WESTERN SYDNEY UNIVERSITY

301035
Environmental Informatics

Dr Jinnat Ali

Unit Co-ordinator, Lecturer, and Tutor
School of Computing, Engineering and Mathematics
Western Sydney University
Room ER.1.o7, Parramatta Campus
E-mail: j.ali@westernsydney.edu.au

301035
Environmental Informatics

Lecture 10
(Types of Spatial Data and Models for Geostatistical Data)

A Brief Revisions - Model Specification

Model Specification:

- Using plots of the data, autocorrelations, partial autocorrelations, and other information, a class of simple ARIMA models is selected.

Model Estimation:

- The autoregressive and moving average parameters are found via an optimization method like maximum likelihood.

Diagnostic Checking:

- The fitted model is checked for inadequacies by studying the autocorrelations of the residual series (i.e., the time-ordered residuals).

A Brief Revisions - Forecast

Forecasting involves making predictions about the future.

It is required in many situations (e.g):

- **Deciding whether to build another power generation plant in the next ten years requires forecasts of future demand;**
- **Scheduling staff in a call centre next week requires forecasts of call volumes;**
- **Forecasts can be required several years in advance, or only a few minutes beforehand.**

Whatever the circumstances or time horizons involved, forecasting is an important aid to effective and efficient planning.

Types of Spatial Data

Spatial Data

Spatial Data known as geospatial data or geographic information.

It is the data or information that identifies the geographic location of features and boundaries on Earth.

For example, natural or constructed features, oceans, and more.

Spatial data is usually stored as coordinates and topology, and is data that can be mapped.

The word ***geospatial*** is used to indicate that **data** that has a *geographic component* to it.

This means, the records in a dataset have locational information tied to them such as geographic **data** in the form of coordinates, address, city, or ZIP code.

GIS data is a form of **geospatial data**.

GIS Spatial Data

There are two components to GIS **data**:

Spatial Information:

- It is coordinate and projection information for **spatial** features.

Attribute Data:

- The **spatial data** is the where and **attribute data** can contain information about the what, where, and why.

Geostatistical Data

Geostatistics is a branch of statistics focusing on spatial or spatiotemporal datasets.

Geostatistics is currently applied in diverse disciplines including:

- Petroleum geology,
- Hydrogeology,
- Hydrology,
- Meteorology,
- Oceanography,
- Geochemistry,
- Geography,
- Forestry,
- Environmental control,
- Landscape ecology,
- Soil science, and
- Agriculture (esp. In precision farming).

Geostatistical Data

- ***Geostatistics*** is applied in varied branches of geography, particularly those involving the spread of diseases (epidemiology),
- The practice of commerce and military planning (logistics), and the development of efficient spatial networks.
- *Geostatistical* algorithms are incorporated in many places, including geographic information systems (GIS) and the **R** statistical environment

Type of Spatial Data

Spatial data have spatial reference:

- They have ***coordinate values*** and a ***system*** of reference for these coordinates.
- **Attribute:**
- It also have the date and time of the observation, this information it is non-spatial in itself,
- But this attribute information is believed to exist for each spatial entity (e.g. an entity is something that exists as itself).
- 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
  - an area or region.

Point of Spatial Data

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.

Point of Spatial Data

Spatial point pattern data have point locations associated with them and the locations themselves are the variable of interest.

For example, the locations of bird nests.

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.

Point of Spatial Data

The Table summarizes the various classifications of spatial data.

Location Type	Spatial Data Type	Examples
Point	Geostatistical	Air quality at monitoring stations TcCB concentrations on a grid Benthic index at irregularly-spaced sampling sites
	Spatial Point Pattern	Locations of bird nests Locations of cancer cases
Areal (Regional)	Lattice	Number of cancer cases for a county Vegetation index for a pixel on a remote-sensing image

A Few Tools for Spatial Statistics

In the following discussions 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.

The Benthic Data

The data frame *benthic.df* in *ENVIRONMENTALSTATS* for S-PLUS contains data from the Long Term Benthic Monitoring Program of the Chesapeake Bay. The data consist of measurements of benthic characteristics and a computed index of benthic health for several locations in the bay.

Variable	Description (Units)
`Site.ID`	Site ID
`Stratum`	Stratum Number (101–131)
`Latitude`	Latitude (degrees North)
`Longitude`	Longitude (negative values; degrees West)
`Index`	Benthic Index
`Salinity`	Salinity (ppt)
`Silt`	Silt Content (% Clay in Soil)

Visualising Spatial Data

Remember 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.

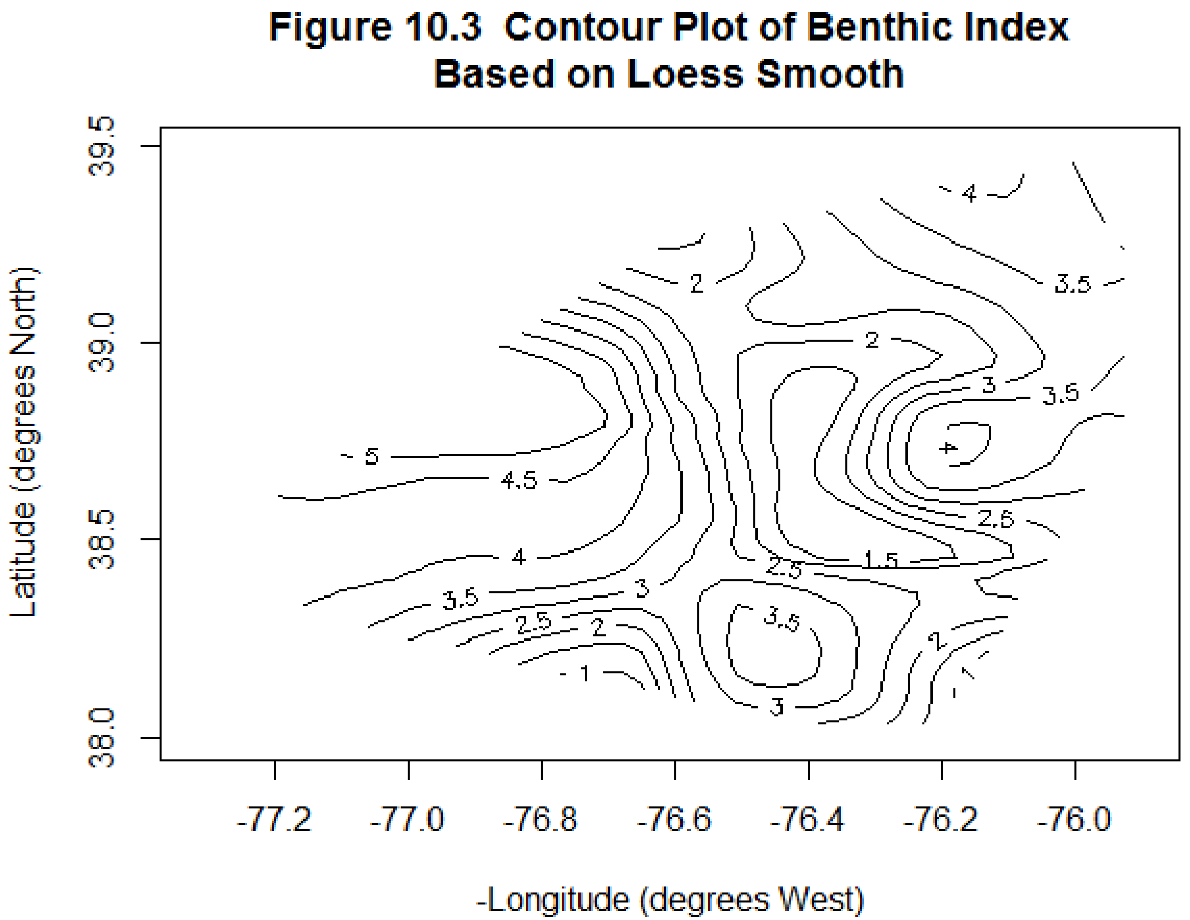
Because of the placement of the sampling locations, the bubble plot is difficult to interpret,

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.

Visualising Spatial Data

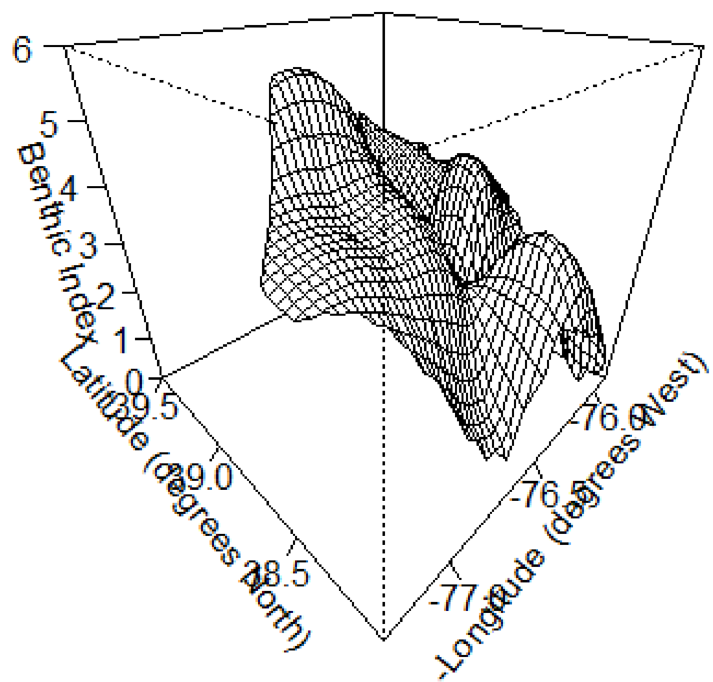
Both of these plots are based on a *loess smooth* (locally weighted scatterplot smoothing) in three dimensions.



Contour is a line on a map joining points of equal height above or below sea level

Visualising Spatial Data

Figure 10.4 Surface Plot of Benthic Index Based on Loess Smooth



Here it is easier to see the areas of low index values.

Visualising Spatial Data

```
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)
```

Visualising Spatial Data

```
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)
```

Recall Figure 1.8

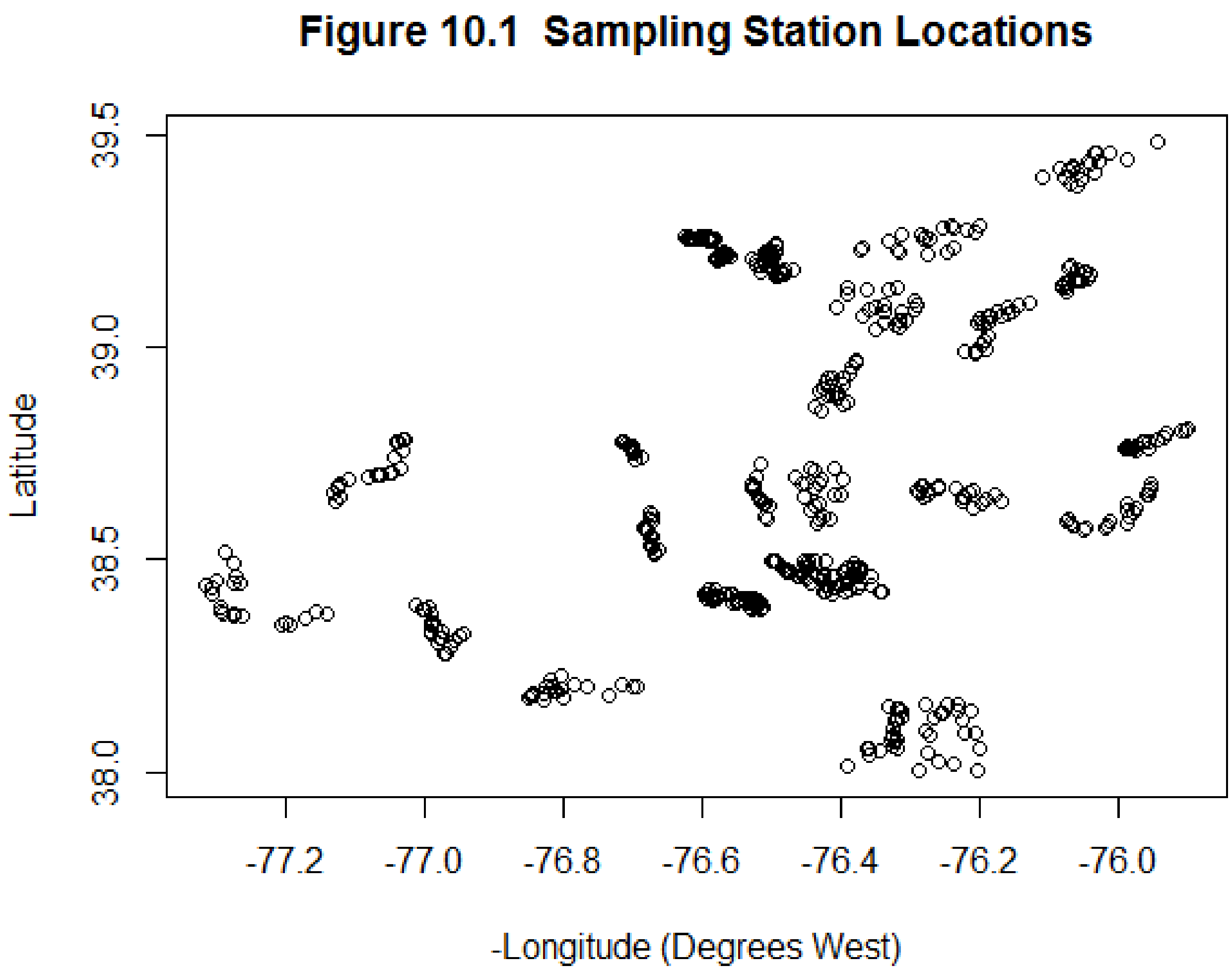


Figure 10.2

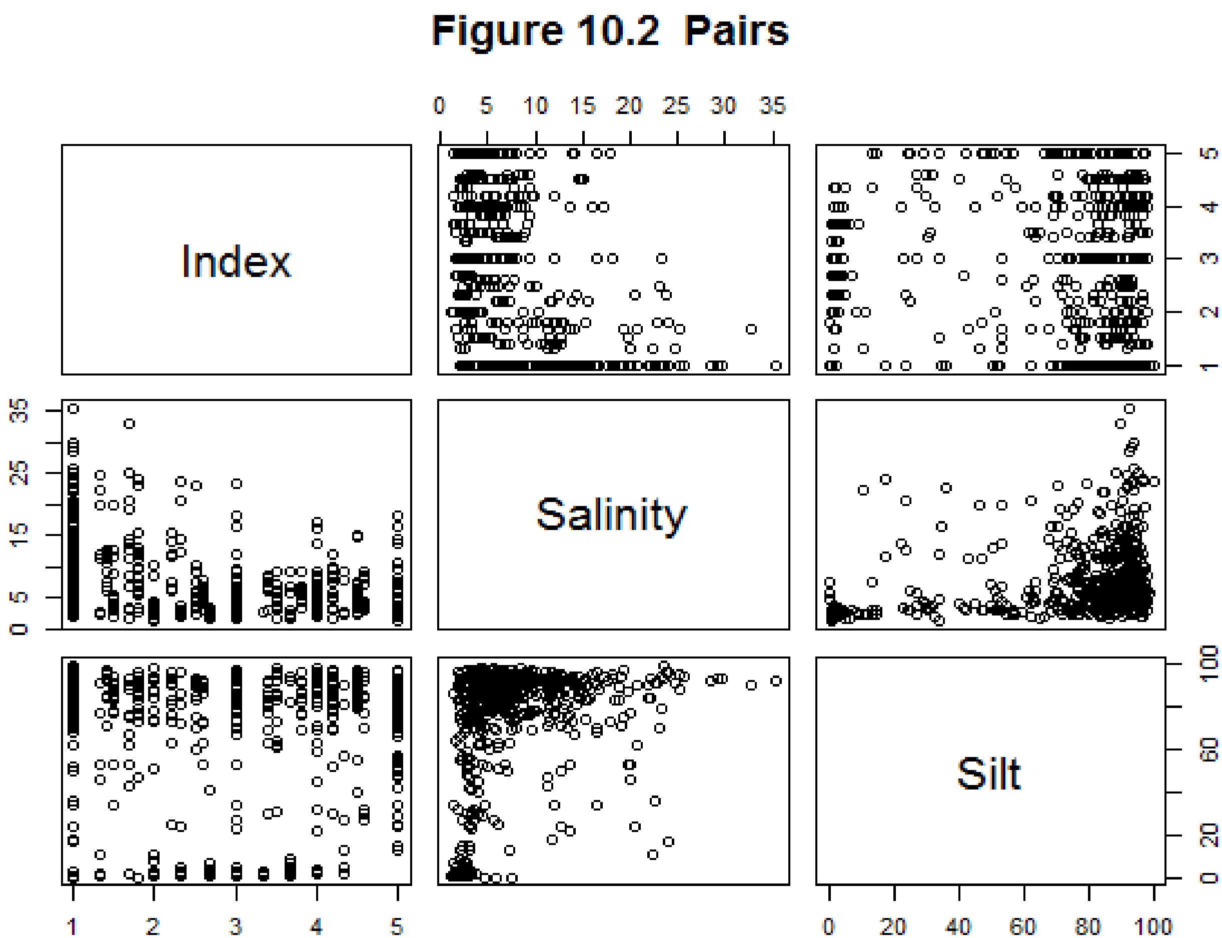


Figure 10.2 Pairs
Index
Salinity
Silt
0 5 10 15 20 25 30 35
1 2 3 4 5
0 20 40 60 80 100

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)
- 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).

Models for Geostatistical Data

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

$$\mathbf{z} = \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{y}.$$

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

$$\mathbf{z} = \beta_0 + \beta_1 \boldsymbol{x} + \beta_2 \boldsymbol{y} + \beta_3 x^2 + \beta_4 y^2 + \beta_5 \boldsymbol{xy}.$$

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.

LOESS Model

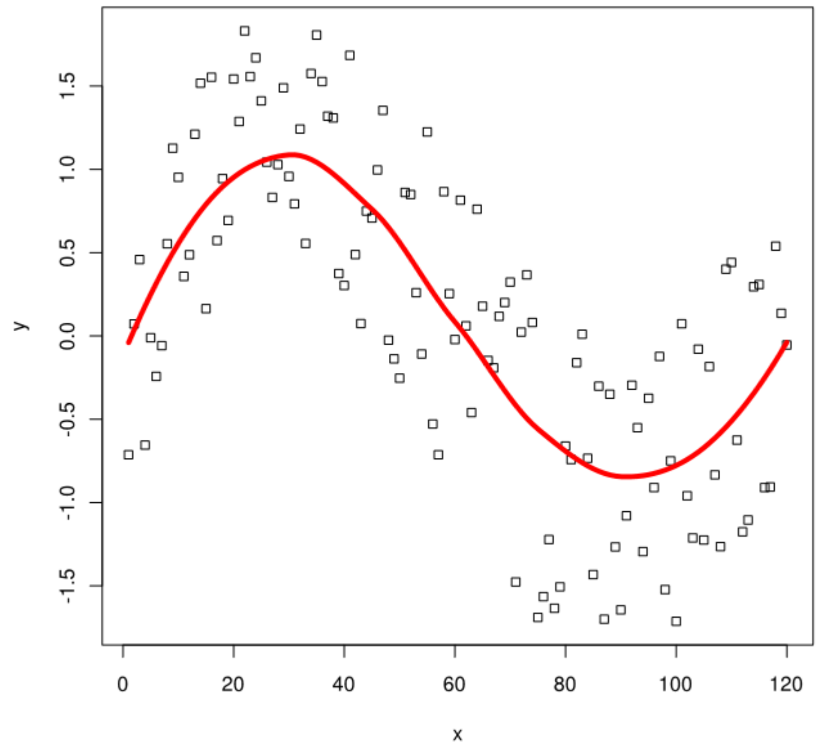
LOESS or **LOWESS** (**locally weighted scatterplot smoothing**) are two strongly related non-parametric regression methods that combine multiple regression models in a *k*-nearest-neighbor-based meta-model.

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

LOESS, originally proposed by Cleveland (1979) and further developed by Cleveland and Devlin (1988), specifically denotes a method that is also known as *locally weighted polynomial regression*.

At each point in the range of the data set a low-degree polynomial is fitted to a subset of the data, with explanatory variable values near the point whose response is being estimated.

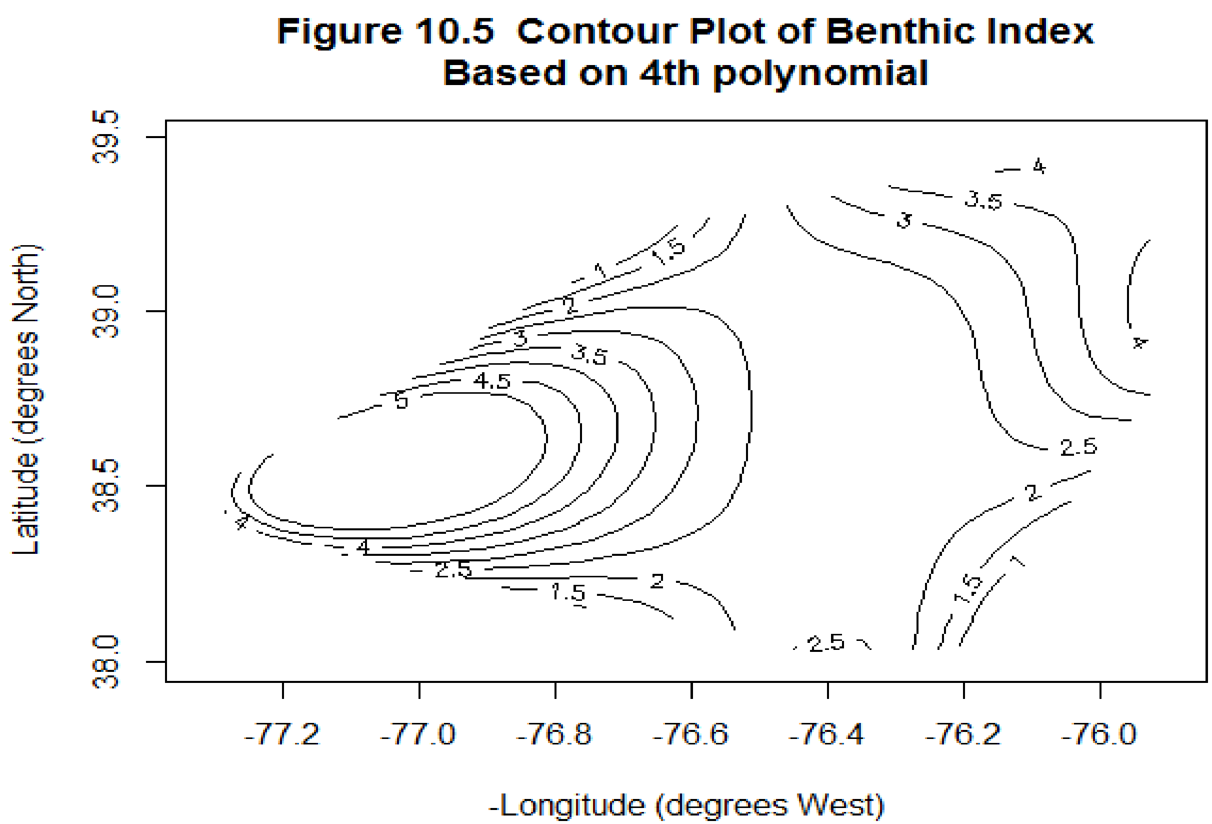
LOESS Model



LOESS curve fitted to a population sampled from a [sine wave] with uniform noise added. The LOESS curve approximates the original sine wave.

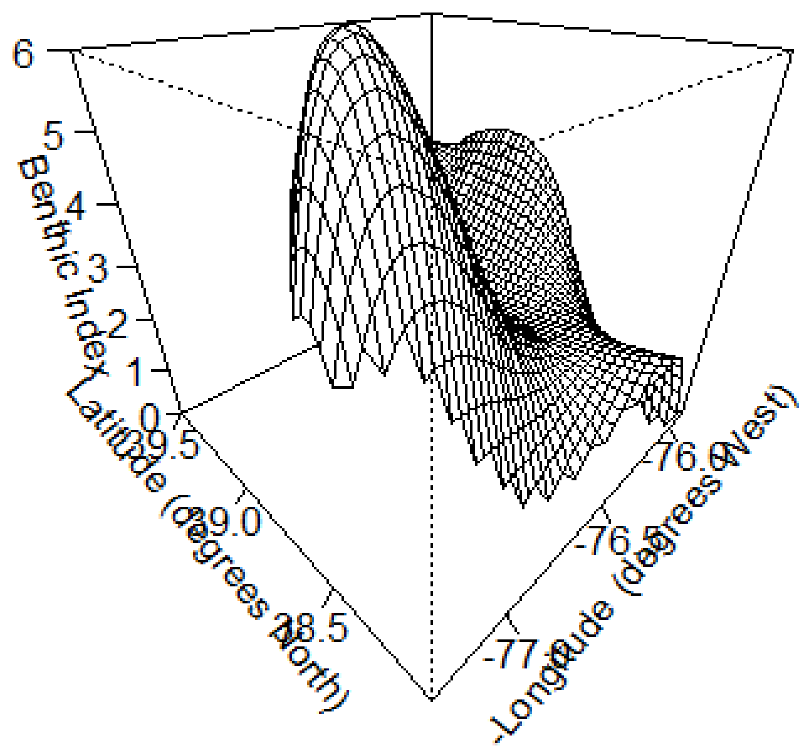
Example 10.2

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.



Example 10.2

Figure 10.6 Surface Plot of Benthic Index
Based on 4th polynomial



Example 10.2

```
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)
```

Example 10.2 (Cont.)

```
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.