# Modeling Areal Data: Intro

## Spatial Smoothing

Spatial smoothing results in a "smoother" spatial surface, by sharing information from across the neighborhood structure.

This smoothing is akin to fitted values (expected values) in a traditional modeling framework.

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 $\alpha$? *This is essentially the exponential smoother from time series.*

## Areal Data Models: Disease Mapping

Areal data with counts is often associated with disease mapping, where there are two quantities for each areal unit:

$$\begin{aligned} Y_i &= \text{observed number of cases of disease in county i} \\ E_i &= \text{expected number of cases of disease in county i} \end{aligned}$$

One way to think about the expected counts is

$$E_i = n_i \bar{r} = n_i \left( \frac{\sum_i y_i}{\sum_i n_i} \right),$$

where $\bar{r}$ is the overall disease rate and $n_i$ is the population for region $i$.

However note that $\bar{r}$, and hence, $E_i$ is a not fixed, but is a function of the data. This is called *internal standardization.*

An alternative is to use some standard rate for a given age group, such that $E_i = \sum_j n_{ij} r_j$. This is *external standardization.*

### Traditional Models

Often counts are assumed to follow the Poisson model where

$$Y_i | \eta_i \sim Poisson(E_i \eta_i),$$

where $\eta_i$ is the relative risk of the disease in region $i$.

Then the MLE of $\eta_i$ is $\frac{Y_i}{E_i}$. This quantity is known as the *standardized morbidity ratio* (SMR).

**Poisson-Gamma Model**

Consider the following framework

$$\begin{aligned} Y_i|\eta_i &\sim Po(E_i\eta_i),\ \text{i} = 1, \ldots, \text{I} \\ \eta_i &\sim Gamma(a, b), \end{aligned}$$

where the gamma distribution has mean $\frac{a}{b}$ and variance is $\frac{a}{b^2}$.

This can be reparameterized such that $a = \frac{\mu^2}{\sigma^2}$ and $b = \frac{\mu}{\sigma^2}$.

For the Poisson sampling model, the gamma prior is conjugate. This means that the posterior distribution $p(\eta_i|y_i)$ is also a gamma distribution, and in particular, the posterior distribution $p(\eta_i|y_i)$ is $Gamma(y_i + a, E_i + b)$.

The mean of this distribution is

$$\begin{aligned} E(\eta_i|\boldsymbol{y}) = E(\eta_i|y_i) &= \frac{y_i + a}{E_i + b} = \frac{y_i + \frac{\mu^2}{\sigma^2}}{E_i + \frac{\mu}{\sigma^2}} \\ &= \frac{E_i(\frac{y_i}{E_i})}{E_i + \frac{\mu}{\sigma^2}} + \frac{(\frac{\mu}{\sigma^2})\mu}{E_i + \frac{\mu}{\sigma 2}} \\ &= w_i SMR_i + (1 - w_i)\mu, \end{aligned}$$

where $w_i = \frac{E_i}{E_i + (\mu/\sigma^2)}$

Thus the point estimate is a weighted average of the data-based SMR for region $i$ and the prior mean $\mu$.

## Poisson-lognormal models

The model can be written as

$$\begin{aligned} Y_i|\psi_i &\sim Poisson(E_i \exp(\psi_i)) \\ \psi_i &= \boldsymbol{x_i^T}\boldsymbol{\beta} + \theta_i + \phi_i \end{aligned}$$

where $\boldsymbol{x_i}$ are spatial covariates, $\theta_i$ corresponds to region wide heterogeneity, and $\psi_i$ captures local clustering.

### Brook's Lemma and Markov Random Fields

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.

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 $\delta_i$ 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 $k = 1, 2, ...$(typically restricted to 2 in spatial settings), $\alpha$ 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.