Analysis of Spatially Continuous Data

Bailey and Gatrell

Chapter 5

Lecture 13
October 20, 2009

1

Continuous Data Analysis

- Focus is on patterns in the attribute values not locations as in the analysis of point patterns
- The locations are simply sites at which attribute values have been recorded within a region
- Attributes are conceptually spatially continuous. Examples include observations on rainfall, temperature, salinity, air quality variables such as ozone, soil variables such as permeability, conductivity, pH,

2

The Data

- Observations on a spatially stochastic process $\{Y(s), s \in R\}$ that varies continuously over a region R and has been sampled at fixed point locations s_i.
- Referred to as $y(s_i)$ for random variable $Y(s_i)$
- Shortened to y_i or $Y_{(si)}$ for the vector **Y** = ($Y_{(s1)}$,…$Y_{(sn)}$) for the random variable Y
- Often referred to as geostatistical data

3

Analysis Objectives:

- Infer the nature of spatial variation in an attribute over the whole of a region R based on sampled point values.
- Examine first order effects – variations in the mean value of surface (large scale), and second order effects (spatial dependence between values at any 2 locations).
- Model the pattern of variability of an attribute and determine factors that might relate to it
- Obtain predictions of a value at un-sampled locations

4

Continuous Data Analysis

- Develop descriptions that capture global trends as well as local variability

Consider first and second order effects

$$E(Y(s)) = \mu(s) \qquad COV(Y(s_i), Y(s_j))$$

Spatial dependence between $Y(s_i)$ and $Y(s_j)$

Propose models consisting of two components

- First order component – representing large (coarse) scale variation
- Second order component – representing fine scale spatial dependence

5

Visualizing Spatially Continuous Data

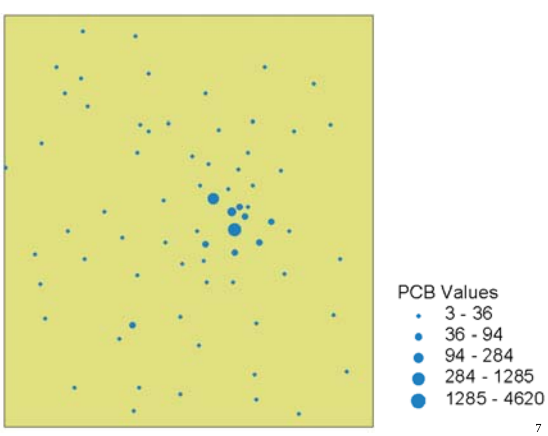
Use symbols that will represent the information on the data values

- Proportional circles or rectangles are often used
- The size of the circle is proportional to the data value

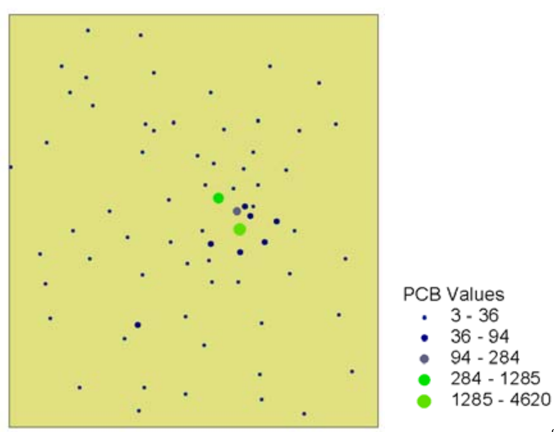
  e.g. radius equal to the square root of data values

- Or height of the rectangle is proportional to the data value
- Colors can be used to reinforce the same data value or add a different variable

6



7



8

Visualizing Spatially Continuous Data

- When mapping symbols to classes the number of classes and the type of class interval “influence the message”

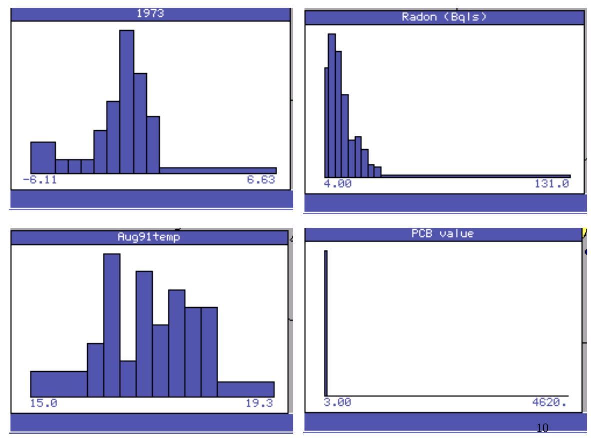
  Larger numbers of data values typically require more classes to cover the range

Rule of thumb

  Number of classes equal to 1+ 3.3 log *n* where *n* is the number of observations

- Start by first examining the distribution of values before selecting class intervals

  For data with very skewed distributions it is useful to transform the data values first.



Visualizing Spatially Continuous Data

- **Equal Intervals**

  The equal interval method divides the range of attribute values into equal sized sub-ranges.

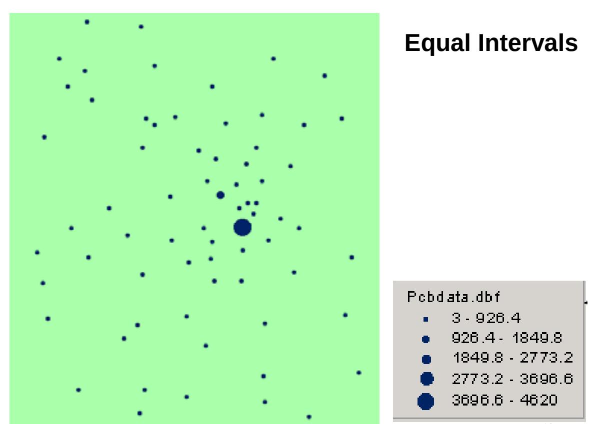
  Good if data values are uniformly distributed over their range

  If data are skewed there will be large number of values in a few classes, and classes with few, if any map features

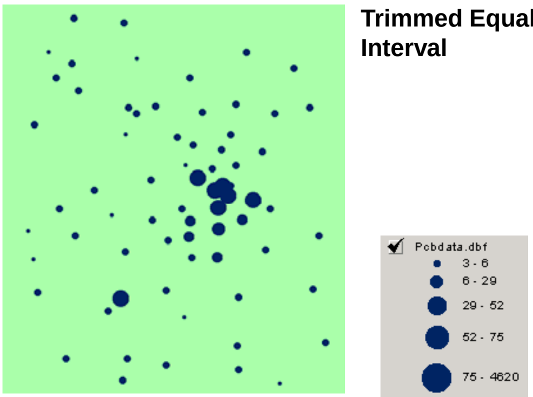
- **Trimmed Equal Intervals**

  Assign top and bottom ten percent to separate classes and equally divide remainder

Equal Intervals



Trimmed Equal Interval



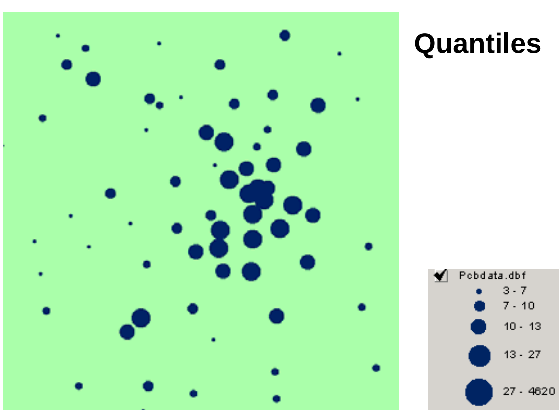
13

Visualizing Spatially Continuous Data

- **Percentiles of the distribution**
  each class contains the same number of features, a good method of classification for evenly distributed data, features with greatly different values can be placed in a single class
- **Standard deviates**
  Class breaks are set above and below the mean at intervals of either 1/4, 1/2, or 1 standard deviations until all the data values are contained within the classes.
- **Natural break intervals**
  Natural breaks find groupings and patterns inherent in the data by minimizing the sum of the variance within each of the classes

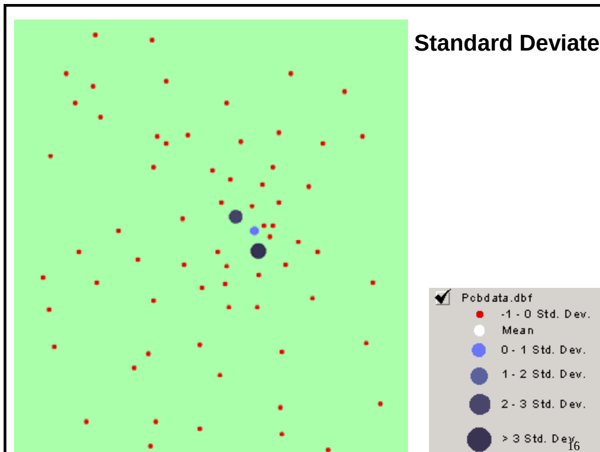
14

Quantiles

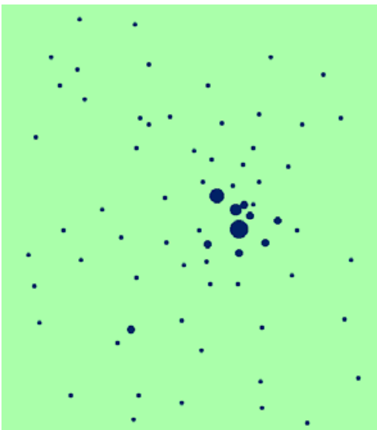


15

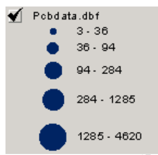
Standard Deviates



16



Natural Breaks



17

Exploring Spatially Continuous Data

- First consider approaches to investigate variation in the mean value $E(Y(s)) = \mu(s)$ over the region
  - Spatial moving averages
  - Interpolation based on tessellations
  - Kernel estimation
- Next consider approaches to investigate second order effect or spatial dependence among values within the region
  - Covariogram, variogram

18

Spatial Moving Averages

- Estimate mean by averaging values at nearby data points

Unweighted average

Average n data values nearest to location s

Weighted average

$$\hat{\mu}(s) = \sum_{i=1}^{n} w_i(s) y_i$$

$\sum w_i(s) = 1$ $\quad w_i(s) \propto h_i^{-\alpha}$ $\quad w_i(s) \propto e^{-\alpha h_i}$

h_i is the distance from s to s_i

α, parameter to change the degree of smoothing

19

Methods Based on Tessellations

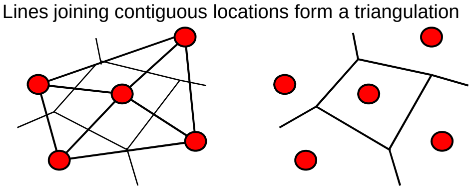
Estimate μ(s) from a tiling of the observed sample locations s_i

Common tessellation is the Delauney Triangulation

n locations in the plane can be assigned a territory closer to the point than to any other

Dirchelet Tessellation, Voronoi, or Theissen polygons

Lines joining contiguous locations form a triangulation



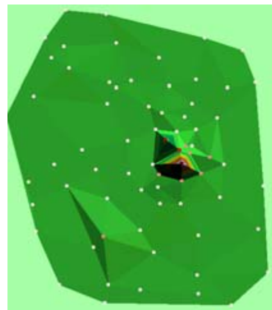
20

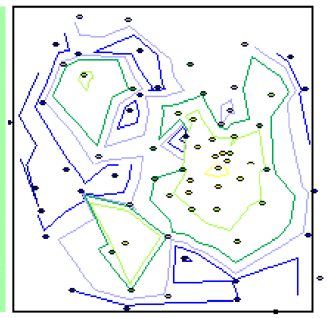
Methods Based on Tessellations

Each triangular face is associated with a function that is used to interpolate values at unknown points within the triangle

Interpolated values are then used to construct isolines

TIN and constructed Isolines





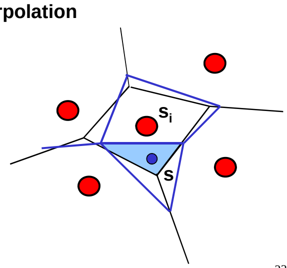
Methods Based on Tessellations

The resulting contours are an exploratory device to examine variation in the mean value

Natural neighborhood interpolation

$$\hat{\mu}(s) = \sum_{i=1}^{n} w_i(s) y_i$$

Weights are now based on areas of Vornoi polygon around s_i "stolen" by the tile around s



Kernel Estimation

$$\hat{\lambda}_\tau(s) = \sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right)$$

We are now interested in an estimate of the mean value for the attribute y_i rather than intensity of events

$$\sum_{i=1}^{n} \frac{1}{\tau^2} k\left(\frac{(s - s_i)}{\tau}\right) y_i$$

Represents the amount of the attribute per unit area

To find the average we need to divide by number of observations per unit area

Kernel Estimation

$$\hat{\mu}_\tau(s) = \frac{\sum_{i=1}^{n} k\left(\frac{(s-s_i)}{\tau}\right) y_i}{\sum_{i=1}^{n} k\left(\frac{(s-s_i)}{\tau}\right)}$$

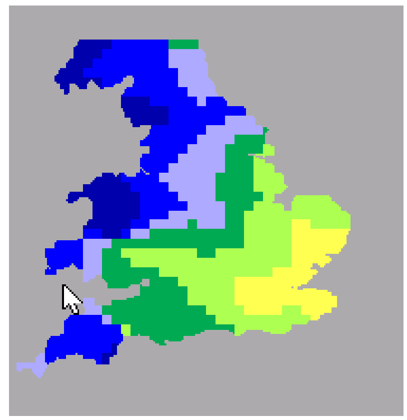
The effect of increasing the bandwidth is to increase the degree of smoothing

Variation on the weighted moving average

$$\sum_{i=1}^{n} w_i(s) y_i$$

$$w_i(s) = \frac{k\left(\frac{(s-s_i)}{\tau}\right)}{\sum_{j=1}^{n} k\left(\frac{(s-s_j)}{\tau}\right)}$$

25



Kernel Estimate for mean August temperature

26

Covariogram and Variogram

- Used to explore the spatial dependence of deviations in attribute values from their mean
- The covariance function is analogous to the K function for analyzing second order properties in point patterns
- In the continuous data case we are interested in the way the deviations of observations from their mean values co-vary over the region
- In most cases we expect positive covariance or correlation at short distances for spatially continuous phenomena

27

Covariogram and Variogram

Assume we have a spatial stochastic process $\{Y(s), s \in R\}$

$$E(Y(s)) as\ \mu(s) \qquad VAR(Y(s))\ \ as\ \sigma^2(s)$$

The covariance of the process at any two points s_i and s_j

$$C(s_i, s_j) = E((Y(s_i) - \mu(s_i))(Y(s_j) - \mu(s_j)))$$

Correlation is:

$$\rho(s_i, s_j) = \frac{C(s_i, s_j)}{\sigma(s_i)\sigma(s_j)}$$

Variance

$$C(s, s) = \sigma^2(s)$$

28

Covariogram and Variogram

The process is stationary if: $\mu(s) = \mu$ $\sigma^2(s) = \sigma^2$

Mean and variance are independent of location and constant throughout the region

and $$C(s_i, s_j) = C(s_i - s_j) = C(h)$$

Covariance depends only on the vector difference ***h***

$C(h)$ Referred to as the **covariogram** or covariance function of the process

$\rho(h)$ Referred to as the **correlogram**

$C(0) = \sigma^2$

29

Covariogram and Variogram

The process is *isotropic* if the dependence is only a function of distance and not direction

That is dependent only on the length of the vector **h**

Then $$C(s_i, s_j) = C(h)$$

and $$\rho(s_i, s_j) = \rho(h)$$

30

Covariogram and Variogram

Intrinsic Stationarity

A weaker assumption of stationarity

There is a constant mean and constant variance in the differences between values at locations separated by a given distance and direction

$$E(Y(s+h) - Y(s)) = 0$$

$$VAR(Y(s+h) - Y(s)) = 2\gamma(h)$$

$\gamma(h)$ is strictly the semivariogram but commonly referred to as the **variogram**

31

Covariogram and Variogram

For stationary processes the covariogram, correlogram and variogram are related

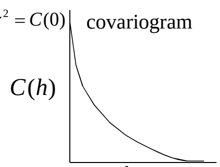
$$\rho(h) = \frac{C(h)}{\sigma^2}$$

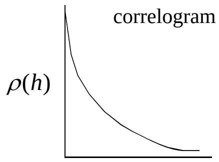
$$\gamma(h) = \sigma^2 - C(h)$$

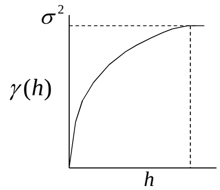
32

Covariogram and Variogram

$\sigma^2 = C(0)$

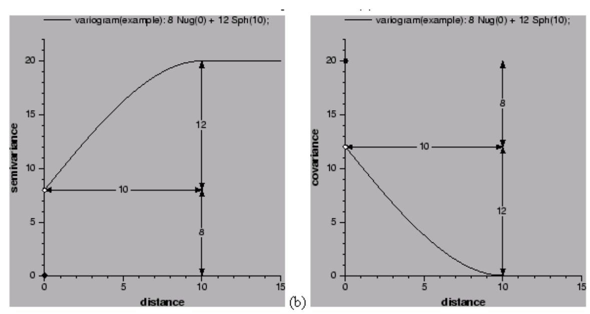






$$\rho(h) = \frac{C(h)}{\sigma^2}$$

$$\gamma(h) = \sigma^2 - C(h)$$



Variogram stops increasing beyond a certain distance and becomes stable at a limit value $\gamma(\infty)$ – called **Sill** $\gamma(\infty) = Var\{Y(s)\} = C(0)$

Range – distance at which this occurs - transition from state in which spatial correlation exists to absence of correlation

Nugget – value of $\gamma(h)$ at h = 0 – due to measurement error or micro-scale variation

Exploratory Analysis Steps

Estimate an isotropic variogram or covariogram

Estimate directional variograms in 2 or 3 broad directions to examine directional effects - anisotrophy

A general exploratory technique is the variogram cloud

Estimation of the Variogram

$$2\hat{\gamma}(h) = \frac{1}{n(h)} \sum_{s_i - s_j = h} \left(y_i - y_j\right)^2$$

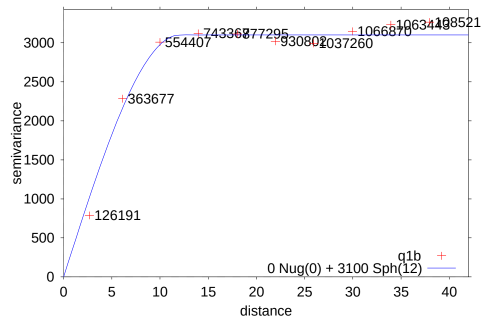
Sum squared differences over all pairs less than or equal to a distance *h*

Assuming isotropy, variogram is estimated over all directions for a given separation distance *h*

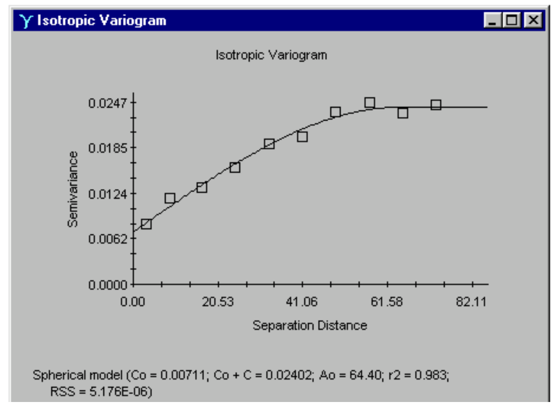
Estimation of the Covariogram

$$\hat{C}(h) = \frac{1}{n(h)} \sum_{s_i - s_j = h} \left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)$$

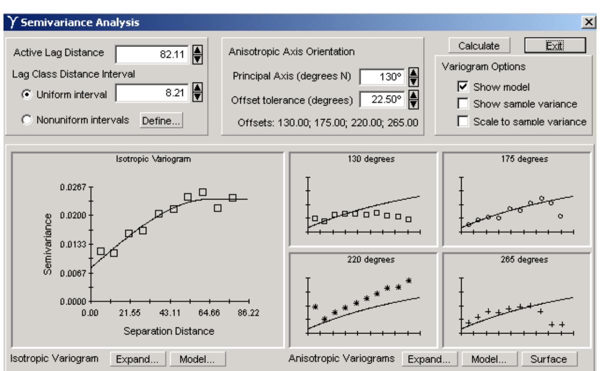
Estimation of the Variogram



n(h) varies as h increases – so the reliability of γ estimates vary



Anisotrophic Variograms



Variogram Cloud

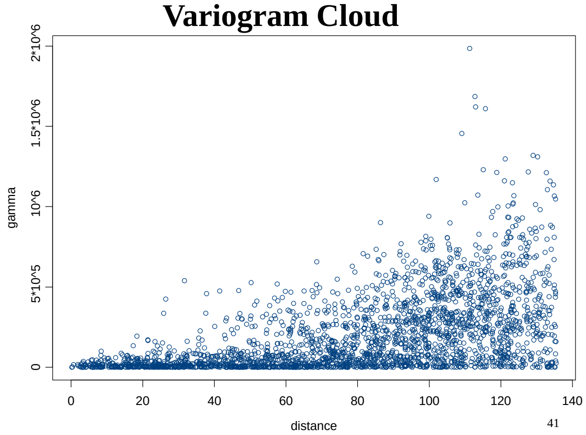
The variogram cloud is the distribution of the variance between all pairs of points at all possible distances (h).

Plot of $(y_i - y_j)^2$ against h ($s_i - s_j$)

or

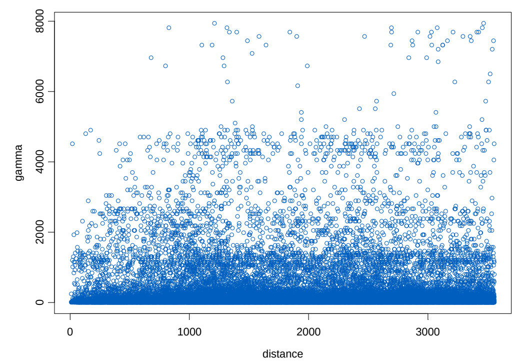
Plot of $\sqrt{(|y_i - y_j|)/2}$ against h

Variogram Cloud



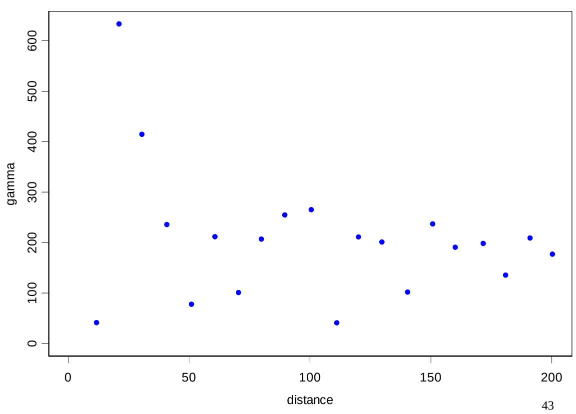
41

Variogram Cloud



42

Radon Data



43