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 11
(Spatial Correlation, Variogram and Models)

A Brief Revisions - 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**.

A Brief - 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.

Spatial Correlation, Variogram and Models

Introduction

- In spatial data, observations close together in space will probably look more similar to one another
- Than observations collected farther away from one another,
- So when we fit any kind of trend surface to these data, the errors from the model may be correlated.
- In this talk we discuss tools for modelling spatial correlation.

Introduction (Cont.)

- In geostatistics the spatial correlation is modelled by the variogram.
- Here, the word variogram will be used synonymously with semi-variogram.
- The variogram plots semi-variance as a function of distance.
- To estimate the spatial correlation from observational data, we need to make ***stationarity*** assumptions.
- One commonly used form of stationarity is *intrinsic* (***weak***) stationarity

Spatial Correlation

- In Lecture 6 we defined the lag-*k* covariance and correlation.
- These measure the covariance and correlation between two observations that are *k* units apart in time for a stationary time series,
- That is, the time series has a constant mean and the covariance depends only on the lag *k*.

Spatial Correlation (Cont.)

- Extending this idea to *spatial data*, we might assume that the observations represent a snapshot from a random process over space.
- Weak *stationarity* implies that the surface has a constant mean and that the *covariance* between two observations depends only on the *distance* between the locations of these observations.

Spatial Correlation (Cont.)

The covariance is ***isotropic*** if it depends only on the distance between the locations and not on the direction:

$$\mathrm{Cov}[Z(s_1), Z(s_2)] = C(h)$$

where h is the distance between location s_1 and location s_2.

The function C is called the ***covariogram*** and is analogous to the auto-covariance function for time series data.

Spatial Correlation (Cont.)

The function

$$\rho(h) = C(h)/C(0)$$

is called the ***correlogram*** and is analogous to the autocorrelation function for time series data.

Rather than looking at covariograms and correlograms, people who deal with spatial data often use the variogram, which is defined as:

$$\begin{aligned} \gamma(h) &= \tfrac{1}{2}\mathbf{Var[Z(}s_1\mathbf{) - Z(}s_2\mathbf{)]} \\ &= \mathbf{Var[Z(}s_1\mathbf{)] - Cov[Z(}s_1\mathbf{), Z(}s_2\mathbf{)]} \\ &= \mathbf{C(0) - C(h)} \end{aligned}$$

Spatial Correlation (Cont.)

The estimator of $\gamma(h)$ is given by

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|}\sum_{N(h)}[z(\boldsymbol{s}_i) - z(\boldsymbol{s}_j)]^2$$

where $N(h)$ denotes the set of all pairs of locations that are h units apart, $|N(h)|$ denotes the number of pairs in this set. There are three features in variograms:

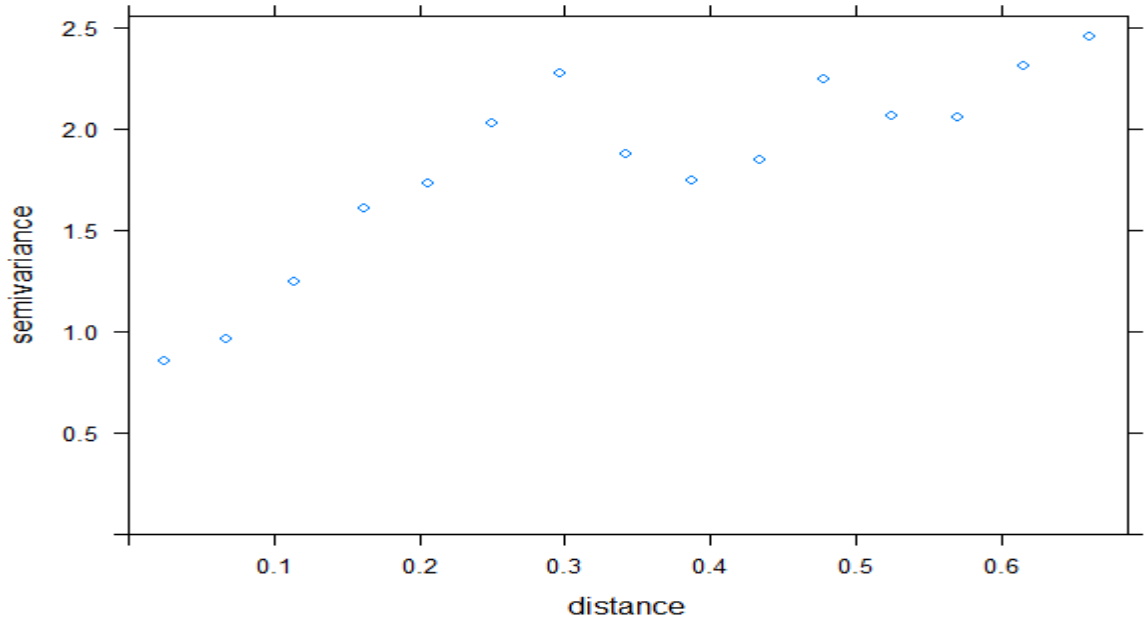
- Sill – The sill is equal to the variance of the process, i.e., the covariance $C(\mathrm{h})$ at distance $h = 0$;
- Range – The range is the distance at which observations are no longer correlated. The range may be finite or infinite;
- Nugget effect – If $\gamma(0) \neq 0$. The nugget effect represents micro-scale variation and/or measurement error.

Example 11.1

Empirical variogram for the benthic data (There appears some nugget effect)

```
library(sp)
library(EnvStats)
library(gstat)
data(Benthic.df)
coordinates(Benthic.df)=~Longitude+Latitude
vg.benthic <- variogram(Index ~ 1, data=Benthic.df)
plot(vg.benthic, main="Figure 11.1  Empirical variogram for Benthic Index
")
```

Figure 11.1 Empirical variogram for Benthic Index



Variogram Modelling

What is Variogram ?

- The variogram characterizes the spatial continuity or roughness of a data set.
- A variogram is a statistically-based, quantitative, description of a surface's roughness.
- A variogram is a function of a separation vector: this includes both distance and direction, or a Δx and a Δy.
- The variogram function yields the average dissimilarity between points separated by the specified vector.
- That is dissimilarity is measured by the squared difference in the Z-values.

Variogram Represent

Consider two synthetic data sets; we will call them **A** and **B**. Some common descriptive statistics for these two data sets are given in Table 1.1.

	A	B
Count	15251	15251
Average	100.00	100.00
Standard Deviation	20.00	20.00
Median	100.35	100.92
10 Percentile	73.89	73.95
90 Percentile	125.61	124.72

Table 1.1 Some common descriptive statistics for the two example data sets.

Variogram Represent

The histograms for these two data sets are given in Figures 1.1 and 1.2. According to this evidence the two data sets are almost identical.

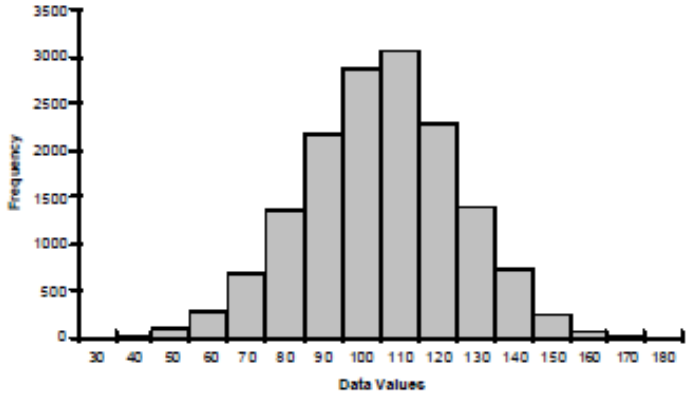


*Figure 1.1 Data Set **A** Histogram*

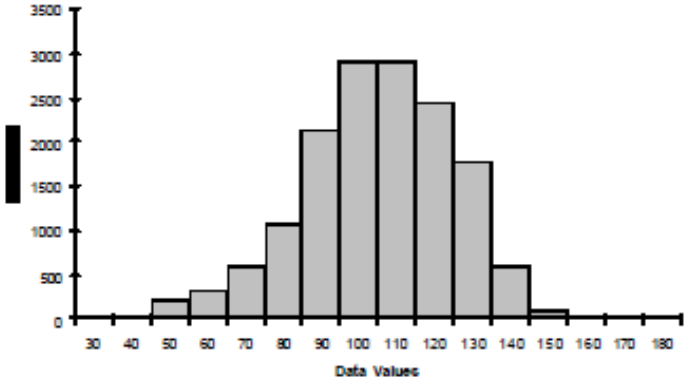


*Figure 1.2 Data Set **B** Histogram*

However, these two data sets are significantly different in ways that are **not** captured by the common descriptive statistics and histograms.

Variogram Represent

As can be seen by comparing the associated contour plots (see Figures 1.3 and 1.4), data set **A** is rougher than data set **B.**

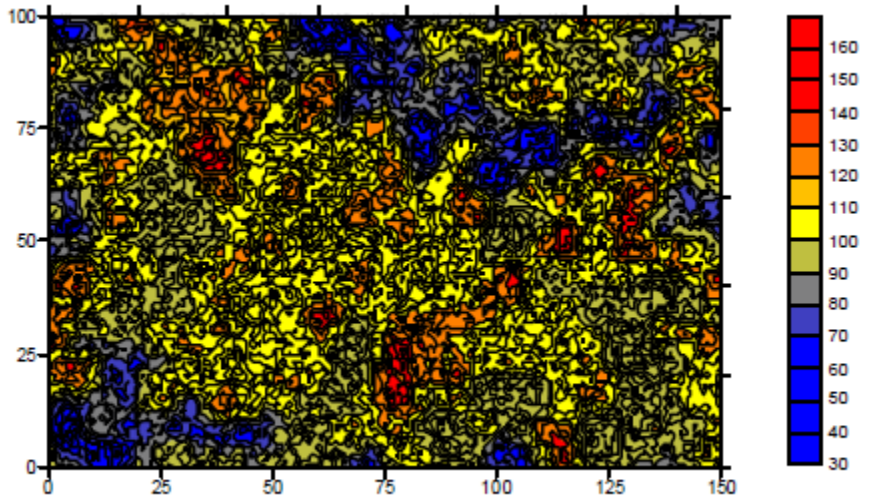


*Figure 1.3 Data Set **A** Contour Plot*

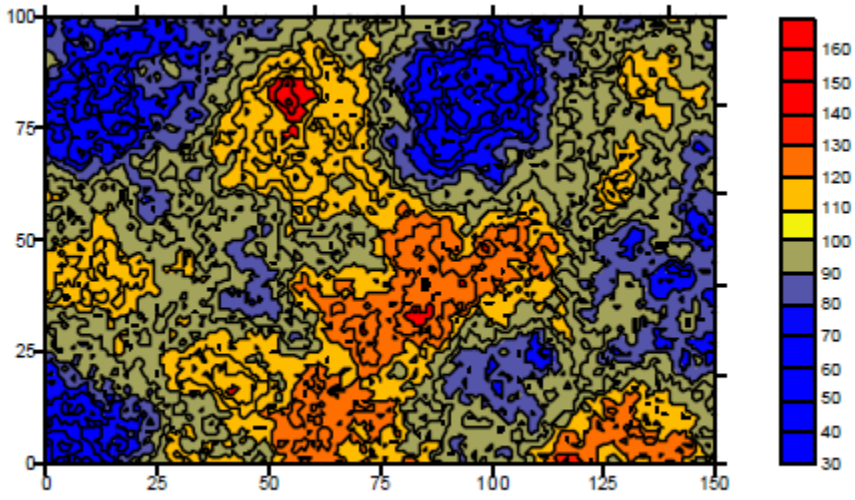


*Figure 1.4 Data Set **B** Contour Plot*

Note that we can not say that data set **A** is "more variable" than data set **B**, since the standard deviations for the two data sets are the same, as are the magnitudes of highs and lows.

Variogram Represent

- The visually apparent difference between these two data sets is one of texture and not variability.
- In particular, data set **A** changes more rapidly in space than does data set **B.**
- The continuous high zones (red patches) and continuous low zones (blue patches) are, on the average, smaller for data set **A** than for data set **B.**
- Such differences can have a significant impact on sample design, site characterization, and spatial prediction in general.

Variogram Represent

- It is not surprising that the common descriptive statistics and the histograms fail to identify, let alone quantify, the textural difference between these two example data sets.
- Common descriptive statistics and histograms do not incorporate the spatial locations of data into their defining computations.

Variogram Represent

- The variogram is a quantitative descriptive statistic that can be graphically represented in a manner which characterizes the spatial continuity (i.e. roughness) of a data set.
- The variograms for these two data sets are shown in Figures 1.5 and 1.6. The difference in the initial slope of the curves is apparent.

Figures 1.5 and 1.6.

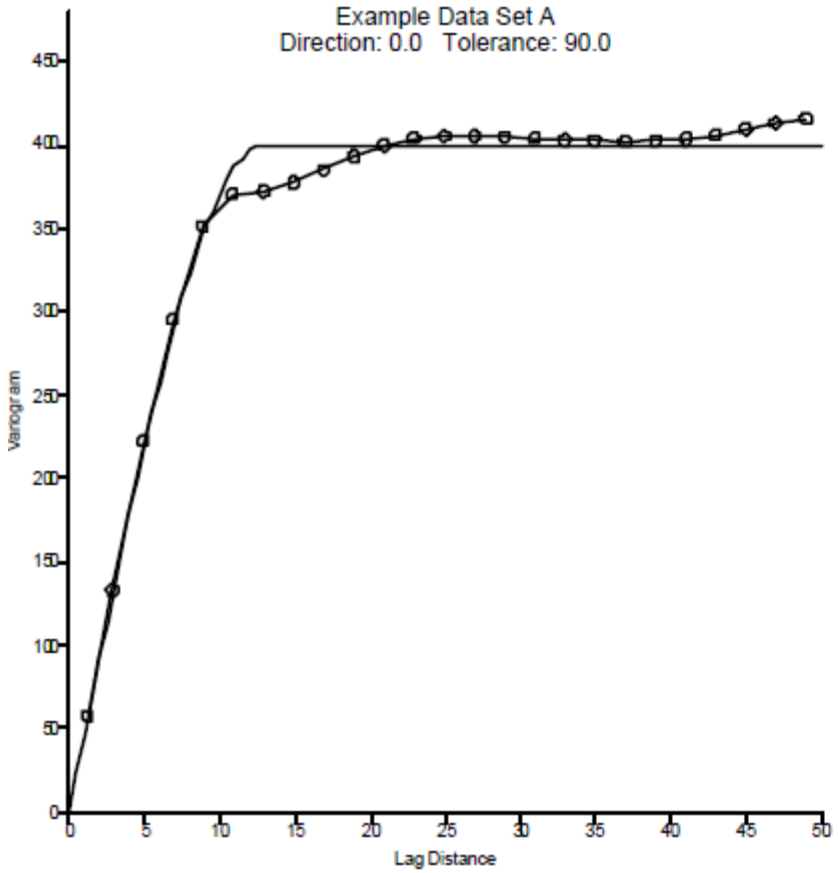


*Figure 1.5 Data Set **A***
Variogram and Model

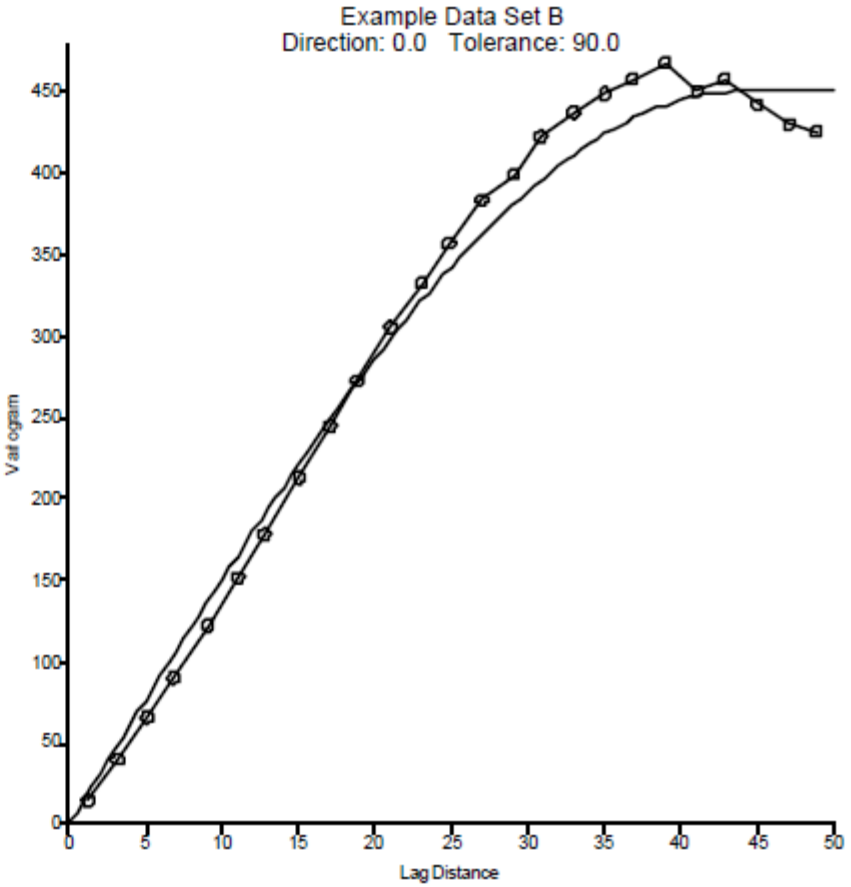


*Figure 1.6 Data Set **B***
Variogram and Model

Variogram Modelling

Three common variogram models are the exponential, Gaussian, and spherical:

Exponential: $C(h) = \sigma^2 e^{-h/r}$

Gaussian: $C(h) = \sigma^2 e^{-(h/r)^2}$

Spherical: $C(h) = \sigma^2 \left(1 - \frac{3h}{2r} + \frac{h^3}{2r^3}\right), h < r$

where σ^2 is the variance of the process and ***r*** is a constant.

Note that the range for the exponential and Gaussian models is infinite, while for the spherical model it is equal to the constant ***r***.

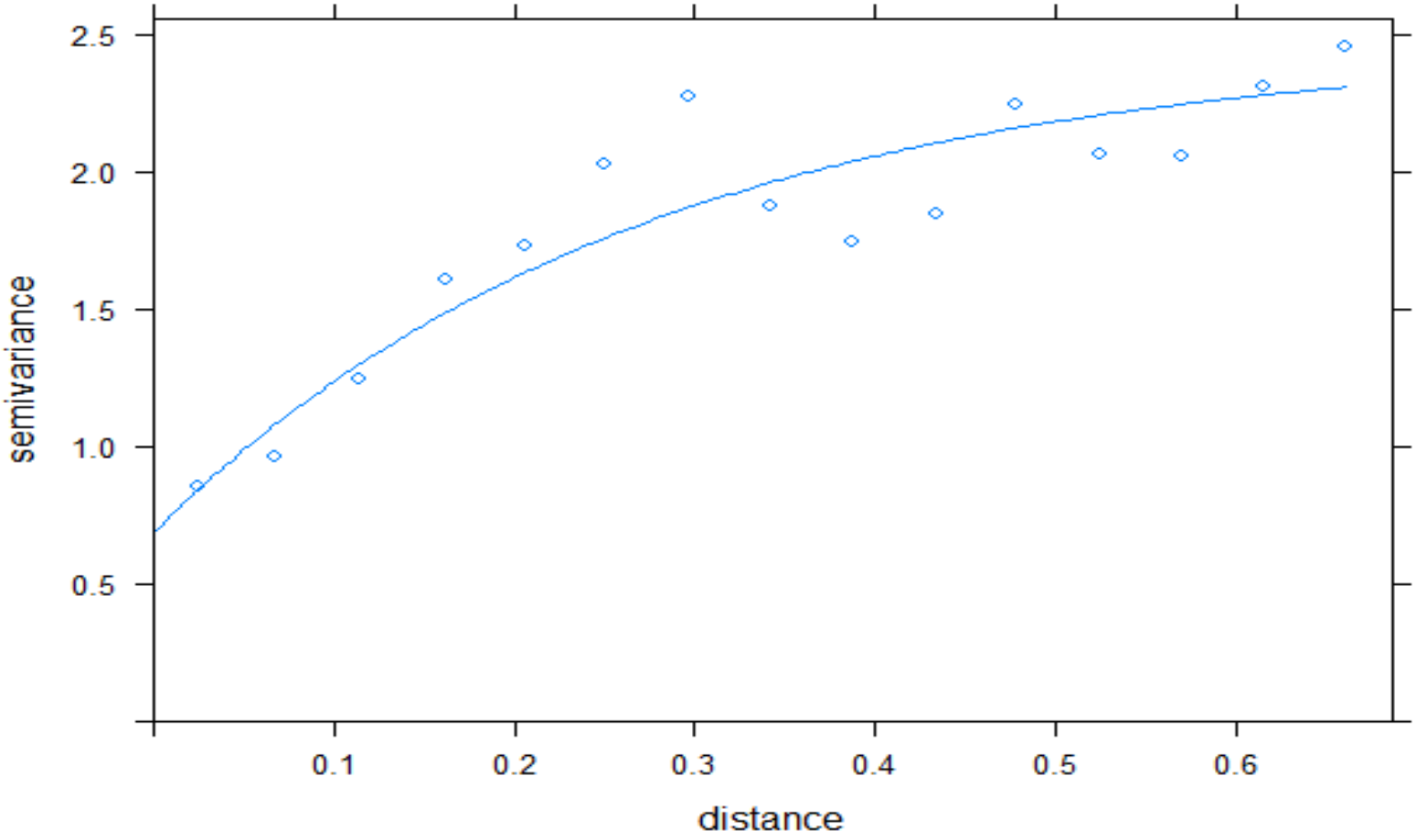
Example 11.2

Fitting variogram models to the benthic data:

```
library(sp)

library(EnvStats)

library(gstat)

data(Benthic.df)

coordinates(Benthic.df)=~Longitude+Latitude

vg.benthic <- variogram(Index ~ 1, data=Benthic.df)

vg.fit.benthic<-fit.variogram(vg.benthic, model=vgm(1,"Exp", 0.5,1))

plot(vg.benthic, vg.fit.benthic,main="Figure 11.2  Exponential variogram f
or Benthic Index")

vg.fit.benthic<-fit.variogram(vg.benthic, model=vgm(1,"Gau", 0.5,1))

plot(vg.benthic, vg.fit.benthic,main="Figure 11.3  Gaussian variogram for
Benthic Index")

vg.fit.benthic<-fit.variogram(vg.benthic, model=vgm(1,"Sph", 0.5,1))

plot(vg.benthic, vg.fit.benthic,main="Figure 11.4  Spherical variogram for
Benthic Index")
```

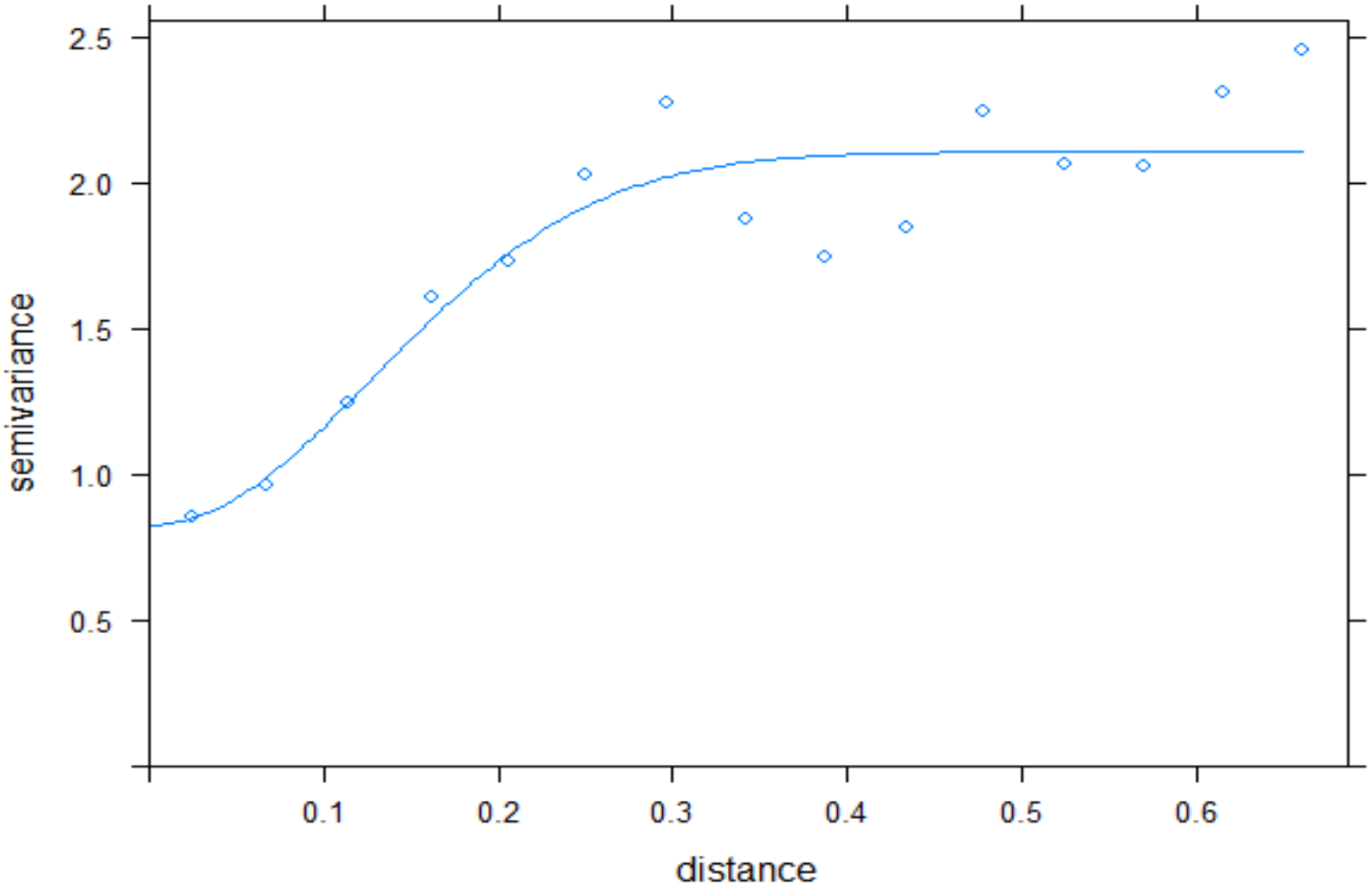
Example 11.2

Figure 11.2 Exponential variogram for Benthic Index



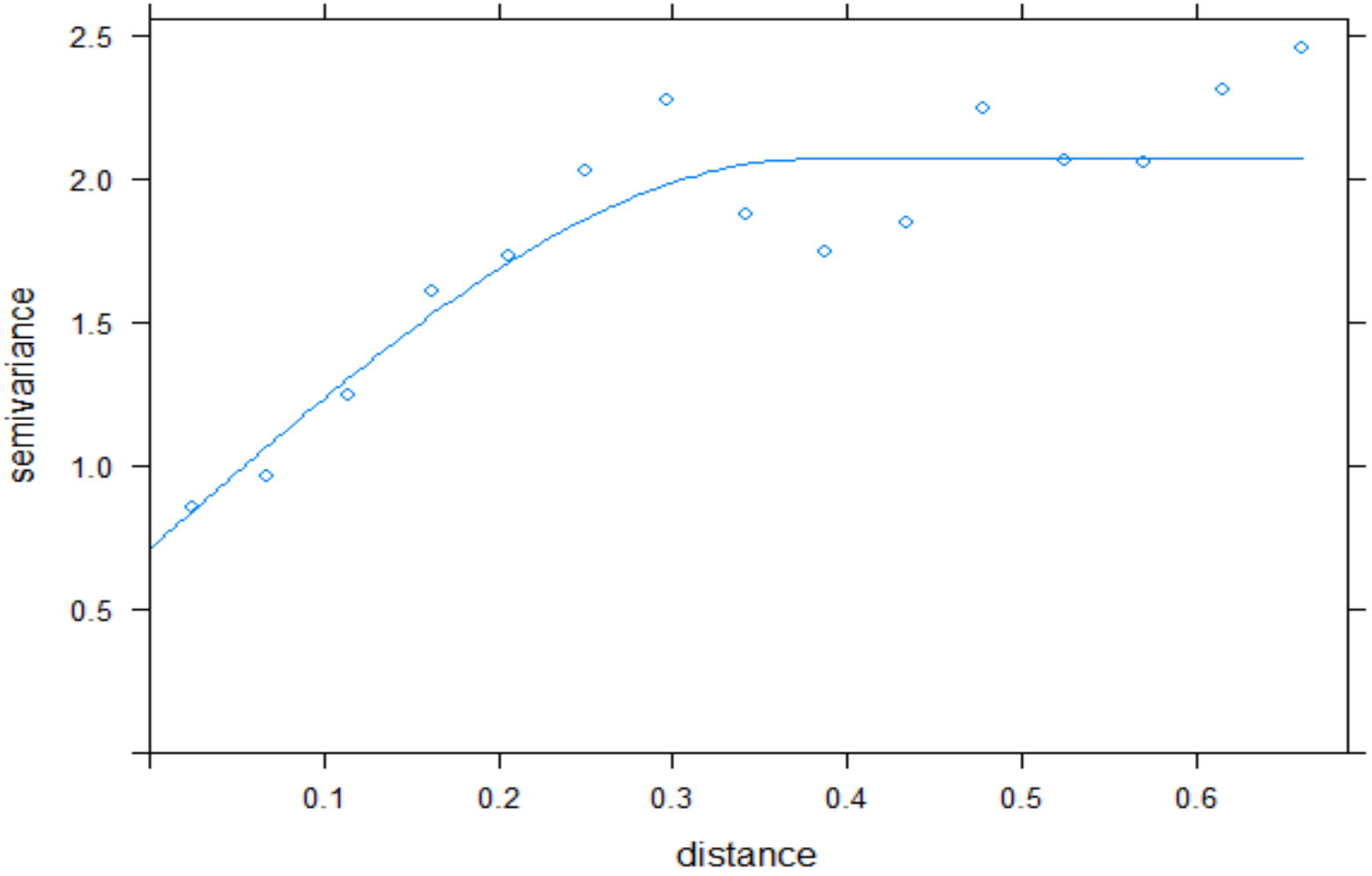
Example 11.2

Figure 11.3 Gaussian variogram for Benthic Index



Example 11.2

Figure 11.4 Spherical variogram for Benthic Index



The Variogram Grid

- If there are n observed data, there are **n(n - 1)/2** unique pairs of observations.
- Thus, even a data set of moderate size generates a large number of pairs.
- For example, if **n** = 500, **n(n - 1)/2** = 124,745 pairs.
- The manipulation of such a large number of pairs can be time consuming, even for a fast computer.
- Note: a *variogram grid* is not the same format as a grid used in creating a map.

Directional Variogram

The Variogram Direction

- The variogram measures dissimilarity as a function of separation distance and direction.
- In many physical settings it is possible for the variable of interest to change more rapidly in one direction than in another.
- For example, the distribution of grain size changes more rapidly in a direction perpendicular to the shoreline than it does parallel.

The Variogram Direction

- Similarly, in an arid climate the prevalence of certain species of plants changes more rapidly as one moves in a direction perpendicular to a river than it does as one moves along the river.
- That is **different** behavior in different directions.
- The direction parameter allows you to investigate the variogram in different directions.

Directional Variograms

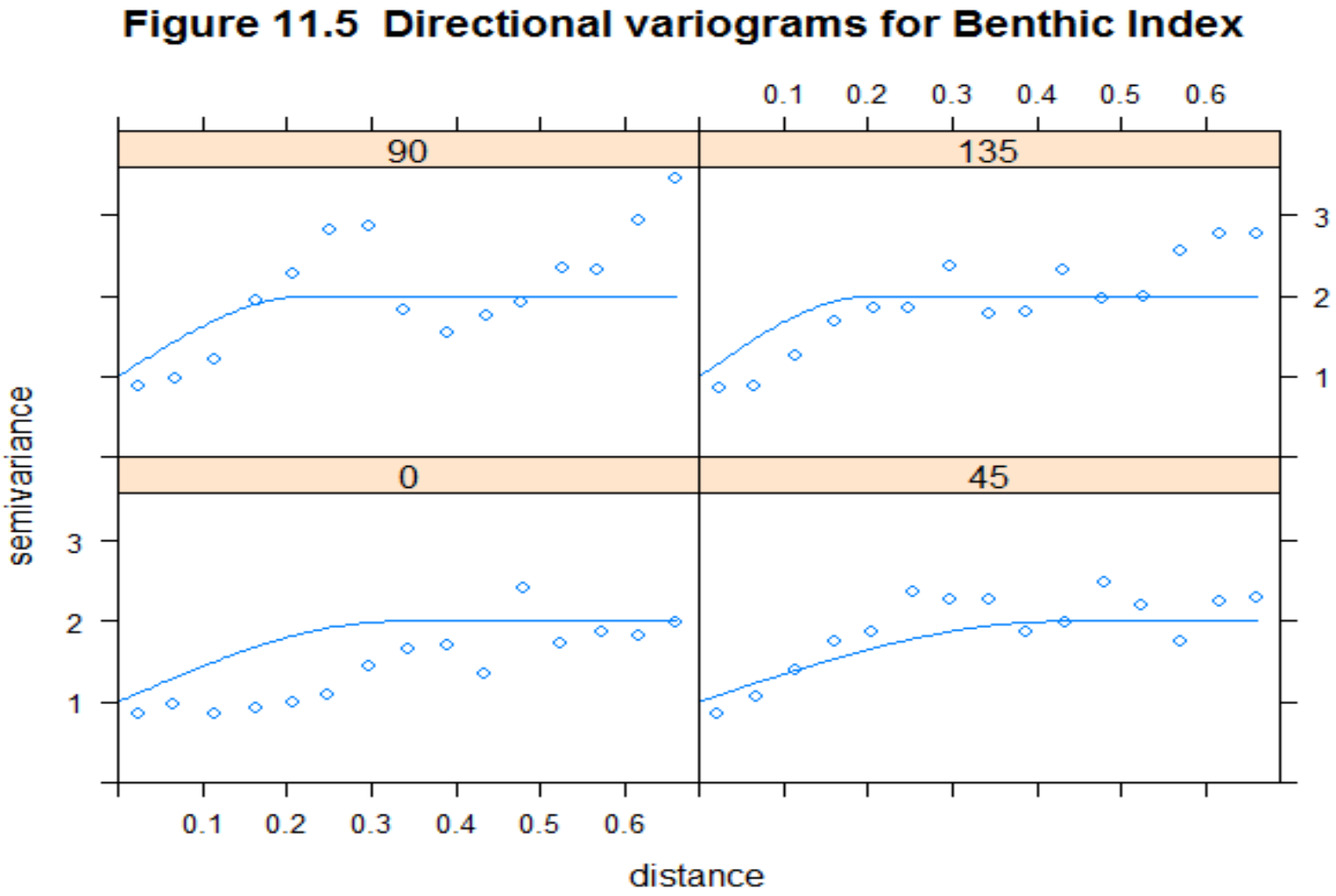
- We have assumed that the spatial process is stationary and that the spatial correlation is isotropic.
- We can use directional variograms explore whether these assumptions appear to be valid for the data.
- Example 11.3 displays the empirical semi-variogram for the benthic index for four difference directions.

Example 11.3

Directional variograms for Benthic Index:

```
library(sp)

library(EnvStats)

library(gstat)

data(Benthic.df)

coordinates(Benthic.df)=~Longitude+Latitude

dvg.benthic <- variogram(Index ~ 1, data=Benthic.df, alpha=c(0, 45,90,13
5))

dvg.fit.benthic<- vgm(1,"Sph", 0.5,1,anis=c(30,0.4))

plot(dvg.benthic,dvg.fit.benthic, main="Figure 11.5  Directional
variograms for Benthic Index")
```

Example 11.3



Note: zero direction is North; 90 degrees is East; and so on.

Directional Variograms

- The form of the variogram appears to be different for different directions based on the raw benthic index data.
- Differences in the variogram in different directions may be caused by the presence of trend and/or anisotropy (different forms of spatial correlation in different directions).

Exercises

11.1 Repeat the examples in this talk.

11.2 Re Example 11.3, obtain variograms with four directions: 22.5, 67.5, 112.5 and 157.5.

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.