LECTURE 17: EXPLORATORY APPROACHES FOR AREAL DATA

CLASS INTRO

INTRO QUESTIONS

- Discuss the differences in the two figures from the third choropleth exercise from Lecture 16.
- For Today:
  - Areal Data Association and Smoothing

EXPLORATORY APPROACHES FOR AREAL DATA

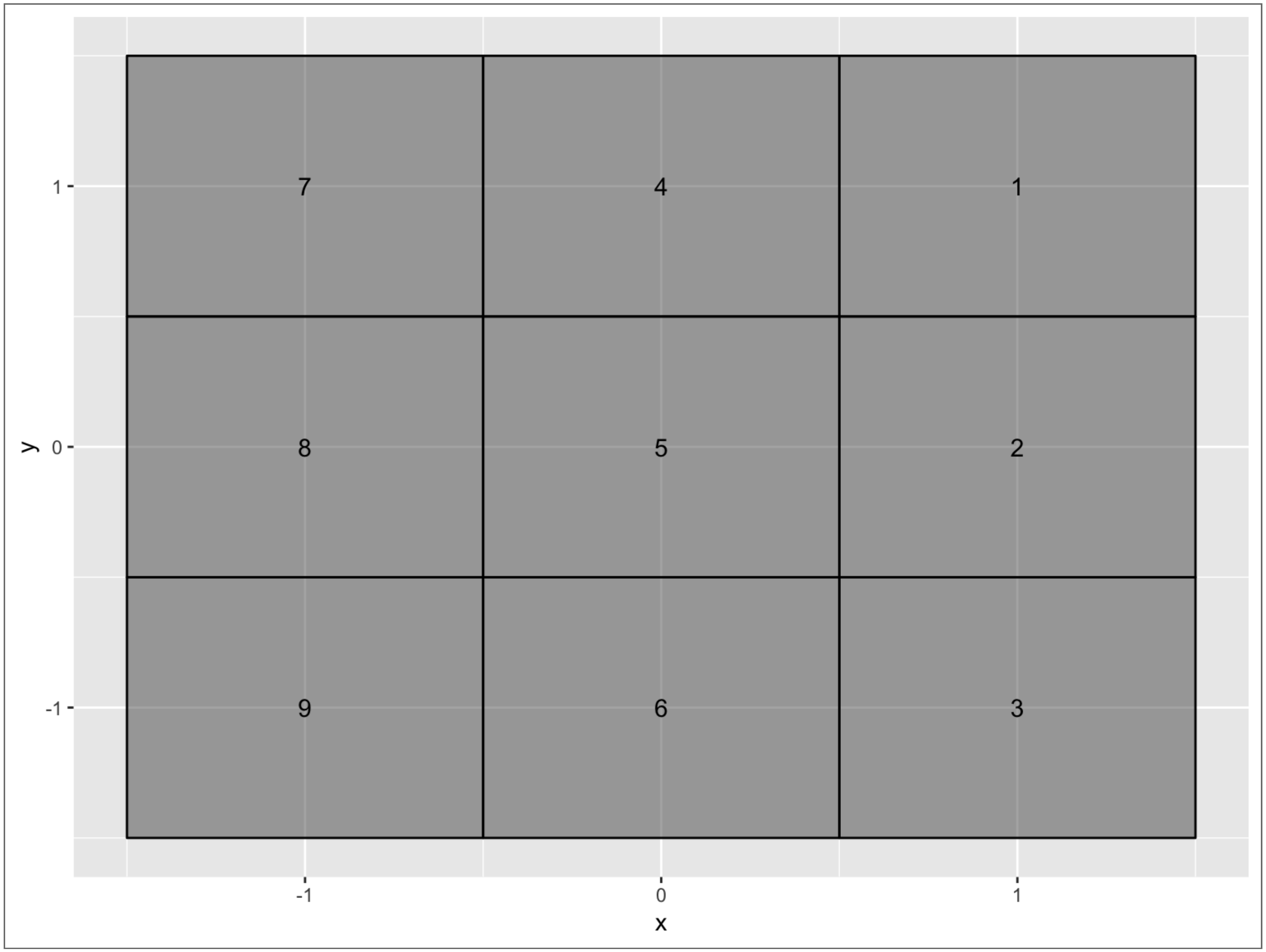
PROXIMITY MATRIX

- Similar to the distance matrix with point-reference data, a proximity matrix W is used to model areal data.
- Given measurements $Y_i, \dots, Y_n$ associated with areal units $1, \dots, n$, the elements of W, w_{ij} connect units i and j
- Common values for w_{ij} are

$$w_{ij} = \begin{cases} 1 & \text{if i and j are adjacent} \\ 0 & \text{otherwise (or if i=j)} \end{cases}$$

GRID EXAMPLE

- Create an adjacency matrix with diagonal neigbors
- Create an adjacency matrix without diagonal neigbors



DISTANCE BASED PROXIMITY MATRIX

- The w_{ij} entries can also be a distance (which is often standardized)
- A set of proximity matrices can also be created that denote first-order neighbors $W^{(1)}$, second-order neighbors $W^{(2)}$, and so on.

SPATIAL ASSOCIATION

MEASURES OF SPATIAL ASSOCIATION

- There are two common statistics used for assessing spatial association:
  - Moran's I
  - Geary's C

MORAN'S I

- Moran's I
  $$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i \neq j} \ w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$
- This is a spatial analogue measuring the lagged autocorrelation.

MORAN'S I

- When the Y_i are iid, I is asymptotically normal with mean = $\frac{-1}{n-1}$
- While a delta method based asymptotic variance is available, for hypothesis testing Monte Carlo procedures are preferred.

GEARY'S C

- Geary's C

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

GEARY'S C

- Geary's C is non-negative with mean of 1 for i.i.d. data.
- Small values, near zero, indicate positive spatial association.
- This is a spatial analogue of the Durbin-Watson Test
- Monte Carlo procedures are preferred.

SPATIAL ASSOCIATION EXERCISE: 1

For the following four scenarios, simulate data on a grid. Plot the grids. Then using a proximity matrix with 1's for neighbors (not diagonal) compute I and G.

Note `geom_raster()` can be used for the figures.

```
df <- expand.grid(x = 1:3, y = 1:3)
df$z <- runif(nrow(df))
# default is compatible with geom_tile()
ggplot(df, aes(x, y, fill = z)) + geom_raster()
```

SPATIAL ASSOCIATION EXERCISE: 2

1. Simulate data for a 3-by-3 grid where the responses are i.i.d. N(0,1). Plot the sampling distribution of I and G.
2. Simulate data for a 6-by-6 grid where the responses are i.i.d. N(0,1). Plot the sampling distribution of I and G.
3. Simulate data and calculate I and G for a 3-by-3 grid and 6-by-6 grid with a chess board approach, where "black squares" $\sim N(-1, 1)$ and "white squares" $\sim N(1, 1)$.

SPATIAL ASSOCIATION EXERCISE: 3

4. Simulate data and calculate I and G for a 6-by-6 grid with the following form.

```
times <- 1:36
rho <- 0.9
H <- abs(outer(times, times, "-"))
V <- rho^H
p <- nrow(V)
V[cbind(1:p, 1:p)] <- V[cbind(1:p, 1:p)]
library(mnormt)
df <- expand.grid(x = 1:6, y = 1:6)
df$z <- rmnorm(1,rep(0,36), V)
# default is compatible with geom_tile()
ggplot(df, aes(x, y, fill = z)) + geom_raster()
```

CORRELOGRAM

- In an analogue to time series data, a correlogram can be constructed.
- The correlogram would have some statistics, say I, on the y-axis and some lag (say $r - th$ order neighbor.)

SPATIAL SMOOTHING

SMOOTHING

- Spatial smoothing results in a "smoother" spatial surface, by sharing information from across the neighborhood structure.
- One option is replacing Y_i with
  $$\hat{Y}_i = \sum_j w_{ij} Y_j / w_{i+},$$
  where $w_{i+} = \sum_j w_{ij}$.
- What are some pros and cons of this smoother?

"EXPONENTIAL" SMOOTHER

- Another option would be to use:
  $\hat{Y}_i^* = (1-\alpha)Y_i + \hat{Y}_i$
- Compare $\hat{Y}_i^*$ with $\hat{Y}_i$.
- What is the impact of α?
- This is essentially the exponential smoother from time series.
- More details about smoothing (or shrinkage) will be discussed from a model-based framework.

BROOK'S LEMMA AND MARKOV RANDOM FIELDS

BROOK'S LEMMA

- To consider areal data from a model-based perspective, it is necessary to obtain the joint distribution of the responses $p(y_1, \ldots, y_n)$.
- From the joint distribution, the *full conditional distribution* $p(y_i | y_j, j \neq i)$, is uniquely determined.
- Brook's Lemma states that the joint distribution can be obtained from the full conditional distributions.

LARGE AREAL DATA SETS

- When the areal data set is large, working with the full conditional distributions can be preferred to the full joint distribution.
- More specifically, the response Y_i should only directly depend on the neighbors, hence,
  $p(y_i | y_j, j \neq i) = p(y_i | y_j, j \in \delta_i)$
  where δ_i denotes the neighborhood around i.

MARKOV RANDOM FIELD

- The idea of using the local specification for determining the global form of the distribution is Markov random field.
- An essential element of a MRF is a *clique*, which is a group of units where each unit is a neighbor of all units in the clique
- A *potential function* is a function that is exchangeable in the arguments.
- With continuous data a common potential is $(Y_i - Y_j)^2$ if $i \sim j$ (i is a neighbor of j).

GIBBS DISTRIBUTION

- A joint distribution $p(y_1, \ldots, y_n)$ is a Gibbs distribution if it is a function of Y_i only through the potential on cliques.
- Mathematically, this can be expressed as:
  $$p(y_1, \ldots, y_n) \propto \exp\left(\gamma \sum_k \sum_{\alpha \in \mathcal{M}_k} \phi^{(k)}(y_{\alpha_1}, y_{\alpha_2}, \ldots, y_{\alpha_k})\right),$$
  where $\phi^{(k)}$ is a potential of order k, $\mathcal{M}_k$ is the collection of all subsets of size $k = 1, 2, \ldots$(typically restricted to 2 in spatial settings), α indexes the set in $\mathcal{M}_k$.

HAMMERSLEY-CLIFFORD THEOREM

- The Hammersley-Clifford Theorem demonstrates that if we have a MRF that defines a unique joint distribution, then that joint distribution is a Gibbs distribution.
- The converse was later proved, showing that a MRF could be sampled from the associated Gibbs distribution (origination of Gibbs sampler).

MODEL SPECIFICATION

- With continuous data, a common choice for the joint distribution is the pairwise difference
$$p(y_1, \ldots, y_n) \propto \exp\left(-\frac{1}{2\tau^2}\sum_{i,j}(y_i - y_j)^2 I(i \sim j)\right)$$
- Then the full conditional distributions can be written as
$$p(y_i | y_j, j \neq i) = N\left(\sum_{j \in \delta_i} y_i / m_i, \tau^2 / m_i\right)$$
where m_i are the number of neighbors for unit i.
- This results in a spatial smoother, where the mean of a response is the average of the neighbors.