Univariate and Multivariate Conditionally Autoregressive Spatial Modeling in the Analysis of Areal Population Data

Kuo-Ping Li, Chirayath Suchindran
University of North Carolina at Chapel Hill

Long Abstract

Areal data such as census data of counties are widely used in population researches. However, the spatial dependence inherent in these data is often ignored by researchers, causing biases in all statistical inferences. Thus, models that consider spatial correlation are desirable. In this study, we applied two hierarchical Bayesian models that account for spatial correlation in the analysis of county population data. The conditionally autoregressive (CAR) model was used to deduct the dependence of population of age 0-17 and age 65+ on the household income of 100 North Carolina counties. This model successfully accounted for spatial clustering effects. The multivariate CAR (MCAR) model was used to access the spatial correlation between two outcomes (the two population proportions). In some situations MCAR models tend to overweight the spatial correlation. Although caution in application should be exercised, MCAR models can be very useful in the analysis of multivariate areal data.

Introduction

Areal data are those data that represent characteristics of discrete areas such as counties and census tracts. Typically the set of areas are defined by geopolitical boundaries. Areal data widely used in population researches, as well as other fields such as public health, political science, and education.

One key property associated with areal data can be stated in the paraphrased “first low of geography”: Everything is related to everything else, but near things are more related than far things”. However, this spatial dependence inherent in areal data is often ignored by researchers, partly due to their unawareness, and partly due to the lack of statistical tools that are able to address it. Nevertheless, this blind ignorance can cause biases in all statistical inferences using the data. Thus, models that consider spatial correlation are very desirable for areal data.

In general, spatial models offer two types of benefit. The first considers the loss of information introduced by spatial correlation relative to independent samples of the same size. In this situation, incorporating the correlation into the model improves the accuracy of statistical inference. Another benefit of a spatial model derives from the view that spatial correlation itself is as an extra source of information. For example, when analyzing incidence counts of a rare disease, spatial correlation can be used to stabilize low counts in an area by “borrowing strength” from nearby areas. In both cases, the researcher recognizes that data came from neighboring areas are often more correlated

than non-neighboring areas. This underlying correlation structure needed to be considered in order to obtain valid inferences.

Methods

We often want to study the relationship between a certain covariate and an outcome. Suppose we have *n* counties, each has an outcome y_i, and we want to relate it to a county characteristic x_i. The simplest way is to write down a linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i ,$$

where β_0 is the intercept and β_1 is the slope. The residue term ε_i is usually treated as being independent and identically distributed. Our goal is to modify the error term so that it accounts for the correlation between neighboring areas.

We can construct a hierarchical model by splitting the error term into two terms:

$$\varepsilon_i = \phi_i + \theta_i ,$$

where ϕi is the spatial random effect term and θ_i is the non-spatial random effect term. The non-spatial random effect term is normally distributed with zero mean:

$$\theta_i \sim \mathrm{N}(0, {\sigma_{nonsp}}^2) .$$

The spatial random effect term ϕi is the essence of the spatial model. We begin to construct a spatial conditionally autoregressive (CAR) model by recognizing that being spatially correlated means a quantity arising from an area is dependent on those from neighboring areas. This concept can be formulated in terms of conditional distributions:

$$Y_i \mid y_{j \in \mathrm{Nbr}} \sim \mathrm{N}(\sum_{j \in \mathrm{Nbr}} b_{ij} y_j , {\sigma_{sp}}^2) ,$$

Where Y_i is the random variable of outcome of area i, $y_{j \in \mathrm{Nbr}}$ are the outcomes of neighboring areas, and coefficients b_{ij} represent the "strength" of the effect of area j to area i.

In practice, the "neighbors" can be defined in many different ways, such as adjacency, distance, whether two areas are connected by main roads, similar socioeconomic status (SES), etc. In this study we define neighbors as those areas that are adjacent to the area in discussion. Also notice that the conditional distribution takes a normal form. This is an extension of the central limit theorem. Because there are a large number of possible pathways or mechanisms that an area can by affected by its neighbors, the distribution of the combined effect can take its asymptotic form of a normal distribution.

The full joint distribution of the spatial random effect of all n areas does not have a closed form and is very difficult to compute by traditional statistical methods. However, it can be quite easily evaluated in the Bayesian framework using Markov chain Monte Carlo (MCMC) simulation, the now-standard method for computational Bayesian analysis.

In the multivariate situation, we extend the univariate CAR to multivariate CAR (MCAR) by replacing the normal distribution in the conditional distribution above by a multivariate normal (MVN) distribution. The variance ${\sigma_{sp}}^2$ is replaced by a variance-

covariance matrix Σ. Suppose we have two outcomes, A and B. Now the outcome A of an area is on condition of not only the outcome A but also outcome B of its neighboring areas. The computation steps is similar to those of CAR, though.

Analysis, Results, and Discussion

We illustrate our CAR and MCAR models using the proportion of young population (age 0-17) and old population (age 65+) of the 100 counties of the state of North Carolina as the outcomes, and these counties' median household incomes as the covariate. These data were from the U.S. Census 2000.

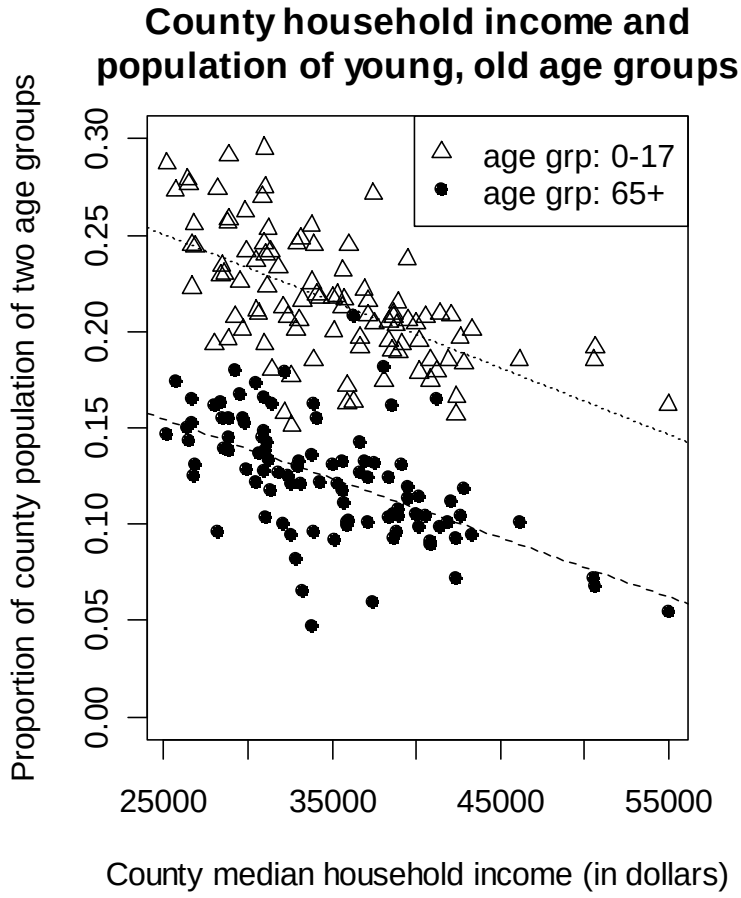

Figure 1. Observed outcomes (proportions of population of young and old age groups) versus the covariate (median household income)

Figure 1 shows the two outcomes versus the covariate. The trend is clear seen on this figure. In general, the higher the median household income, the less young and old population in a county. We used two separate CAR models to estimate the slopes for the two outcomes. The mean and 95% CI of the estimated slope for the young population is -3.8 (-5.9, -2.0), all per million dollars. The estimated slope for the old population is -2.9 (-5.0, -0.99). For both outcomes the null hypothesis is rejected. The proportion of young population has slightly stronger negative dependence on the median household income. This is consistent with Figure 1.

We estimated the relative contribution to the total random effect from spatial random effect. It was estimated that spatial random effect accounted for 68% (95% CI: 61%, 75%) of the total random effect for the proportion of the young population, and 67% (59%, 73%) for the old population. For both outcomes the spatial random effect is more profound than non-spatial random effect.

We used the MCAR model to evaluate both slopes in a single model. Unfortunately it yielded poor results: both slopes are close to zero. This can be explained by noting that MCAR put more weights on the spatial correlations, because it relates an outcome to not only itself but the other outcome of neighboring areas. This probably forced the MCAR model to absorb as much variations of outcomes into the spatial random effect as possible, and therefore overlooked the effect of the covariate. This suggests that caution should be exercised when using MCAR for inferences for individual outcomes.

However, we found that the MCAR model is useful in assessing the correlation coefficients of two outcomes. By removing the regression term from the MCAR model but retain the spatial term, we are effectively “spatially smoothing” the outcome surfaces. The resulted correlation coefficient between the spatially smoothed outcomes is 0.94. For comparison, the correlation coefficient calculated from the observed data is 0.13. The value of 0.94 after spatial smoothing is quite big compared to the value of 0.13 for the observed data. This is because the after-spatial-smoothing correlation coefficient is not a measure of the correlation between individual outcome (data points in Figure 1), but the two spatially smoothed trends (the two dashed lines in Figure 1). This suggests MCAR is effective in filtering out noises while retaining the underlying spatial trends and detecting the relationship between the trends of different outcomes. Therefore, MCAR can be very useful in the analysis of multivariate areal data.