Spatial Methods Memo

Ethan Roubenoff

September 2020

In this reading list, I am focusing on a series of statistical models to handle spatial data that are common in demographic and public health data. Many non-spatial statistical models, including simple linear regression, assume that data are independently and identically distributed (iid). In this case, each observation in the data has the same distribution they are mutually independent. However, for a number of reasons specific to each research topic, we may have reason to believe that this iid assumption is not valid; this may be especially true when data have a strong spatial pattern and data may be conditional on neighboring observations. Methods that falsely assume iid can lead to inflated variance estimates or biased mean estimates if the spatial pattern is strong.

Spatial non-independence can be summarized by Tobler's first law of geography (Tobler, 1970): "Everything is related, but close things are more related than far things." In modern human populations, people are more likely to work, eat, and socialize near where they live; subsequently working, eating, and socializing with people who live near them as well. This scale can be very small, on the neighborhood level (considering work on geography of opportunity) or entire cities, states, or countries. There are a number of forces at play here: first, a source of unobserved heteogeneity could be spatially patterned, for example in an environmental risk; second, humans can naturally cluster together along certain traits and identities, including language, income, and employment among many others; third, human contact patterns may in turn cause other phenomena to be spatially patterned, like accents, information dissemination, and disease outbreaks. Often all three, possible many more, may be present—they may act circularly, be self-reinforcing, or otherwise create nonseparable dynamics.

Voss argues that all demography is spatial, to a degree. Administrative records are specific to locales at all levels of resolution; estimations are often implicitly, if not explicitly, about a population in a certain place (Voss, 2007). Progress in accessible computational techniques like Geographically Weighted Regression and Spatial Econometrics saw a wild expansion in the 1990s as GIS became a common tool for data analysis (Matthews and Parker, 2013). All migration studies are necessarily spatial; people flow from one area to another (Wachter, 2005).

My personal interest in geography as a demographic consideration began with theories of space-place identity construction. People collectively assign meaning to space—for example, (-37°, 122°)—turning it into a place—Berkeley, CA. These place-based boundaries (sometimes called 'platial' by Michael Goodchild, as opposed to spatial boundaries, like coastlines) may ultimately be arbitrary, their effect is nonetheless tangible (Goodchild, 2000). Tax funding, schools, social services, and legal jurisdiction all behave according political boundaries; neighborhood identity and culture becomes associated with these tangibles in combination with the compositional demography of the people who live there. This segregation is often self-reinforcing. I am interested now in how health inequalities are spatially patterned: how mortality, disease incidence, underlying health are geographically distributed. This is elaborated more in my other exam in public health in demography.

The models presented in brief here and during my Qualifying Exam focus on 'polygon' data (like mortality rates in counties). Most geostatistical work focuses on 'point' data; while mortality and health phenomena are technically point referenced (occur in a single location with coordinates), the data I will mostly be working with are often summarized to the administrative level. As such, I have chosen to focus on hierarchical models, autoregressive distributions, spatial econometrics, and space-time models, with their applications in disease mapping.

1 Spatial Autocorrelation and the Neighborhood Matrix

Spatial Autocorrelation refers to a property of spatial data wherein nearby observations may have similar values. This is sometimes seen as the 'clumpiness' of the data. Autocorrelation can be calculated as a global measure, showing how all observations either are or aren't autocorrelated, or locally, considering only the neighboring values. The latter is often used in cluster analysis, such as Local Indicators of Spatial Association (developed by Haining, but elaborated in Banerjee, Carlin, and Gelfand (2003) and Haining (2004)).

The most common estimator of autocorrelation is Moran's I:

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

The numerator represents a weighted sum of pairwise deviations from the mean; the denominator the weighted squared deviations from the mean. Commonly, weights are inverse-distances. A Moran's I value equal to 0 indicates no autocorrelation; one can conduct a hypothesis test. Positive values of Moran's I indicate positive autocorrelation (assimilation, or clumpiness); negative values indicate dispersion. **In determining if data warrant a spatial model, it is often recommended to run the model without any spatial effects and conduct a hypothesis test on the autocorrelation of the residuals; if Moran's I of residuals is significantly different from 0, a spatial model is justified.**

Critical in any spatial statistcal work is the concept of the neighborhood matrix: a mathematical representation of geographic adjacency. For example, this 3x3 grid could be representing by binary neighborhood matrix W:

1	2	3
4	5	6
7	8	9

$$W = \begin{bmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{bmatrix}$$

A symmetric binary matrix like this is most common for representing adjaceny, but can easily be extended to include reciprocal distance weights, higher-order neighbors, or measures of connectivity that are not strict adjacency (for example transit networks). While estimates will change between different matrices W, the following distributional properties remain the same.

2 Autoregressive Models

There are two classes of autoregressive models: conditionally autoregressive (CAR) and simultaneously autoregressive (SAR) models. The CAR model, also called the BYM model after authors Besag, York, and Mollie, is specified for each geographical unit as a normal distribution with expectation equal to the average of its neighbors:

$$\phi_i|\boldsymbol{\phi_{j\sim i}} \sim N\left(\frac{1}{n_i}\sum_{j\sim i}\phi_j, \frac{\tau^2}{n_i}\right) \tag{2}$$

Where ϕ_i is a unit, the set $\boldsymbol{\phi_{j\sim i}}$ indicates the neighbors of i, n_i is the number of neighbors, and τ^2 is a specified variance. This model is conditionally specified for each geographic unit. This distribution is improper, however, meaning that it cannot be integrated to a valid probability distribution; as such, it can only be used as a prior in a Bayesian model [1]. The CAR model is known to show poor performance when spatial autocorrelation is not very strong, otherwise it will oversmooth random variation in the data. Work by Cressie (1993) and Leroux attempt to modify the CAR model to improve performance. The Leroux model (Leroux, 2000) has been shown through simulation to be superior, and is employed by many in disease mapping (Lee, 2011). The CAR model can also be extended to poisson, binomial, and logistic distributions as well (Haining, 2004).

Some versions of the CAR model include a parameter for autocorrelation, usually ρ. When this parameter is included, it is possible for the CAR model to be specified properly as a valid probability distribution. In this way it can be used in a frequentist model (Dormann, 2007) Banerjee, Carlin, and Gelfand (BCG) caution against this for a number of reasons. First, ρ does not map clearly onto any other measures of spatial autocorrelation, like Moran's I or Geary's C. Instead it describes a weighted sum of the grand mean of all areas and the local mean of the neighbors, weighted by ρ, which can lead the researcher to incorrect conclusions. **BCG recommends to just use the CAR model as a prior for random effects in a Bayesian model.** A simple, two-level model for disease counts with CAR random effects could take the form:

$$Y_i \sim Pois(E_i e^{\psi_i}) \tag{3}$$

$$\psi_i = \boldsymbol{x_i'\beta} + \theta_i + \phi_i \tag{4}$$

$$\theta_i \sim N(0, \sigma^2) \tag{5}$$

$$\boldsymbol{\phi} \sim CAR(\tau) \tag{6}$$

Where Y is the counts of disease, which are assumed the be Poisson distributed with relative risk ϕ_i; θ_i is unstructured error and $\boldsymbol{\phi}$ is equivalent to the distribution in equation (2) (BCG, p 153).

The SAR model is more straightforward. Beginning with neighborhood matrix W, the SAR distribution has general form:

$$\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \sigma^2[(I-W)(I-W)^T]^{-1}) \tag{7}$$

Computation of SAR models can be complex due to matrix inversion, but the theory behind them is simpler. They do not suffer from impropriety like CAR models and can be used in MLE or

[1] In order for this to be valid, ρ must be bounded by $1/\lambda_{max} < \rho < 1/\lambda_{min}$, where λ_{max} and λ_{min} are the largest and smallest eigenvalues of W, respectively. This guarantees that $(I - \rho W)$ is positive semi-definite (BCG p. 81).

regression models (see next section for examples). SAR models tend to pick up on more 'global' autocorrelation that CAR models, which are more 'local' in their sensitivity[2]. As such, deciding between these two main families of models depends on Bayesian or Frequentist framework and the local or globally autocorrelative nature of the data.

3 Spatial Econometrics and Spatial Regression

Spatial Econometrics refers to a series of regression models that attempt to remove autocorrelation from variables, pioneered by Luc Anselin. Confusingly, sometimes spatial econometric models are called Spatially Auto Regressive models (SAR). This is the most common use case of the SAR distribution

Beginning by considering a simple general linear model (GLM) with no spatial effects:

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{u}$$

Autocorrelation can be present in any combination of the Y, X, and error terms. The idea is to incorporate the neighborhood matrix $\boldsymbol{W}$ as an adjustment factor, or 'spatial multiplier', to the autocorrelated variable in order to remove any spatial effect. In other words, values are removed (either through subtraction or multiplication) according to their spatial adjacency structure.

In choosing which model is most appropriate, there are two main considerations: first, if the spatial heterogeneity is observed or unobserved; second, if the effects of externalities (spillovers, or perturbences to the system) are felt locally or globally. Addressing the latter point first: global specifications assume that every location is related, but near locations are more related; local models only assume the first-order neighbors are related. Unobserved spatial heterogeneity is accounted for with spatial multipliers in the error term; observed heterogeneity prompts use of a spatially lagged X model.

Anselin (2003), Golgher and Voss (2016), and Kissling and Carl (2007) elaborate on a number of models, but I will only outline two here: the Spatial Durbin model and the Spatial Moving Average (SMA) model. The Spatial Durbin model assumes global autocorrelation in the error terms (also called the SAR Error model, for example by Dormann). Anselin gives the examples of modeling the effect of air quality on house prices where only proxy measures of air quality are available, such as vehicle traffic or manufacturing (a hedonic model). The spatial Durbin model has the form:

$$\begin{aligned} y &= X\beta + (I - \lambda W)^{-1}u \\ &= \lambda W y + X\beta - \lambda W X\beta + u \end{aligned}$$

In the first form of the model, the spatial multiplier $(I - \lambda W)^{-1}$ can be thought of as giving structure to the the error term u. This term is SAR distributed. Through some rearranging, the more common expanded form is derived. There are two spatial multipliers in the expanded form of the model: first, term $\lambda W y$ present on the right hand side 'removes' a proportion of neighboring values from the left hand side of the equation; second, term $\lambda W X\beta$ subtracts a proportion of neighboring values from the X variables. An important note is that this error model can be defined equivalently in terms of X and Y variables, as the first and second lines above are equivalent.

[2] Consider the variance-covariance matrices for the two distributions. The proper CAR variance is specified as $var(\boldsymbol{\phi}) = \tau^2(I - W)^{-1}$ where the SAR variance is $var(\boldsymbol{\phi}) = \sigma^2[(I - W)(I - W)^T]^{-1}$. Since the SAR variance involves multiplying W by its transpose, the structure of the variance is more complex than in the CAR model.

The spatial moving average model, used for local correlation, is defined more straightforwardly:

$$y = X\beta + u + \gamma W u$$

Anselin gives a use-case here of if heterogeneity in house prices was unmodeled and also did not affect more than two houses away. The difference between the SMA model and the Spatial Durbin model lies in the in the spatial multiplier: the SMA multipler $\gamma W u$ affects only the adjacencies represented in W and is CAR distributed, but the Durbin multiplier $(I - \lambda W)^{-1} u$ has more complex global dynamics due to the matrix inversion.

4 Geographically Weighted Regresison and other models

Geographically Weighted Regression (GWR; Fotheringham, 1998) is a fairly straightforward technique popular in social science disciplines (Matthews and Yang, 2012). GWR is analogous to a moving window regression in time-series analysis. Essentially, for each geographic area, a regression model is fit to all of the observations within a certain distance of that observation. The set of parameters estimated are assigned to that specific area; parameters are estimated for the next area and it's neighbors, and so on. The difficulty with GWR lies in determining how far that distance should be; this is called the 'bandwidth.' It can be difficult to choose bandwidth *a priori*, and as a result bandwidth selection is often performed as part of model selection.

Dormann et al 2013, in addition to SAR models, discuss Spatial Eigenvector Mapping (SEVM) and GAMs in addition to autoregressive and other GLM frameworks. SEVM works by decomposing a matrix of associated values into principal components. The main advantage of SEVM is that it does not assume stationary[3] of the data, like the autoregressive models do; however, it can be computationally prohibitive on even medium-size datasets. These models seem to perform well under simulation, although with larger confidence intervals than autoregressive models. GAMs as well do not require stationarity.

Best et al 2005 additionally discuss partition models, which split the data into clusters where local trends can be analyzed using the Potts model. They also describe a gamma moving average model, which unlike the SMA model described above, attempts to fit a smoothing function to the data.

5 Lawson et al 2012: Bayesian 2-Stage Space-Time Mixture Modeling with Spatial Misalignment of the Exposure in Small Area Health Data

Lawson et al present a "two-stage" space-time mixture model for estimating the association between air pollution (PM2.5) and asthma hospitalizations. Although they come to a contradictory conclusion—more air pollution leads to a lower asthma hospitalization rate, inconsistent with most literature on the topic—other models they test give similar results on their dataset, suggesting the discrepency comes from the data rather than from the model. I particularly like this article

[3] One assumption made by all autoregressive models is that the distribution is stationary over the area. This is often relaxed to say that the variance follows some well-defined deterministic function. Non-parametric smoothing and non-stationary approaches are not addressed here, but are discussed in BCG.

because it brings together a number of techniques I have studied: hierarchical and autoregressive models; spatial, temporal, and spatio-temporal random effects models; multivariate spatial distributions; and disease mapping. They also address spatial misalignment: in this study, exposure (air pollution) is measured at point locations and outcome (asthma hospitalizations) is recorded at the county administrative level. I find their comparison models, especially the independent space-time random effects model (model 2), to be a useful baseline in justifying the more complex model.

A concise overview of the model: the two-stage space-time mixture model follows a number of ordered steps. Initially, space and time trends of PM2.5 data collected at a series of points are aggregated to the county levels. Next, the (county, year) specific relative risk is regressed on an intercept term, the PM2.5 counts, and socioeconomic covariates. This is the first stage. In the second stage, the residuals from the first stage are decomposed into a series of latent temporal components; coefficients on these temporal components are estimated as a Multivariate Intrinsically AutoRegressive (MIAR) distribution, a multivariate analogue of the CAR model discussed earlier. The model is then re-estimated, using these weighted components as adjustments. This procedure can be summarized as:

$$y_{ij} \sim Pois(e_{ij}\theta_{ij}) \tag{8}$$

Where y_{ij} is the count of asthma hospitalizations in county i in year j, e_{ij} is the expected number of cases, and θ_{ij} is the relative risk. [4] The first stage is then:

$$log(\theta_{ij}) = \alpha_0 + Z^*_{ij}\gamma_{ij} + \boldsymbol{X^T_{ij}\beta_{ij}} \tag{9}$$

Where α is an intercept, Z^*_{ij} is the estimated PM2.5 count with coefficient γ_{ij}, $\boldsymbol{X}$ and $\boldsymbol{\beta}$ are socioeconomic values and coefficients, respectively. In the second stage, the residuals $\hat{r}_{ij}$ are assumed to follow:

$$\hat{r}_{ij}|\hat{\theta}_{ij}, y_{ij}, e_{ij} \sim N(\alpha_r + \Lambda_{ij}, \sigma^2_{r_{ij}}) \tag{10}$$

Where α_r is an intercept and Λ_{ij} is a space-time random effect, representing space-time patterns in the residuals. Λ_{ij} is then decomposed into a series of L weighted temporal component trends[5]:

$$\Lambda_{ij} = \sum_{l=1}^{L} w_{il}\chi_{il} \tag{11}$$

these weights w are assumed to follow an MIAR distribution, which specifies all components for a county as normally distributed at the average of their neighbors. These estimated weights are plugged back into the model equation (9):

$$log(\theta_{ij}) = \alpha_0 + Z^*_{ij}\gamma_{ij} + \boldsymbol{X^T_{ij}\beta_{ij}} + \sum_{l=1}^{L} \hat{w}_{il}\hat{\chi}_{il} + \eta_{ij} \tag{12}$$

[4] Banerjee, Carlin, and Gelfand discuss that using expected cases in the same line as observed counts violates an endogeneity assumption: the expected number of counts is a fictitious quantity in referece to the observed. They recommend a model that follows $y_{ij} \sim Pois(P_{ij} \cdot \mu_{ij})$, where P_{ij} is the county-year specific population at risk and μ_{ij} is the asthma hospitalization rate, which is to be estimated in the following steps. Alexander et al (2017) do the same in their Bayesian mortality estimation model.

[5] I skipped a step here; an extensive weight-normalizing step is used to filter out unused temporal components. The authors set L to be a high number of components and then let the model select the optimal number of components.

Where η_{ij} is unstructured error.

This paper's main contribution to the literature is on this second stage, where the residuals are decomposed into space-time trends and used as adjustments in re-estimating the model. Banerjee, Carlin, and Gelfand (2003) and Cressie and Wikle (2011) give numerous examples of disease mapping that focus on proper distributional forms for equation (9), using CAR or other spatial priors (the Leroux (2000) model seems to be preferred). Lawson et al's model outperforms this approach in simulation (in Lawson's paper, this is the second model, referenced as Knorr-Held (2000)).

The MIAR distribution, also called the MCAR (multivariate conditionally autoregressive) distribution, suffers from an estimation issue not addressed by the authors. In brief, the MIAR distribution represents an entire vector of values, each of which is normally distributed at the average of its neighbors. This technique would be used to model, for example, multiple types of cancer jointly through a CAR model. Jin, Banerjee, and Carlin (2007) note that in coding this model in a Bayesian estimation software like BUGS or JAGS, this vector of variables has to be given an arbitrary order, which can affect the posterior estimates. They, along with Banerjee, Carlin, and Gelfand (2003), recommend more generalized models that lack this arbitrary ordering.

This paper demonstrates a 'change-of-support,' where data are collected at two different spatial resolutions. In this situation, going from point data where PM2.5 levels are measured to counties is a simple integration. the BCG and Haining books discus other situations where it may be necessary to re-aggregate data. This is a common problem in spatial analysis, especially with different data sources; while administrative data (mortality, socioeconomics) are often at county or state levels, ecological or environmental data are often recorded at point locations. Sometimes data are collected at different, non-nesting levels; for example, zip codes frequently do not align with municipal boundaries and commonly overlap areas.

The authors compare their new model with a series of competitors: a simple Poisson model, a space-time random effects model, and a mixture model. All of these models yield a negative coefficient for PM2.5, indicating that the data variables chosen are responsible for this contradictory result instead of the modeling technique. They remark that past medical research often indicates a positive association between PM2.5 and asthma. It is possible that the socioeconomic variables they control for (race, median income, unemployment rate) are co-linear with and masking the tree effect of PM2.5 rates.

I am interested in extending Alexander, Zagheni, and Barbieri's (2017) Bayesian mortality model to include spatial effects in addition to temporal effects. Their model considers age, year, and location patterns in mortality; however, their hierarchical model considers each geographic subunit to be iid within the higher level (for example, counties within states). One option is to put a CAR prior on county error terms; however, this would fail to incorporate space-time interactions. Using Lawson et al's approach could provide one option for estimating spatio-temporal trends. They argue, in agreement with Wakefield et al (2019) and Wikle et al (1998) that space and time need to be considered jointly as opposed to additively, with separate random effects for space and time. Especially in a disease context, where diseases have an ordered temporal process of transmission to people in their proximity, considering space and time as a single process is necessary at a fine areal and temporal resolution.