Spatial Statistics
Areal Data Unit 3

PM569 Spatial Statistics

Lecture 7: October 21, 2016

Areal Data

- We have examined neighbourhood relationships and defined weights applied to areal units
- Weights connect our areal units in a particular way
- We have examined evidence of global and local spatial (auto)correlation in the areal units through Moran's I, Geary's c, Local Moran's I and the Getis-Ord G* statistics
- We now move more formally to fitting spatial regression models for areal data, most commonly called autoregressive models
- Simultaneous and Conditional Auto-Regressive models (SAR and CAR)

Areal Data

- Recall from geostatistics $E[Z(s)] = \mu$, so we can define the spatial process as a Gaussian process, $Z(s) \sim N(\mu, \Sigma(\theta))$.
- $\Sigma(\theta) = \tau^2 I + \sigma^2 C(\theta; h)$
- We can let the mean depend on a set of covariates:

$$Z(s) \sim N(X\beta, \tau^2 I + \sigma^2 C(\theta; h))$$

- In terms of areal data, we typically use Y(s) to represent the response variable of interest as multivariate normal with mean $\mu = X\beta$ and variance-covariance matrix Σ

Areal Data: Autoregressive Models

Autoregressive Models

- Similar to universal kriging or regression kriging.
- Use regression on values from neighbouring areal units to account for spatial dependence.
- Autocorrelation reflects self regression where you use observations of the outcome at other locations as additional covariates in the model.
- $Y(s) \sim MVN(X\beta, \Sigma)$
- In universal kriging we modeled Σ as a parametric function of distance.
- In areal modeling, we restrict distances to those between our areal units.
- In R we use the spdep library and spautolm function to fit autoregressive models.

Areal Data: Simultaneous Autoregressive Models

- Given that we use a regression on the values from other areas to account for spatial dependence, we model the error terms $\epsilon(s_i)$

$$\epsilon(s_i) = \sum_{j=1}^{n} b_{ij}\epsilon(s_j) + \nu(s_i)$$

- Essentially, we regress the $\epsilon(s_i)$ on all other residual terms
- Recall $b_{ii}=0$
- Here, $\nu(s_i)$ are the "residual residuals" and $\nu(s_i) \sim N(0, \sigma^2 I)$
- The b_{ij} are used to represent the spatial dependence between areas

Areal Data: Simultaneous Autoregressive Models

- We represent Σ as our residual errors $\epsilon(s_i) = \sum_{j=1}^{n} b_{ij}\epsilon(s_j) + \nu(s_i)$ and apply spatial correlation to these residuals through b_{ij}

$$Y(s_i) = x(s_i)\beta + \sum_{j=1}^{n} b_{ij}\epsilon(s_j) + \nu(s_i)$$

$$Y(s_i) = x(s_i)\beta + \sum_{j=1}^{n} b_{ij}[Y(s_j) - x(s_j)\beta] + \nu(s_i)$$

Areal Data: Simultaneous Autoregressive Models

- The degree of spatial dependence is through the term $\sum_{j=1}^{n} b_{ij}[Y(s_j) - x(s_j)\beta]$
- This is a weighted sum of the deviation of the jth observation from its modeled mean value
- Controlling for the covariate effects within areal unit i
- If we think of it another way, we can examine covariate effects while controlling for spatial dependence

Areal Data: Simultaneous Autoregressive Models

- SAR models are often represented in matrix form
- From the equation on the previous slide,

$$Y = X^T\beta + B(Y - X^T\beta) + \nu$$

$$(I - B)(Y - X^T\beta) = \nu$$

- B (n x n) contains the spatial dependence parameters b_{ij}, I is the identity matrix
- For the SAR model to be defined, the matrix I-B must be non-singular (i.e. matrix must have an inverse, if a square matrix the determinant is non-zero)

Areal Data: Simultaneous Autoregressive Models

- $Y(s) \sim MVN(X\beta, \Sigma)$
- Where $E[Y] = X^T\beta$
- Var[Y]=$(I - B)^{-1}\Sigma_\nu(I - B^T)^{-1}$
- Since $\Sigma_\nu = diag[\sigma_1^2, ..., \sigma_n^2]$
- We can simplify to $Var[Y] = \sigma^2(I - B)^{-1}(I - B^T)^{-1}$
- From this we have the variance-covariance matrix of $\mathbf{Y}$, namely Σ_Y

Areal Data: Simultaneous Autoregressive Models

- The matrix of spatial dependence parameters, B is clearly important
- We have too many parameters if we include all b_{ij}
- For estimation and inference we reduce the number of spatial dependence parameters through a parametric model for the b_{ij}
- We use our weight matrices of proximity for this purpose

Areal Data: Simultaneous Autoregressive Models

- We can define our matrix of spatial dependence parameters, $B = \rho W$
- Where W is our weights matrix as before and ρ is a spatial autocorrelation parameter
- $Var[Y] = \sigma^2(I - \rho W)^{-1}(I - \rho W^T)^{-1}$

$$Y(s_i) = x(s_i)\beta + \rho \sum_j w_{ij}[Y(s_j) - x(s_j)\beta] + \nu(s_i)$$

- There are i=1,...,N regions and j is a neighbour of region i
- Again in matrix form, SAR models will be written as

$$Y = X\beta + \epsilon$$
$$\epsilon = \rho W\epsilon + \nu$$

Areal Data: Simultaneous Autoregressive Models

$$Y = X\beta + \epsilon$$
$$\epsilon = \rho W\epsilon + \nu$$

$$\begin{aligned} Y &= X\beta + (I - \rho W)^{-1}\nu \\ &= X\beta - \rho WX\beta + \rho WY + \nu \end{aligned}$$

- We can see where the spatial lags come into play with $\rho WX\beta$ and ρWY
- $\rho WX\beta$ represents the spatial lag in the covariates (neighbouring covariates affect the outcome in region i)
- ρWY represents spatial lag in the outcome (neighbouring observations impact the outcome in region i)

Areal Data: Simultaneous Autoregressive Models

$$\begin{aligned} Y &= X\beta + (I - \rho W)^{-1}\nu \\ &= X\beta - \rho WX\beta + \rho WY + \nu \end{aligned}$$

- The term $(I - \rho W)^{-1}\nu$ is where a regular linear model and an autoregressive linear model differ

Areal Data: Simultaneous Autoregressive Models

- For the SAR model to exist, the matrix $(I - \rho W)$ must be invertible (non-singular)
- So we ensure $\det[I - \rho W] \neq 0$
- This is achieved by imposing conditions on W and ρ
- Decompose W into eigenvalues/eigenvectors
- Eigenvalues $\omega_{max} > 0$ and $\omega_{min} < 0$ gives $1/\omega_{min} < \rho < 1/\omega_{max}$

Areal Data: Simultaneous Autoregressive Models

- Let W=
$$\begin{bmatrix} -1 & 2 \\ 0 & 3 \end{bmatrix}$$
- W is a nxn matrix satisfying $Wx = \omega x$
- Eigenvalues are the scalars ω and vectors x are the eigenvectors
- $(W - \omega I)x = 0$, meaning x is the eigenvector associated with eigenvalue ω
- To get the eigenvalues, solve $\det[W - \omega I] = 0$

$$\begin{vmatrix} -1-\omega & 2 \\ 0 & 3-\omega \end{vmatrix}$$

- $(-1 - \omega)(3 - \omega) - 0$, so $(\omega + 1)(\omega - 3)$ and thus ω =-1 or 3

Areal Data: Simultaneous Autoregressive Models

- Estimation is analogous to what we saw for GLS of geostatistical data
- Recall:
- The probability model for the data is $Y \sim MVN(X\beta, \Sigma)$
- Maximize likelihood
  $L(\beta, \Sigma) = -log|\Sigma|^{1/2} - (1/2)(Y - X\beta)^T\Sigma^{-1}(Y - X\beta)$
- Our covariance matrix $\Sigma_{ij} = Cov(Y(s_i), Y(s_j))$ is spatial as before, with parameters σ^2, τ^2, ϕ
- The covariance matrix Σ dependent upon spatial parameters θ, so we need to estimate $\Sigma(\hat{\theta})$

Areal Data: Simultaneous Autoregressive Models

- In terms of the SAR model,
  $\Sigma_{SAR}(\theta) = \sigma^2(I - B)^{-1}\Sigma_\nu(I - B^T)^{-1}$
- So the θ are the spatial dependence parameters b_{ij} and the parameters of Σ_ν
- As in the geostatistical case, we do not know the spatial parameters, so they must be estimated
- Find $\hat{\beta}$, $\hat{\rho}$ and $\hat{\sigma}^2$ that maximize the likelihood

$$\begin{aligned} \hat{Y} &= X\hat{\beta} - \hat{\rho}WX\hat{\beta} + \hat{\rho}W\hat{Y} \\ &= \hat{\rho}W(Y - X\hat{\beta}) \end{aligned}$$

Areal Data: Simultaneous Autoregressive Models

- We can test the SAR vs OLS regression in terms of testing for $H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$
- Furthermore, we can test for residual spatial autocorrelation by using Moran's I

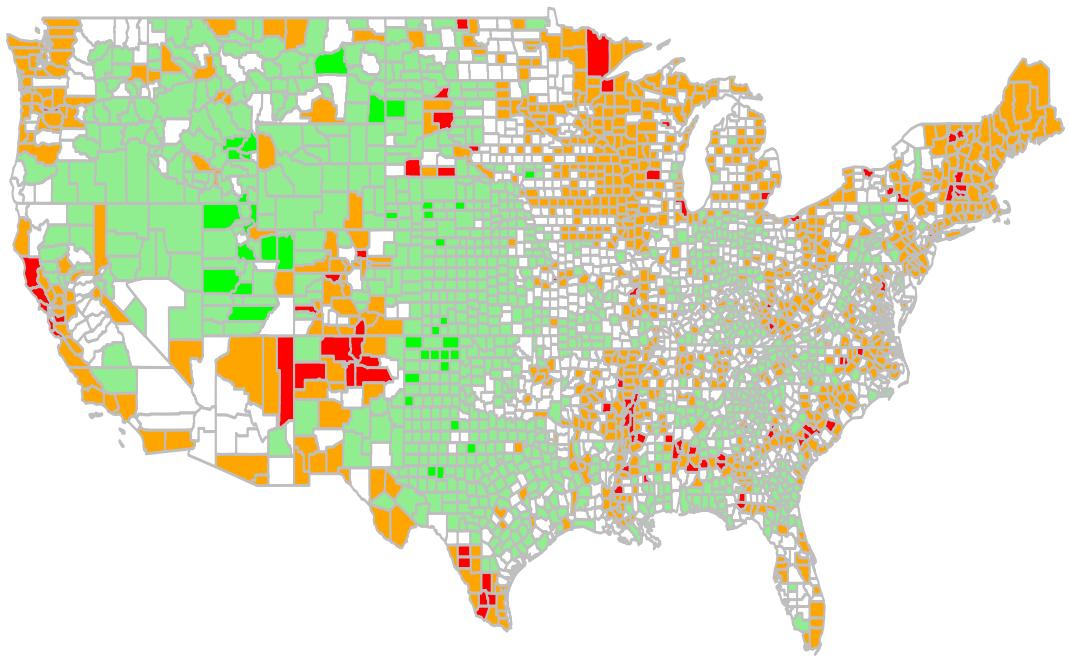
Areal Data: OLS vs SAR

Let's look at the 2004 US election data, which was Bush (Rep.) vs Kerry (Dem.). We want to see if there is an association between the number of voters who voted for Bush and median household income.

- Take rates first: percent of total voters who voted for Bush, and per capita income
- Fit a linear model to look at effect estimates and residuals.

Areal Data: OLS vs SAR

Fitting a simple linear model with no spatial autocorrelation, we see there is still spatial autocorrelation in the residuals (therefore they are not iid).



Areal Data: OLS vs SAR

To test the residual spatial autocorrelation, we look at the Moran's I of the residuals:

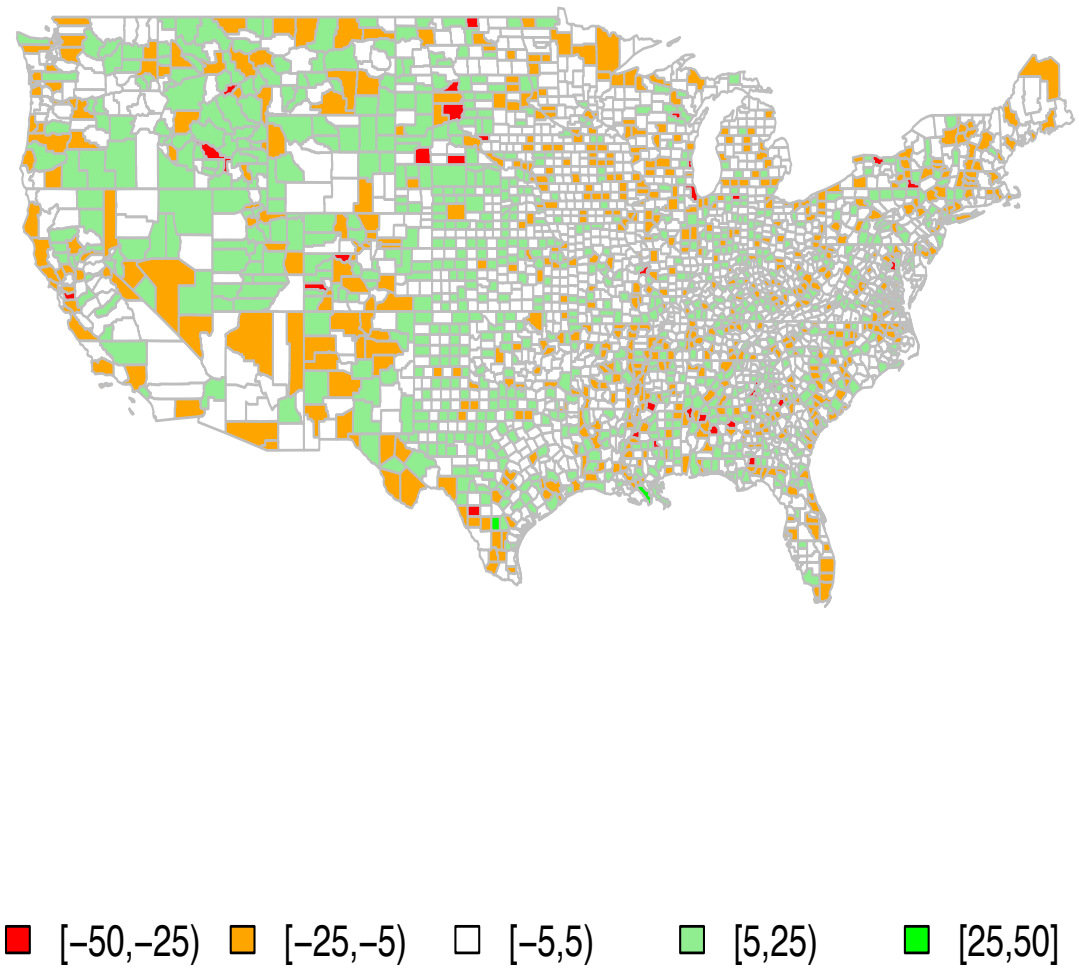
```
data:  res.lm
weights:  W_cont_el_mat

Moran I statistic standard deviate = 51.1383, p-value
=0.00001
alternative hypothesis:  greater
sample estimates:
Moran I statistic Expectation Variance
0.5501302998 -0.0003219575 0.0001158637
```

The null hypothesis of no spatial autocorrelation is rejected.

Areal Data: OLS vs SAR

Fitting a SAR model, the residuals are no longer spatially autocorrelated



Areal Data: OLS vs SAR

To test the residual spatial autocorrelation, we look at the Moran's I of the residuals:

```
data:  res.sar
weights:  W_cont_el_mat

Moran I statistic standard deviate = -4.7192, p-value
=1
alternative hypothesis:  greater
sample estimates:
Moran I statistic Expectation Variance
-0.0510920669 -0.0003219575 0.0001157378
```

We fail to reject the null hypothesis of no spatial autocorrelation.

Areal Data: Conditional Autoregressive Models

- SAR models are in terms of the joint probability distribution function $p[Y(s_1), Y(s_2), ..., Y(s_n)]$
- SAR spatial dependence is through the variance-covariance matrix $\Sigma_{SAR} = Var(\mathbf{Y})$
- Assume $Y(s_i)$ depends $Y(s_j)$ only if s_j is in the neighbourhood set of s_i
- CAR models are in terms of the conditional probability distribution
- Each observation $Y(s_i)$ is conditional on the values of all other observations
- CAR models are thus defined by $p[Y(s_i)|Y_{-i}]$ where Y_{-i} is the vector of all observations except $Y(s_i)$
- This model is applied for each observation in turn

Areal Data: Conditional Autoregressive Models

- As in SAR models, we assume $Y(s_i)$ depends $Y(s_j)$ only if s_j is in the neighbourhood set of s_i
- We also assume that each conditional distribution is Gaussian. As a result, we only need to specify the conditional mean and variance of each observation
- The CAR mean and variance are specified as:

$$E[Y(s_i)|Y_{-i}] = x(s_i)\beta + \sum_j c_{ij}[Y(s_j) - x(s_i)\beta]$$

$$V[Y(s_i)|Y_{-i}] = \sigma_i^2$$

- The spatial dependence parameters are shown as c_{ij} for the CAR model as opposed to b_{ij} from our SAR model
- Again, $c_{ii} = 0$ since we do not want to regress observed values on itself and c_{ii} are non-zero as long as s_j is in the neighbourhood of s_i

Areal Data: Conditional Autoregressive Models

- The joint pdf $p[Y(s_1), Y(s_2), ..., Y(s_n)]$ determines the full conditional distribution $p[Y(s_i)|Y(s_j), j \neq i]$
- However, when does $p[Y(s_i)|Y(s_j), j \neq i]$ uniquely determine $p[Y(s_1), Y(s_2), ..., Y(s_n)]$?

Areal Data: Conditional Autoregressive Models

- Through Brook's lemma we can recover the joint distribution of $Y(s_i)$ from the full conditionals
- Let $y_0 = (y_{10}, ..., y_{n0})$

$$p(y_1, ..., y_n) = \frac{p(y_1|y_2, ..., y_n)}{p(y_{10}|y_2, ..., y_n)} \frac{p(y_2|y_{10}, y_3, ..., y_n}{p(y_{20}|y_{10}, y_3, ...y_n)} ... \frac{p(y_n|y_{10}, ..., y_{n-1,0})}{p(y_{n0}|y_{10}, ..., y_{n-1,0})} p(y_{10}, ..., y_{n0})$$

- The conditions needed for a set of conditional distributions to define a valid joint distribution are less straightforward than for joint distributions to define conditional distributions (in the Gaussian case)

Areal Data: Conditional Autoregressive Models

- A set of conditional distributions defined over spatial neighbourhoods and meeting the conditions defines a Markov Random Field, where each observation given the other observations depends only on values at neighbouring locations

Areal Data: Conditional Autoregressive Models

$$E[Y(s_i)|Y_{-i}] = x(s_i)\beta + \sum_j c_{ij}[Y(s_j) - x(s_i)\beta]$$

$$V[Y(s_i)|Y_{-i}] = \sigma_i^2$$

- In terms of the CAR model, $\Sigma_{CAR}(\theta) = \sigma^2(I - C)^{-1}\Sigma_c$
- The difference from SAR is that $\Sigma_c = diag[\sigma_1^2, ..., \sigma_n^2]$ and we impose the symmetry constraint that $\sigma_j^2 c_{ij} = \sigma_i^2 c_{ji}$

Areal Data: Autoregressive Models

Is there a particular reason why we would choose one over the other?
As the Encyclopedia of GIS states, the conditional autoregressive model (CAR) is appropriate for situations with first order dependency or relatively local spatial autocorrelation, and simultaneous autoregressive model (SAR) is more suitable where there are second order dependency or a more global spatial autocorrelation.
The CAR model obeys the properties of a Markov random field, namely it assumes that the state of a particular area is influenced only by its neighbours and not neighbours of neighbours, etc. (i.e. it is spatially memoryless), whereas SAR does not assume such. This is due to the different ways in which the variance-covariance matrices are specified. So, when the Markov random field property holds, CAR provides a simpler way to model autocorrelated geo-referenced areal data.

Areal Data: Autoregressive Models

Relationship with Mixed Effects Models

While mixed effects models are seen in a different context than spatial models, they are central to multi-level models and small area estimation, which can be used in the analysis of spatial data. The errors in the SAR and CAR models are used to account for between-area variation, following a specified correlation structure. This is typically known as random effects (random effects can change from area to area).

Mixed effects models can be formulated as:

$$Y = x\beta + Ze + \epsilon$$

Where e represents the random effects which are $N(0, \Sigma_e)$and Z accounts for the structure in the random effects. We can set Z to reproduce SAR or CAR specification, but it also must depend on ρ, the spatial parameter.

In R, we can use the nlme package (example in R script).

Areal Data: Geographic Weighted Regression

Geographic weighted regression (GWR) is a technique where locally weighted regression coefficients are applied by moving a weighted window over the data. A bandwidth must be chosen for an isotropic spatial weights kernel (usually gaussian). GWR re-writes the linear model in a slightly different form:

$$y_i = X\beta_i + \epsilon$$

where i is the location at which the local parameters are to be estimated. Parameter estimates are solved using the weighting scheme:

$$\beta_i = (X^t W_i X)^{-1} X^t W_i y$$

where the weights W_{ij} are calculated with a Gaussian function such as

$$w_{ij} = exp(\frac{-d_{ij}}{h})^2$$

Areal Data: Geographic Weighted Regression

The distance in the weighting function d_{ij} is the Euclidean distance between the location of observation i and j, and h is the bandwidth. The bandwidth may be user defined or by minimization of the root mean square prediction error.

The Modifiable Areal Unit Problem (MAUP)

Areal data inherently involve aggregation. The modifiable areal unit problem is the analytical issue of different scales of aggregation.

- The issue was first documented by Gehlke and Biehl (1934) when they found statistical inference changed when contiguous census tracts were grouped to form larger areas. The magnitude of correlation coefficient between two variables increased. When random census tracts were grouped, the correlation was unaffected by group size.
- MAUP was coined by Openshaw and Taylor (1979)
- Spatial correlation changes depending on the size of the areal unit: when smaller areal units were aggregated to form larger units, correlation changed.
- Boundaries of areal units are often created artificially and can be changed. When boundaries are drawn in a different way, you are likely to get different results.

The Ecological Fallacy

This is a term used in epidemiology to explain the process of deducing individual behavior from aggregate data.

- Ecological (aggregate) and individual correlations are almost always not equal. Conclusions on individual behavior drawn from grouped data are therefore questionable.
- Ecological fallacy occurs when analyses based on grouped data lead to conclusions different from those based on individual data. The resulting bias is often called "ecological bias".
- It is a special case of the MAUP.

Change of Support and Spatial Misalignment

Spatial transformations may be required to 'align' data. For example, meteorological and air pollution data occur at points, but census data occur as areal units. If we want to examine associations between point and areal data, we have to change the support of the point data, and represent it more generally as an average over an areal unit.

- Spatial "support" refers to the size or volume associated with each data value. The complete specification also includes the geometrical size, shape and spatial orientation of the regions associated with the measurements.
- Changing the support of a variable is typically done by averaging or aggregation.
- The changed variable will have different statistical and spatial properties.

Solutions to Spatial Misalignment

The geostatistical solutions to spatial misalignment are used for point data

- Kriging
- Co-Kriging
- Block Kriging (kriging where the average in an area is generated rather than the value at a point)

Preserving the original scale of the data is desirable, and multiscale spatial methods are one way to do this. This is an active area of research.
See Gotway and Young (2002) for a statistical description of the problem.