Lecture 9: Raster Data and Introduction to Spatial Dependence

Raster Data

• Easiest to think of it as a matrix of values in which each number represents a color shade for a square on a map – though entries can be missing

• A raster data set may have multiple "bands" that hold information on different things though usually they come with only one band

• Raster data exist on most observable things that are continuous across space

  • rainfall
  • soil quality
  • average january temperature
  • land use (water, forest, farmland, built up, transportation, etc.)
  • elevation

• Individual pixels can have information about attributes as well

• There are many raster data formats but ArcGrid is most common

• Most picture formats work fine too – ArcMap will separate them into bands for you

Basics of Raster Data

• Unlike vector data, which has coordinate points for various vertices or points, raster data sets have information for every single grid point in a region

  • If they are at all detailed, raster data sets are very large!! Most common is 1 second (about 30X30m) resolution

  • Spatial location information in the file includes parameters defining the edges of a rectangle and cell size plus projection

• USGS is a clearinghouse for most of the best raster data for the US, and some other parts of the world as well

  • http://seamless.usgs.gov/index.php -- You can interactively choose what layers to download

  • http://sedac.ciesin.columbia.edu/gpw/index.jsp has population density data worldwide

  • Many other available sources

• Symbolizing the data is similar to how you deal with feature data except not that you can make the data show partly transparent in the Display tab

Basic Tools for Raster Data

• Converting to and from Rasters

  • Conversion Tools in the toolbox

  • Can convert categorical rasters to polygons and all rasters to points

  • Interpolate raster values using a provided field for point and polygon vectors converted to rasters

• Exporting Raster Subsets, etc.

  • Data Management – Raster – Raster Processing – Clip lets you clip a raster by a given rectangle

  • Conversion Tools – From Raster – Raster to ASCII lets you get an ascii file of the Raster grid

Doing Calculations With Raster Data

• The Spatial Analyst section of the toolbox is full of tools that let you do many calculations on raster data usually to create new raster data sets

• You can do many of these things using the Spatial Analyst toolbar which you can add by selecting View – Toolbars in ArcMap

• Raster Calculator – Perform many calculations on raster data

  • Many of the tools in the Spatial Analyst tool box do the same functions: Conditional, Overlay, Math

• Distance – Create various distance measures from empty grid points to closest grid point known as a "source"

  • Example: Want to create a raster of distance to nearest transit station from each point in the view to make a nice map. Generally you would have the source in vector format, but the result is a new raster data set.

• Density - Take a feature data set and use it to calculate a raster that is density (smoothed or not)

Doing Calculations With Raster Data

• Surface Analysis

  • Create vector datasets using something about rasters

• Cell Statistics – Do some standard calculation using multiple raster data sets

• Neighborhood Statistics – Do some standard calculation but smoothing over a larger neighborhood rather than cell by cell

• Zonal Statistics – Calculate some statistic from your raster data for each "zone". Zone can be given by a categorical raster data or a feature dataset

  • Calculate average elevation in each census tract

  • Calculate standard deviation of elevation for each type of land cover

  • Calculate fraction of land in each census tract that is built up

Interpolating Surfaces

• This is about creating raster data from point data

  • Weather stations
  • Soil samples
  • Hydrological measurements

• Basic assumption is that there is some spatial dependence in the data – otherwise the information at location (x,y) would not be useful for inferring anything about location (x+e,y+u) where e and u are small.

• Spatial Analyst Toolbar – Interpolate to Raster

  • Inverse Distance Weight – based on point observations z_i, interpolate z(x)

$$z(x) = \frac{\sum_i w_i z_i}{\sum_i w_i} \text{ commonly } w_i = \frac{1}{d^2{}_i}$$

  • Kriging - Spatial interpolation by an error minimization criterion

Spatially Dependent Data

• Standard statistical theory is set up to handle independent and identically distributed ("iid") data

• This means that the underlying assumption is that each data point is independently drawn from the same distribution

• Analysis of spatial data generally assumes that data is not independently drawn – and in particular that there is more information about me from my neighbor than from someone far away – or that there is some sort of distance decay in the dependence of the data

• Often when observations are dependent, you cannot make as precise inference about the estimate you care about as if they were independent

• Generally, in order to have any statistical power with spatial data, we need to make a promise that this dependence goes towards 0 as observations get arbitrarily far apart.

Example

• Suppose you want to calculate the mean number of bedrooms of housing units in Chicago. But you sample 100 units all from the same building of 1 bedroom apartments.

• Suppose in addition for the sake of argument that each apartment in a given building always has the same number of units

  • Classical statistics would tell you that your estimate is 1 with a standard deviation of 0

  Estimate of mean = $\frac{1}{N}\sum_{i=1}^{N} x_i$

  Standard deviation of estimated mean = $\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2}$

  • But in reality, this is a case in which we have only one real observation, because if we know the number of bedrooms in for one of the apartments in the building, we know them all. Therefore while data point 1 is informative, data points 2-100 have no information in them at all. Therefore, in some sense the true N=1 and true standard deviation of the estimated mean is infinity.

Measuring Spatial Dependence

• First, it's important to determine how you want to capture distance in your data

  • Adjacency – most often used for polygon data -- requires calculation of an NXN adjacency matrix showing which observation is next to which other observations

  • Continuous – most often used for point data – requires calculation of pairwise distances between data points

• Nonparametric Covariance Function

  • Suppose you have data on a lattice for all observations i,j at distance d_{ij}

  • Equals the variance at distance 0

$$C(d_{ij}) = \frac{1}{N_{ij}}\sum_i\sum_j (x_i-\bar{x})(x_j-\bar{x})$$