Analyzing Bus Ridership with a Spatial Direct Demand Model
Raven McKnight; Eric Lind, PhD
TRBAM-21-04281

1. Novel Demographic Predictors

Metro Transit is the primary transit service provider in the Twin Cities (Minneapolis-Saint Paul, MN) region. We fit a spatial Bayesian direct demand model of Metro Transit’s bus ridership. The motivations for this type of modeling are (1) conduct geographic smoothing to handle known spatial challenges with transit data and (2) to leverage more of what we know about ridership and the region, both in terms of demographics and spatial information.

The model was fit using 2018 American Community Survey (ACS) variables and 2019 Automatic Passenger Count (APC) data for ridership. The APC data was aggregated to Census block groups to get average weekday boardings per Census block group. We found three major benefits to modeling transit ridership with a spatially explicit model.

- Predictors, it can be reproduced/repeated in a matter of hours at low cost. Because the model is fit using open-source software and readily available predictors, it can be reproduced/repeated in a matter of hours at low cost.
- Spatial Bayesian models can be repeated more easily than many traditional transit demand models.

2. Solving Spatial Problems

Spatial methods solve several common problems with direct demand models. Bus ridership data is inherently spatial because boardings happen at fixed locations. Additionally, we expect some spatial autocorrelation in ridership data. Therefore, traditional statistical methods, which assume independence, are not suitable for modeling transit data.

Statistically, spatial models better represent transit data by explicitly modeling the spatial autocorrelation present.

Another problem occurs when we aggregate bus stops into Census geographies, like block groups. Census geographies tend to be bounded by major roadways. This means that bus stops are often aggregated counterintuitively (Figure 2).

3. Reproducible and Repeatable

Because the model is fit using open-source software and readily available predictors, it can be reproduced/repeated in a matter of hours at low cost.

Methods

This is a spatial Bayesian model fit using Stan via R. The model is a reparameterization of the Besag-York-Mollie (BHM) model, which is a spatial extension of a Poisson regression. The BHM can be written

$Y_i \sim \text{Poisson}(\lambda_i)$

$\log(\lambda_i) = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_n x_{ni} + \theta_i + \phi_i$

where $Y_i$ refers to the average number of weekday boardings in block group $i$ and $\lambda_i$ refers to the average number of weekday departures. Each $x_{ji}$ refers to a predictor, such as population density, and each $\beta_j$ is the corresponding coefficient estimate. In this model, $\theta_i$ is an error term and $\phi_i$ is a spatial error term.

The coefficient estimates and non-spatial errors all get standard Normal priors. The spatial error term gets a Conditional Autoregressive (CAR) prior, which performs the geographic smoothing described in Figure 3. The CAR prior is based on an adjacency matrix $W$ which encodes spatial relationships between block groups. For block groups $i$ and $j$, $w_{ij} = 1$ if the block groups share a boundary and 0 if they do not. This is how spatial information comes into the model and adjusts predictions based on neighbors.

Ongoing Work

- Updating the model to reflect ridership during the COVID-19 pandemic
- Using the model to analyze specific routes and geographies
- Incorporating new demographic predictors (percent renters) into agency-wide surveys and analyses

Acknowledgements

Thank you to Professor Alicia Johnson and the Macalester College Mathematics, Statistics, and Computer Science Department for supporting this research as an honors thesis.