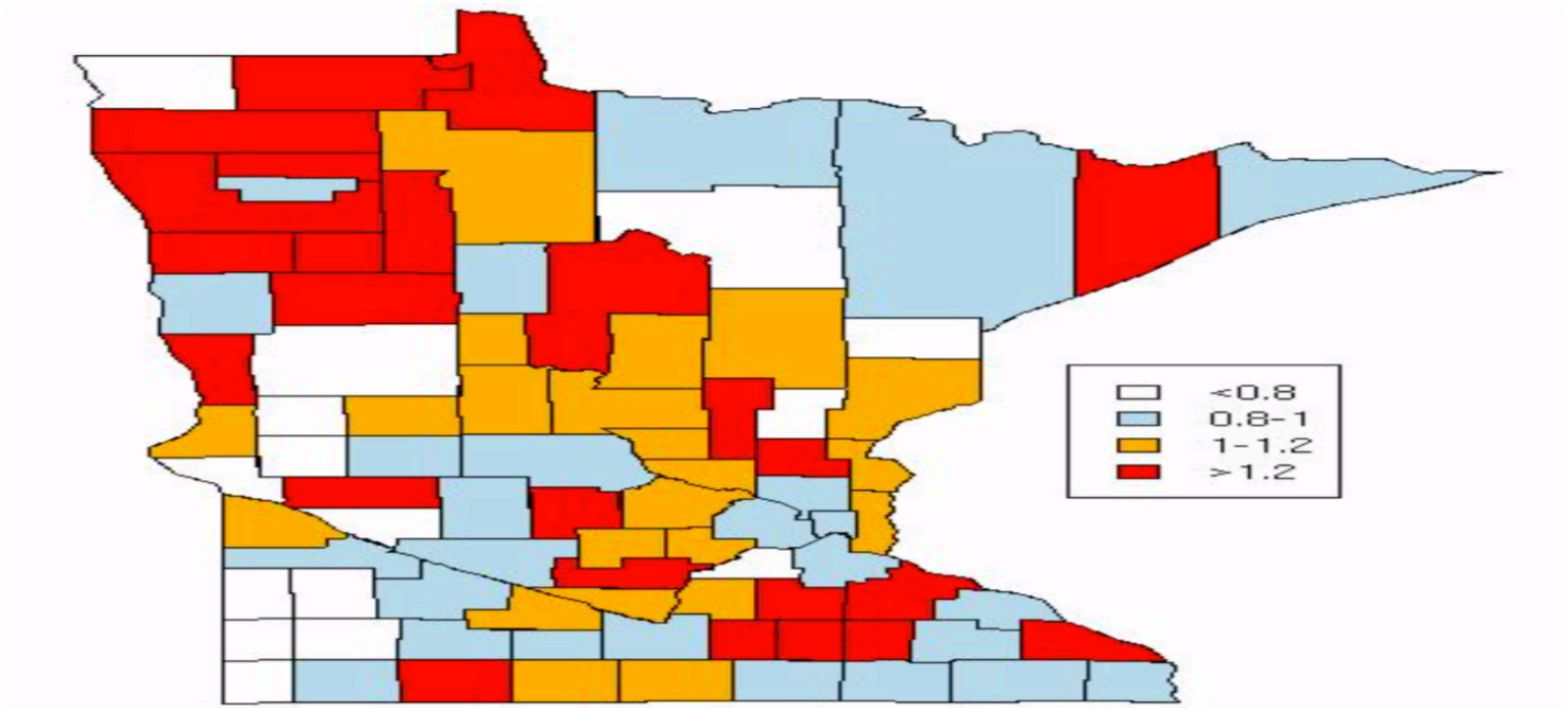
Lecture 4. Areal Data

Spatial Big Data Analysis with GIS

Korean Statistical Society, Winter School, February 24, 2023

Areal data

Figure below is called a **choropleth map**, produced by classifying values into a few different bins (like a histogram) and shading a map accordingly.



Areal data basic

- Do not observe process at sites $\mathbf{s}_1, \ldots, \mathbf{s}_n$ but instead a block-level $b_1, \ldots, b_n$.
- Notion of distance is not obvious anymore.
- One distance-based approach: use distances between centroids of regions and use geostatistical models.

  Distances between centroids treats the data as if they were concentrated at a single point in the centroid of the region. Also, with some irregular shapes, centroid of a region may lie outside or on the border of the region.
- Easier to look at adjacencies: regions sharing a border are adjacent to each other.

Inference for areal data

- Is there a significant **spatial pattern** (dependence)?
- Are there **clusters** with elevated/depressed levels ('**hotspots**')?
- Our information may be exhaustive in many cases (we may have data for all cells), but we hope to produce better estimates by taking spatial relationships into account. Deciding how to model spatial relationships impacts how much spatial smoothing is done.
- Change of support issues: (a) inference for areal units that are different from the ones for which data are observed (e.g., county versus zip code information), (b) inference that involves areal data and point level data.

Inference for areal data (conti-)

- Julian Besag: Spatially smoothed maps are exploratory in that they can provide motivation for looking for appropriate missing (spatial) covariates.
- Typically use **Markov random field (MRF)** models - powerful methodology that is useful for modeling a variety of problems.
- MRF models have computational advantages (say over Gaussian process models) that make them useful for large data sets and complicated problems.
- **Lattice data**: The discussion that follows largely carries over to modeling and inference for both areal and lattice data.

Measures of spatial association

- Moran's I is like 'lagged autocorrelation' in time series:

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\sum_{i \neq j} w_{ij} \sum_i (Y_i - \bar{Y})^2}$$

  I is not necessarily in $[-1, 1]$.

- Let $w_{ij}^{(1)} = 1$ if i and j are first-order neighbors, else 0. We can use these weights to obtain $I^{(1)}$. Similarly, define $w_{ij}^{(r)}$ according to whether i and j are rth order neighbors and obtain $I^{(r)}$. Plot of $I^{(r)}$ versus r is a **correlogram**.
- Plot of correlogram: if there is spatial dependence, it will decrease with increase in r until the distance where there is no appreciable dependence (after that it should vary around 0).
- This is the areal equivalent of an empirical covariogram.

Measures of spatial association (conti-)

- Geary's C, like Durbin-Watson test statistic for first-order serial correlation:

$$C = \frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{2\sum_{i \neq j} w_{ij} \sum_i (Y_i - \bar{Y})^2}, C \geq 0$$

- Under the null hypothesis of independence: $E(C) = 1$.
- Small C, $C \in (0, 1)$ indicates positive spatial association.

  $C > 1$ indicates negative spatial association.
- Ratio of quadratic forms, asymptotically normal if Y_is are iid (but Y_is are not iid!). Better to do significance testing by permutation tests.
- Areal equivalent of an empirical variogram.

Moran's I vs Geary's C

- Both are two special cases of the general cross-product statistic that measures spatial autocorrelation.
- Moran's I is produced by standardizing the spatial autocovariance by the variance of the data.
- Geary's C uses the sum of the squared differences between pairs of data values as its measure of covariation.
- Geary's C is inversely related to Moran's I, but they are not identical.
- Moran's I is a measure of global spatial autocorrelation, while Geary's C is more sensitive to local spatial autocorrelation.

Local Indicators of Spatial Association (LISA)

- Anselin (1995) proposed the ideas of LISAs.
    1. The LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around the observation.
    2. The sum (or mean) of LISAs for all observations is proportional to global indicator of spatial association.
- The local Moran's I statistic can be used to detect some local clusters and outliers.

Spatial autoregressive models

- In time series, autoregressive models express the data at time t as a linear combination of the values in the past.
- For example, if Z_t is a time series of interest, an AR(p) model is in the form

$$Z_t = c + \sum_{i=1}^{p} \phi_i Z_{t-i} + \epsilon_t$$

- We can do similar things with spatial data, e.g., SAR and CAR models.

Indirect versus direct modeling of dependence

- To illustrate why modeling lattice/areal data processes is inherently different, it is useful to consider the **Simultaneous Autoregressive (SAR)** Model.
- Assume $Z(\mathbf{s}_i) = \mu(\mathbf{s}_i) + w(\mathbf{s}_i)$ where $\mu(\mathbf{s}_i)$ is the mean function, and the dependent error is modeled via $w(\mathbf{s}_i)$ for $\mathbf{s}_1, \ldots, \mathbf{s}_n$.
- In geostatistics, we could simply define the covariance of the $w(\mathbf{s}_i)$s via a covariance function that is continuous in space. However, we cannot do this now.
- Instead, model $w(\mathbf{s}_i)$s as some combination of contributions from other sites.

Indirect modeling of dependence: SAR model

- Modeling $w(\mathbf{s}_i)$s as a combination of contributions from other sites using the SAR model:

$$w(\mathbf{s}_i) = \sum_{j=1}^{n} b_{ij}\{Z(\mathbf{s}_j) - \mu(\mathbf{s}_j)\} + \epsilon(\mathbf{s}_i),$$

  where $\epsilon(\mathbf{s}_i)$s are independent error. Similar to the measurement error/nugget term in geostatistics models (but no micro-scale variation interpretation here).

- The b_{ij}s denote the spatial connectivity of the sites. If all b_{ij} are 0, we are back to the iid scenario.

- $\mathbf{B} = \{b_{ij}\}$ is a parameter of the model, typically $\mathbf{B} = \rho\mathbf{W}$ where $\mathbf{W}$ is a spatial connectivity matrix, ρ is a parameter.

SAR model covariance

- Even if $b_{ij} = 0$, $Z(\mathbf{s}_i)$ and $Z(\mathbf{s}_j)$ can be dependent. Let $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^\top$. Write the model as:
  - $Z(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s})$
  - $\mathbf{w} = \mathbf{B}\mathbf{w} + \epsilon$
- The above two stage model can be expressed as:

$$(\mathbf{I} - \mathbf{B})(\mathbf{Z} - \mu) = \epsilon,$$

  where $\mathbf{B} = \{b_{ij}\}$ and $\epsilon \sim N(\mathbf{0}, \Sigma_e)$, $\Sigma_e = \psi\mathbf{I}$.
- Hence, $\text{Var}(\mathbf{Z}) = \psi(\mathbf{I} - \mathbf{B})^{-1}(\mathbf{I} - \mathbf{B}^\top)^{-1}$.
- So, even if $b_{ij} = 0$, we can have $\text{Cov}(Z(\mathbf{s}_i), Z(\mathbf{s}_j)) \neq 0$.
- Covariance has been defined *indirectly* via $\mathbf{B}, \Sigma_e$.

Proximity matrices for areal data

- Define an $n \times n$ matrix $\mathbf{W} = \{w_{ij}, i = 1, \ldots, n, j = 1, \ldots, n\}$ (n: number of regions). Let $w_{ii} = 0$ but define spatial proximity between region i and j by w_{ij}, for example:
  - $w_{ij} =$ inverse distance between centroids of i, j.
  - $w_{ij} = 1$ if regions i, j are adjacent, else 0.
  - $w_{ij} = 1$ if distance between i, j is $< \delta$ (some fixed δ).
  - $w_{ij} = 1$ if region j is among m closest regions of i.
- Note: not necessarily symmetric.
- Can define first-order neighbors $\mathbf{W}^{(1)}$, second-order neighbors $\mathbf{W}^{(2)}$.

Markov property

- In time series, if the conditional distribution of $Z(t+1)$ given $Z(s)$, $s = 1, \ldots, t$ is the same as that of $Z(t+1)$ given $Z(t)$, we say the process has *Markov property*.
- We can extend this to spatial data. A spatial random field $Z(\mathbf{s})$ is a *Markov random field* if $Z(\mathbf{s}_i)$ only depends on its neighbors $\mathcal{N}_i$.

Conditional autoregressive (CAR) models

- CAR model uses the concept of Markov property.
- We consider $f(Z(\mathbf{s}_i)|\mathbf{Z}(\mathbf{s})_{-i})$ where $\mathbf{Z}(\mathbf{s})_{-i}$ denotes the vector of all the data except $Z(\mathbf{s}_i)$.
- Specifically we assume each of the conditional distribution is Gaussian and we let

$$\begin{aligned} E(Z(\mathbf{s}_i)|\mathbf{Z}(\mathbf{s})_{-i}) &= \mu(\mathbf{s}_i) + \sum_{j=1}^{n} b_{ij}\{Z(\mathbf{s}_j) - \mu(\mathbf{s}_j)\} + \epsilon(\mathbf{s}_i), \\ \text{Var}(Z(\mathbf{s}_i)|\mathbf{Z}(\mathbf{s})_{-i}) &= \sigma_i^2, \quad i = 1, \ldots, n. \end{aligned}$$

  Here, b_{ij} is nonzero only if $\mathbf{s}_i \in \mathcal{N}_i$ and $b_{ii} = 0$.

- Given conditional distributions, it is not easy to construct joint distribution to do estimation and inference.

SAR and CAR models

- SAR model is very popular in spatial econometrics.
- SAR models are more suitable where there are second order dependency or a more global spatial autocorrelation.
- Conditional autoregressions (Beags, 1974) or CAR models are appropriate for situations with first order dependency or relatively local spatial autocorrelation.

  Set $\text{Var}(\mathbf{Z}) = (\mathbf{I} - \mathbf{C})^{-1}$ where $\mathbf{C} = (\mathbf{B} + \mathbf{B}^\top) - \mathbf{B}^\top\mathbf{B}$ to obtain CAR representation of the SAR model (Not the same weights structure).
- SAR model is only defined for multivariate Gaussian distribution while CAR model is not.

Inference/parameterization for SAR and CAR models

- A large portion of the work appears to proceed by using a Bayesian formulation of the problem, since it appears more natural to assume conditional independence of the data given the parameters and then assume the parameters follow an MRF (Cressie, 1991).
- R package to carry out inference for simple SAR and CAR models: `spdep`. Also, does exploratory data analysis (Moran's I, Geary's C etc.) which can be helpful for selecting the neighborhood structure, which in turn is used to specify an appropriate model.

Reference

- Cressie, N. Statistics for Spatial Data. Wiley. Chapter 1.
- Banerjee, S., Carlin, B., and Gelfand, A. Hierarchical Modeling and Analysis for Spatial Data (2nd). CRC Press.
- Jun, M., Genton, M. G., and Jeong, J. Lecture Notes for Spatial Statistics. UH, KAUST, and HYU.