Lecture 10: Basics of Spatial Statistics

Calculating a Confidence Interval for an Estimate of a Mean

• Suppose that you have data on outcome Z with data points in space

• if data is independent and identically distributed (iid)

$$\overline{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i \text{ and}$$

$$\mathrm{V}(\overline{Z}) = \frac{1}{n^2}\sum_{i=1}^{n} V(Z_i) = \frac{n}{n^2} V(Z_i) = \frac{{\sigma_0}^2}{n}$$

so 95% confidence interval is

$$(\overline{Z} - 1.96\frac{\sigma_0}{\sqrt{\mathrm{n}}}, \overline{Z} + 1.96\frac{\sigma_0}{\sqrt{\mathrm{n}}})$$

Calculating a Confidence Interval for an Estimate of a Mean

• However, if there is some positive spatial dependence in the data such that

$$\mathrm{cov}(Z_i, Z_j) = {\sigma_0}^2 \rho^{d_{ij}}, 0 < \rho < 1, \mathrm{d_{ij}} = \text{distance}$$

• Observations are dependent with the dependence decaying with distance. "Distance" doesn't have to be in the traditional sense – could be on a network

• Now the calculation of the variance of the estimated mean is more complicated

$$\mathrm{V}(\overline{Z}) = \frac{1}{n^2}\mathrm{V}(Z_1 + Z_2 + ... + Z_N)$$

$$= \frac{1}{n^2} E[(\sum_{i=1}^{n} Z_i)^2] - [E[\sum_{i=1}^{n} Z_i]]^2$$

$$= \frac{1}{n^2} E[(\sum_{i=1}^{n}\sum_{i=j}^{n} Z_i Z_j] - [E[\sum_{i=1}^{n} Z_i]]^2$$

$$= \frac{1}{n^2} (\sum_{i=1}^{n}\sum_{i=j}^{n} E[Z_i Z_j] - [\sum_{i=1}^{n} E[Z_i]]^2$$

Calculating a Confidence Interval for an Estimate of a Mean

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i=j}^{n} (E[Z_i Z_j] - E[Z_i]E[Z_j])$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{cov}(Z_i, Z_j)$$

$$= \frac{{\sigma_0}^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \rho^{d_{ij}} > \frac{{\sigma_0}^2}{n}$$

• Therefore the confidence interval is wider because of the dependence in the data

• Notice that $n / \sum_{i=1}^{n}\sum_{j=1}^{n} \rho^{d_{ij}}$ is the equivalent number of independent observations you would need to achieve the same precision of the estimate

Measuring Spatial Dependence

• The problem with the previous example is that we don't know what ρ is. In general the spatial dependence is something that has to be estimated.

• We might care about this estimate in its own right

  • Is crime spatially correlated? If so, knowing how much might help us understand sources of crime

  • Is gentrificaction spatially correlated?

• Or it might be an input to helping us do a statistical task better

  • Fixing standard errors in regression analysis

  • Prediction for interpolation over space (Kriging) or density estimation

Moran's I

• Probably the most commonly used measure of spatial dependence

• Have spatial data on outcomes z_i

• Weights matrix W with elements w_{ij}

  • $w_{ii} = 0$ always
  • Sometimes will use an adjacency matrix such that $w_{ij}=1$ if observations i and j are adjacent, 0 otherwise
  • In all cases, w_{ij} declines in distance between observations i and j
    • $w_{ij} = a/d_{ij}$ and $w_{ij} = \exp(-ad_{ij})$ are common

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(z_i - \bar{z})(z_j - \bar{z})}{\sum_i (z_i - \bar{z})^2}$$

• Can think of this as a spatial covariance function that has been weighted to account for closeness of observations and adjusted for overall variance in the data

Moran's I

• When you estimate Moran's I, you get a "Z score" which tells you whether there is statistically significant positive or negative spatial autocorrelation in your data

• This is relative to the 0 autocorrelation value which is technically

$$I_0 = \frac{-1}{n-1}$$

• And just matches up to a cumulative normal

  • Z< -1.96 means statistically negative spatial autocorrelation

  • Z > 1.96 means statistically positive spatial autocorrelation

• Note the assumptions going into the calculation of Moran's I. In particular, it is not clear at all how to determine the weights matrix

• In ArcToolbox, calculate Moran's I in Spatial Statistics – Analyzing Patterns – Spatial Autocorrelation

• In Stata, install spatial package by typing findit spatgsa

Spatial Stationarity

• As came up with calculation of Moran's I and many other spatial statistical problems, we want a systematic way of forming the weights matrix

  • For some problems simple adjacency makes sense

  • For many problems we want a continuous weights matrix
    • Set weight > 0 if this is a distance at which observations are correlated

• Generally we want a weights matrix such that the weight between any two observations distance d apart is the same. That is, we want to apply our statistical analysis to a data set that is **stationary** and **ergodic**

  • Distribution function looks the same at every location

  • E[Z(s)] = E[Z(s+h)] and V[Z(s)] = V[Z(s+h)]

  • Additionally, we also generally want C[Z(s),Z(s+h)] = C[Z(m),Z(m+h)], or Covariance Stationarity

Spatial Covariance Functions

• To fill in that weights matrix, we need to know about the covariance function. For thick data on a lattice, we can estimate it as follows

$$C(d) = \frac{1}{N_{ij}} \sum_{i,j,d_{ij}=d} (z_i - \bar{z})(z_j - \bar{z})$$

• However, in practice we may have point data that is not on a lattice. To deal with this we need some way to smooth the data so that data points of distance s and s+e both contribute to the estimation of the covariance function for distance s.

• Typical method is to use a "kernel" to estimate the covariance nonparametrically

• Common is to use a flat kernel and to calculate covariance using the above formula for each distance range of a given bandwidth (0-1km, 0.1-1.1km, 0.2km-1.2km, etc.)

• From this we learn over what distance range the data is spatially dependent

Details of Kriging

• The purpose of Kriging is to interpolate points (often to create a raster data set) when we only have observations at a few points in space

$Z(s) = \mu + \delta(s)$ data is for known spatial locations s

$p[Z(s_{unobs})] = \sum_i \lambda_i Z(s_i)$ predictor is a weighted average

Problem is to calculate the weights λ_i on all other observations

Goal : Minimize mean squared prediction error at new location s_0

such that weights sum to 1

$$\min_{\lambda} E[Z(B) - p(Z(B))]^2$$

$$\text{s.t.} \sum_i \lambda_i = 1$$

• This is just a boring math problem – but what is cool is that it has an analytical solution that depends heavily on something called the **variagram**

Variagrams

• A Variagram that tells us the difference in the variance as a function of distance. It is closely related to the spatial covariance function

$$2\gamma(d) = \frac{1}{N_{ij}} \sum_{i,j,d_{ij}=d} [z_i - z_j]^2$$

• In practice, this variagram will usually have to be estimated using some sort of kernel technique like the spatial covariance function above

Solution

• The solution to the above problem turns out to be

$$\lambda_0 = \Gamma_0^{-1} \gamma_0 \text{ where}$$

$$\lambda_0 = [\lambda_1, \lambda_2, ..., m]'$$

$$\gamma_0 = [\gamma(s_0 - s_1), \gamma(s_0 - s_2), ..., 1]'$$

$$\Gamma_0 = \begin{bmatrix} 0 & \gamma(s_1 - s_2) & ... & 1 \\ \gamma(s_2 - s_1) & 0 & ... & 1 \\ ... & ... & ... & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

• Kriging takes into account both proximity and can be organized to take into account direction as well, depending on how you define locations s

• There is also a standard error that you can put on the predicted value at each point that is an ugly formula