Analyzing Bus Ridership with a Spatial Direct Demand Model

Raven McKnight; Eric Lind, PhD

TRBAM-21-04281

Metro Transit

Background

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.**

1. Novel Demographic Predictors

This model uses a selection of predictors from the 2018 American Community Survey as well as land-use characteristics such as employment density. Using a wider array of demographic variables allowed us to identify novel predictors of bus ridership in the region. For example, the strongest predictor of ridership in the model is **percent of housing units occupied by renters** (Figure 1).

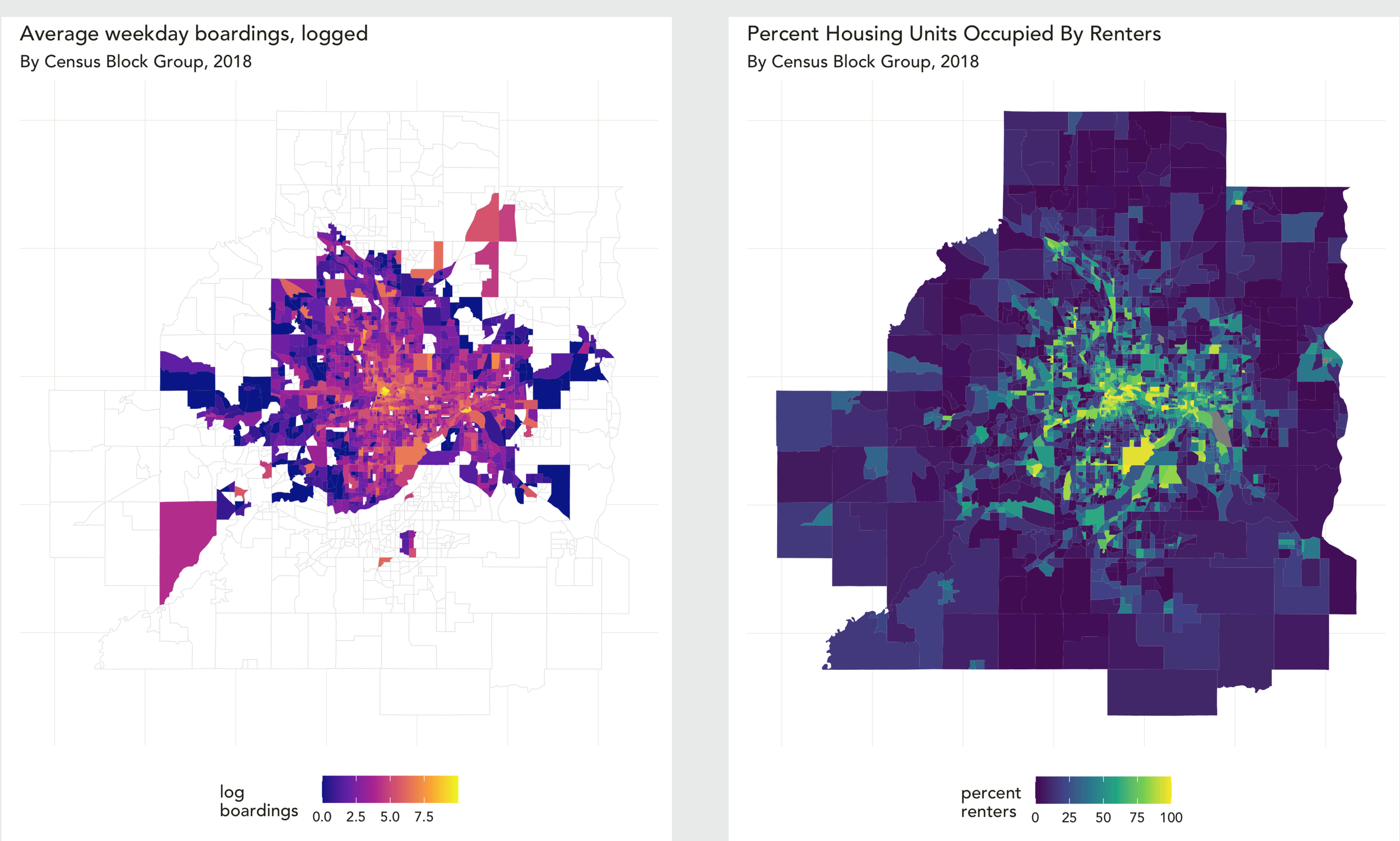

Figure 1, Left: Map of average weekday boardings in the 7 County Minneapolis-Saint Paul Metropolitan Area (logged due to high outliers in downtown Minneapolis). Block groups with no fill were not served by Metro Transit in 2019; Right: Percent of housing units occupied by renters in 2018. Note the correlation with boardings.

Percent of housing units occupied by renters is a **strong predictor** of bus ridership.

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).

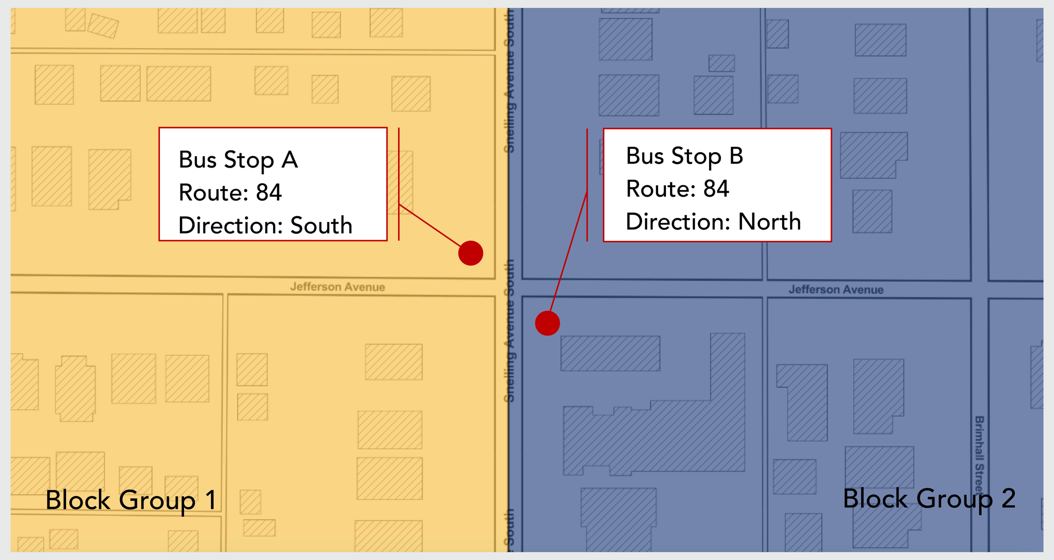

Figure 2: One example of counterintuitive aggregation in the Minneapolis-Saint Paul metropolitan area.

This problem is called the **Modifiable Areal Unit Problem (MAUP)** and means that the results of an analysis may change depending on the unit of aggregation. Our model **"shares information"** between block groups (Figure 4). This is also called **"geographic smoothing"** and it can help solve the MAUP, build a stronger model, and better represent rider behavior.

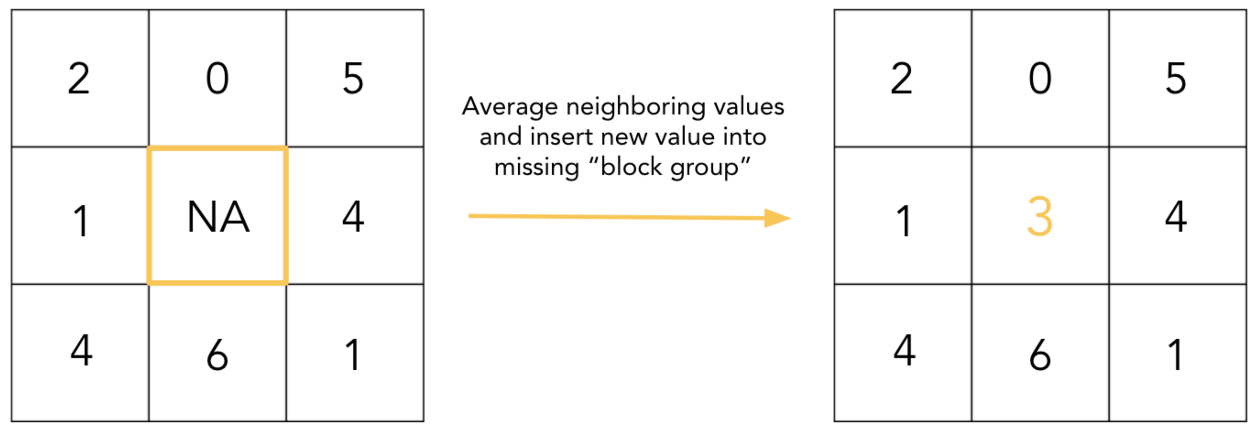

Figure 3: Simple representation of how the direct demand model shares information between block groups.

Spatial models **address the Modifiable Areal Unit Problem** when transit data is aggregated to Census geography.

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.

Spatial Bayesian models can be repeated more easily than many traditional transit demand models

Methods

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

$$Y_i \sim \text{Poisson}(E_i \lambda_i)$$
$$\log(\lambda_i) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \theta_i + \phi_i)$$

where Y_i refers to the average number of weekday boardings in block group i and E_i refers to the average number of weekday departures. Each x_i refers to a predictor, such as population density, and each β_i is the corresponding coefficient estimate. In this model, θ_i is an error term and ϕ_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.