude <- seq(min(lat), max(lat), length=50)  Longitude <- seq(min(lon), max(lon), length=50)  predict.list <- list(Longitude=Longitude,  Latitude=Latitude)  predict.grid <- expand.grid(predict.list)  library(gam)  predict.fit <- predict.gam(poly4.fit, predict.grid)  index.chull <- chull(lon, lat)  inside <- point.in.polygon(point.x = predict.grid$Longitude,  point.y = predict.grid$Latitude,  pol.x = lon[index.chull],  pol.y = lat[index.chull])  predict.fit[inside == 0] <- NA  contour(Longitude, Latitude, predict.fit, levels=seq(1, 5, by=0.5), labcex=0.75,xlab="-Longitude (degrees West)", ylab="Latitude (degrees North)")  title(main=paste("Figure 10.5 Contour Plot of Benthic Index", "Based on 4th polynomial", sep="\n"))  persp(Longitude, Latitude, predict.fit, xlim = c(-77.3, -75.9), ylim = c(38.1, 39.5), zlim = c(0, 6), theta = -45, phi = 30, d = 0.5, xlab="-Longitude (degrees West)",  ylab="Latitude (degrees North)",  zlab="Benthic Index", ticktype = "detailed")  title(main=paste("Figure 10.6 Surface Plot of Benthic Index", "Based on 4th polynomial", sep="\n"))  detach("Benthic.df")  rm(loess.fit, lat, lon, Latitude, Longitude, predict.list, predict.grid, predict.fit, index.chull, inside) |
|  |

**█Exercises**

10.1 Re Example 10.1, use the loess smooth with span = 0.10 to construct the contour and surface plots and comment on your findings.

10.2 Re Example 10.2, use a 2nd degree polynomial to construct the contour and surface plots and comment on your findings.

**References**

* Bivand, R. S., Pebesma, E. and Gómez-Rubio, V., (2013), *Applied Spatial Data Analysis with R*, SPRINGER
* Millard, S.P. and Neerchal, N. K. (2000), *Environmental Statistics with S-PLUS*, Chapman & Hall.