Twin Cities Bus Ridership: A Spatial Bayesian Analysis

Raven McKnight

Faculty Advisor: Professor Alicia Johnson

Department of Mathematics, Statistics, and Computer Science

May 5, 2020

Abstract

Transit agencies often use simple models to predict transit demand. In this study, we explore a series of models for Metro Transit's bus ridership in the Twin Cities region of Minnesota which potentially improve upon the current industry standard. Specifically, we implement Bayesian models which incorporate (1) demographic predictors; and (2) spatial components which are not often included in transit demand models. We consider regularized horseshoe priors to determine which demographic predictors are most relevant to ridership but ultimately do not include the shrinkage priors. We incorporate spatial structure using the BYM2, a reparameterization of the classical Besag-York-Mollié spatial Bayesian model. We present an application of the BYM2 to 2017 Metro Transit bus ridership. The BYM2 is an areal data model which uses an Intrinsic Conditional Autoregressive (ICAR) prior and allows us to build a more nuanced understanding of bus ridership. We discuss specific interpretations of the BYM2 for Metro Transit and the benefits of statistically rigorous models at transit agencies.

Acknowledgements

First and foremost, I would like to thank Alicia Johnson, Eric Lind, and Brianna Heggeseth for their help on this project. This thesis would not exist without their teaching and encouragement. I would also like to thank Suzanne Burr and the entire MSCS department for offering so much support throughout this project and beyond. Before finding MSCS, I would never have imagined myself doing work like this. Thank you to my friends, classmates, roommates, and fellow honors students for being part of this crazy journey. Thanks especially to Angel for doing all of this with me! Finally, a huge thank you to my parents, grandparents, and entire extended family for all the support that got me here. Who would've thought?

Contents

List of Figures **4**

List of Tables **5**

1 Introduction **7**

2 Background **8**
2.1 Transit Market Areas 9
2.2 Motivation 9

3 Data **10**
3.1 Ridership Data 10
3.2 Covariates 11

4 Models **12**
4.1 Model 1: Baseline Poisson Regression 13
4.2 Model 2: Horseshoe Priors 14
4.3 Model 3: Overdispersed Poisson Regression 18
4.4 Model 4: BYM2 20
4.4.1 (Intrinsic) Conditional Autoregressive Priors 20
4.4.2 BYM2 22

5 Results **24**

6 Conclusions **26**

7 Appendix **28**
7.1 Model 1 28
7.2 Model 2 29
7.3 Model 3 31
7.4 Model 4 32
7.5 Model 5 34

References **37**

List of Figures

1 Annual ridership by mode. 8
2 Metro Transit's Transit Market Areas from Metropolitan Council's Transportation Planning Policy, Appendix G. See References for link to resource. 9
3 Average weekday boardings by Census block group, colored on the log scale. Boardings occur in 1,495 of 2,085 block groups in the 7-county metro area. 11
4 Posterior predictive check for Model 1. 14
5 The horseshoe prior resembles a Beta(1/2, 1/2) distribution for fixed values $\tau = \lambda_k = 1$ 15
6 Posterior predictive check for Model 2. 17
7 Posterior predictive check for Model 3. 20
8 Posterior predictive check for Model 4. 24
9 Observed versus Fitted boardings in Saint Paul, Model 4. 26
10 Residuals for Model 4. 26

List of Tables

1 List of covariates 12
2 List of covariates "shrunk" by Model 2. 18
3 List of positively and negatively correlated covariates, Model 4 25
4 Stan output for Model 1 29
5 Stan output for Model 2 31
6 Stan output for Model 3. 32
7 Stan output for Model 4. 34
8 Stan output for Model 5 36

1 Introduction

Public transit ridership has been in decline around the country for years, but particularly since the mid 2010s ("Public transport is in decline in many wealthy cities" 2018). Following a spike in ridership related to the 2008 housing market crash, ridership has been trending steadily downwards in most metropolitan areas over the past decade. Local bus ridership is particularly affected compared to light rail and rapid bus lines, which tend to offer more frequent service. Transit agencies are naturally interested in understanding the specific causes of these declines, as well as more generally being able to predict transit demand in their region.

Transit ridership, or demand, is often forecast using urban planning models. For example, four-step transportation models simulate trips and their spatial distribution in terms of origin and destination before assigning them modes, such as bus versus bicycle, and routes. Activity-based models predict transportation demand based on when and where riders are likely to conduct various tasks (commuting, shopping, recreation, etc). Other transportation models are based largely on land-use characteristics under the general assumption that different types of development (or lack thereof) are more or less likely to support ridership. Models of these types are generally used by transit agencies, including Metro Transit.

Metro Transit is the primary transit provider in the Minneapolis-Saint Paul Metropolitan Area. The agency is a division of the Metropolitan Council and the majority of its service is within the 7-county metro area (Anoka, Carver, Dakota, Hennepin, Ramsey, Washington, and Scott counties). The agency operates one commuter rail line (the Northstar, extending into Sherburne and Wright counties), two light rail lines (Blue Line and Green Line), and over 100 bus routes.

Since the 2014 opening of the Green line, light rail ridership has increased each year (Figure 1). Between the Blue and Green Lines, Metro Transit averaged approximately 72,000 average daily weekday rides in 2017 (Rider's Almanac Blog 2017). In the same year, buses, including the "bus rapid transit" A-Line, averaged around 190,000 daily weekday rides. Despite making up a larger portion of total rides, bus ridership is in decline. According to the Metro Transit Riders' Almanac blog, ridership loss has been particularly acute on the busiest urban local routes in downtown Minneapolis. The Riders' Almanac blog also states that situational factors, such as lower-than-average gas prices and a system-wide fare increase in 2017, have been major causes of bus ridership loss (Rider's Almanac Blog 2015, Rider's Almanac Blog 2016, Rider's Almanac Blog 2017, Rider's Almanac Blog 2018, Rider's Almanac Blog 2019).

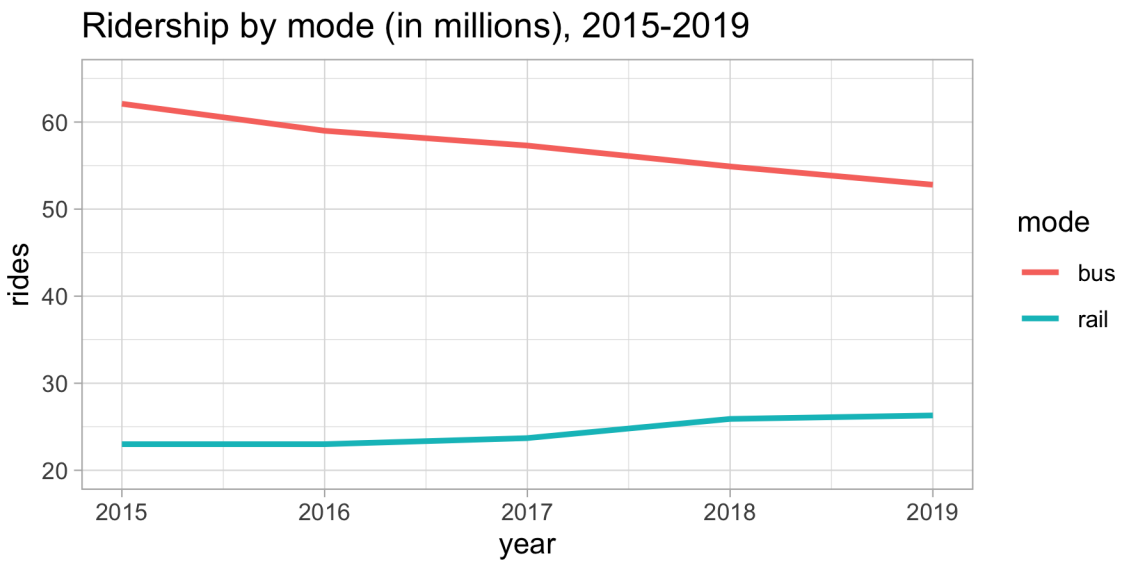

Figure 1: Annual ridership by mode.

Of course, Metro Transit is interested in better understanding these ridership trends. Unfortunately, temporal trends are challenging to model beyond basic patterns. Many temporal variables which may affect ridership loss are difficult or impossible to collect. For example, data scientists at Metro Transit suspect that increasingly flexible work schedules (working from home 2 days per week, teleworking full time, etc) may contribute to ridership loss because bus service caters specifically to 9-to-5 commuters. Other temporal variables, such as more commuters biking to work or choosing ride sharing apps over transit are similarly hard to quantify. Although detailed temporal models are limited by data availability, there are still ways to improve upon the urban planning models described above.

This study focuses on incorporating two often overlooked features of ridership into rigorous statistical models. First, this study incorporates demographic predictors not traditionally included in urban planning models. It is valuable to understand who is utilizing transit service and where ridership will occur. Second, this study incorporates spatial information. Bus ridership is inherently spatial: boardings and alightings can only occur at fixed bus stops. This reality has not often been leveraged in ridership models at the agency. By incorporating these two features, we hope to build a more nuanced understanding of bus ridership in the Twin Cities.

Section 2 of this paper outlines an existing method for predicting ridership at Metro Transit and provides further motivation. Section 3 describes the data used in this study. The bulk of this paper is in Section 4, which explores a series of models of bus ridership. We start with a Bayesian Poisson regression and build to a reparameterization of the spatial Besag-York-Mollié model. Section 5 discusses the results of each model and Section 6 provides brief conclusions about what transit agencies can learn from rigorous modeling. Finally, the Appendix contains full Stan computational code and output for each model [1].

2 Background

Metro Transit has various models of transit ridership for internal and external use. This section describes one spatial approach to modeling ridership published in Metro Transit's Transportation

[1] All of the code used to conduct this study is available in a Github repository at https://github.com/ravenmcknight/honors-project-bym2

Planning Policy (TPP) and offers motivation for expanding upon the existing methodology.

2.1 Transit Market Areas

Transit Market Areas (Figure 2) are written into Metro Transit's official Transportation Planning Policy, or TPP (2040 Transportation Policy Plan). The Market Areas are a method of predicting transit ridership across space. Market Area 1 is expected to support fast, frequent, all-day service, Market Area 2 slightly less service, and so on. Market Area 5 is expected to support peak-hour express commuter services, or park-and-ride lots.

Figure 2: Metro Transit's Transit Market Areas from Metropolitan Council's Transportation Planning Policy, Appendix G. See References for link to resource.

The methodology to determine Market Areas is fairly straightforward. A linear regression of average weekday boardings and alightings (where passengers enter and exit the bus, respectively) was fit in 2013. The resulting coefficients were used to define a formula which determines a "Transit Market Index" (TMI) for each Census block group in the 7-county area. Census block groups are an areal unit often used in transit modeling. Each Census tract, a more familiar geography, is made up of several block groups. Once each block group is assigned a TMI value, they are binned into 5 groups, each corresponding to one Market Area. Following the notation of the TPP, the formula for determining the TMI for each block group is expressed as

$$\begin{aligned}\text{TMI} =& 0.65 * \text{Population Density} + 0.20 * \text{Employment Density} + \\ & 0.11 * \text{Automobile Deficit} + 0.23 * \text{Intersection Density}\end{aligned}$$

where each predictor is logged and scaled by developed land acreage. This TMI model is largely a land-use model which determines riderhsip based on the type of development in each block group.

Population and employment density are straightforward predictors: ridership generally increases with density of people and jobs. Here, automobile deficit refers to the total number of adults over age 16 less the total number of automobiles available in a block group. Intersection density is a proxy for walkability, with the assumption that well-gridded streets with many regular intersections are conducive to walking, and therefore, to transit ridership. There is an additional indicator variable, omitted in this notation, for the Census block group containing the MSP International Airport.

Each block group in the 7-county metropolitan area is assigned a Market Area by the TMI process above. Once each block group has been assigned, the Market Areas are smoothed using Geographic Information Systems (GIS) to produce more or less concentric Market Areas. This geographic smoothing yields the map in Figure 2. Further documentation of the official Transit Market Areas and their interpretations are available in the TPP's Appendix G.

2.2 Motivation

There are several reasons we are interested in improving the methodology described above. First, there is evidence to suggest that demographic predictors beyond simple population density may be informative in predicting bus ridership. Previous studies of Metro Transit's riders have shown that transit riders tend to be more diverse than the background population of the Twin Cities in terms of race, income, age, and so on (Metro Transit 2019). While it is challenging to collect demographics from individual bus riders, we hope to leverage demographic characteristics of the Twin Cities to predict areas of high ridership. Spatially aggregated data is the most feasible way to include such demographic information.

Including spatial information has other benefits. One primary goal of spatial modeling is to conduct geographic smoothing. Rather than producing results and later smoothing, as described above,

we integrate spatial information into the structure of the model itself. This may be particularly helpful in the case of bus ridership data because boardings can only occur at existing bus stops. We hope to improve predictions in block groups which are outliers or which currently have no bus stops by incorporating information about neighbouring block groups.

Census geography itself is a limitation in modeling bus ridership. Geographies such as block groups tend to be bounded by major roadways. Unfortunately, bus routes tend to lie on those same roadways. Because of these bounds, bus stops may be split into block groups counterintuitively. For example, bus stops on either side of the same street travelling in opposite directions are often in different geographies. In these cases, boardings occur on the boundaries of block groups. There may be flaws in assigning demographics to ridership based on this method of spatial aggregation. The full implications of this limitation are not explored in this study, but some work has been done to define "transit-specific geographies" which could alleviate this problem (Viggiano 2017). In this study, we use spatial smoothing via modeling to mitigate bounding issues.

3 Data

This analysis relies on ridership data provided by Metro Transit as well as demographic information from the American Community Survey (ACS) and the Longitudinal Employer Household Dynamic Survey (LEHD). In combination, these data sources allow us to explore the demographics of bus ridership in the Twin Cities. This section briefly describes the data provided by Metro Transit and outlines the data preparation process.

3.1 Ridership Data

Metro Transit provided three data sources for this analysis. The primary data source, **Automatic Passenger Count (APC)** data, corresponds to each time a bus stops and lets passengers on or off the vehicle. When a bus opens its doors, two beams of light detect movement and count the number of riders entering and exiting.

This analysis began by retrieving APC data from January 1, 2015 through December 31, 2018, which yielded 493 billion cases. In theory, these 493 billion cases correspond to *all* trips run from 2015 to 2018, and all stops on those trips. However, we know that, for a variety of reasons, this is untrue. The automatic data reporting technology is flawed and can fail to report data unpredictably. Additionally, the trips with missing data may not be random: all buses of one "series", for example, may fail to report at once. This could lead to the APC data set underestimating boardings on a particular set of routes or locations. Therefore, we use two additional Metro Transit data sources to create a more complete APC data set.

Metro Transit's Schedule details the *planned* schedule for each day of service. We use this data set to determine which trips *should* have run on a given day and locate trips which are missing APC data. **Automatic Vehicle Location (AVL)** data is reported by in-service vehicles approximately every eight seconds. The vehicle reports its GPS location continuously while in service. We can use this data set to determine which scheduled trips actually occurred, even if they are missing from the APC data set. If a trip appears on the schedule but has no APC or AVL data, we assume the trip did not run. If the trip is missing APC data but appears in AVL records, we estimate boardings for that trip by taking the average number of boardings for trips of that type. We interoplate based on trips with the same stop, route, direction, service type (weekend vs weekday), time of day, season, and, when possible, date.

After this interpolation, each row in the augmented APC data set corresponds to a bus stopping, opening its doors, and allowing passengers to enter or exit. We aggregate this data from individual bus stops to Census block groups. This is, in large part, to match the granularity of demographic data available for modeling. Additionally, many transit-demand studies in the literature use Census block groups as the spatial unit of analysis.

In addition to spatial aggregation, we aggregate boardings to an average number of weekday boardings in 2017. The resulting distribution is right-skewed with a few large outliers as we see below (Figure 3, Figure 4). The aggregation a) greatly reduces the number of rows of data in play, and therefore computation time, and b) allows us to build a spatial model without simultaneously incorporating temporal trends. We use 2017 data because at the time of this study, 2017 was the most recent year with complete ACS data available. This average number of weekday boardings per block group will be the response variable for all models presented in this paper. Figure 3 shows this response variable on a map of the 7-county metro area. Note that the response variable is logged in Figure 3 for visual clarity, and that block groups missing from the map had no ridership in 2017.

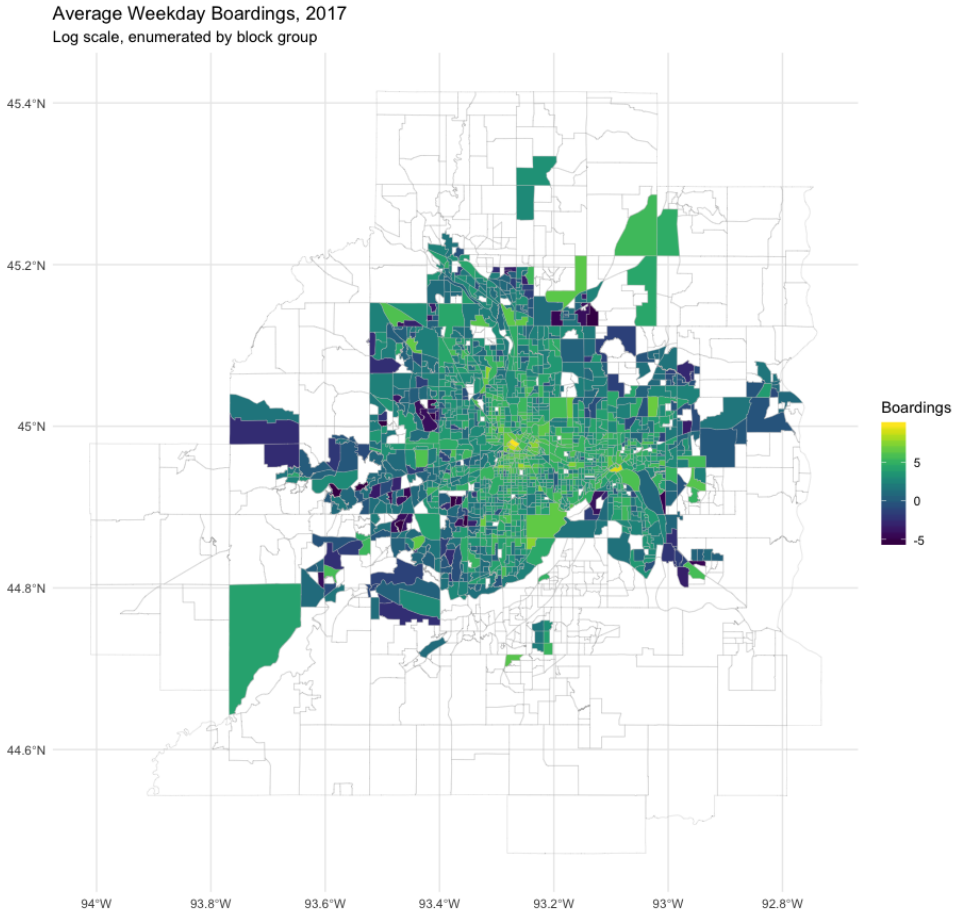

Figure 3: Average weekday boardings by Census block group, colored on the log scale. Boardings occur in 1,495 of 2,085 block groups in the 7-county metro area.

3.2 Covariates

All other data sources used in this study are from non-Metro Transit sources. Census block groups are sourced from TIGER/Line Shapefiles via the `tigris` R package. Model covariates, listed in Table 1, are primarily from the 2017 American Community Survey (ACS) 5-Year Estimates. Additional variables related to employment are sourced from the 2017 Longitudinal Employer-Household Dynamic Survey (LEHD). All covariates are scaled to have a mean of zero and a standard deviation of one. This step is necessary because covariates included in this study are on very different scales.

Covariate	Meaning	Source
x_1	median household income	ACS 2017
x_2	percent residents identified as white alone	ACS 2017
x_3	area of 10 minute walk isochrone from population-weighted block group centroid	OpenStreetMap
x_4	percent residents with a high school diploma	ACS 2017
x_5	percent residents with a bachelor's degree	ACS 2017
x_6	percent of residents who rent their residence	ACS 2017
x_7	percent of residents without a vehicle	ACS 2017
x_8	percent of residents who speak only English	ACS 2017
x_9	percent of residents identified as foreign	ACS 2017
x_{10}	employment density	LEHD 2017
x_{11}	percent of jobs for white employees	LEHD 2017
x_{12}	percent of jobs for male employees	LEHD 2017
x_{13}	percent of jobs for employees with no college degree	LEHD 2017
x_{14}	percent of jobs for employees who make less than $40,000	LEHD 2017
x_{15}	percent of jobs for employees under age 30	LEHD 2017
x_{16}	median age	ACS 2017
x_{17}	percent residents who use transit to commute to work	ACS 2017
x_{18}	population density	ACS 2017

Table 1: List of covariates

These covariates were chosen based on existing Metro Transit knowledge. Several of the variables correspond directly to the Transit Market Area predictors described in Section 2: population and employment density, for example. We replace the intersection density variable with x_3, a measure of walkability generated in OpenTripPlanner. A 10 minute walk isochrone measures the area a person on foot can cover in 10 minutes on the existing street network. We use the area of that isochrone to approximate walkablity. Unfortunately, this metric can be fooled by rural areas where you can walk uninterrupted for 10 minutes down a single roadway, for example. An improved "walkability" metric could improve granular models of transit ridership.

Nearly all studies of transit ridership incorporate some measure of employment density. Few incorporate anything more detailed than number of jobs per acre or square mile. This study incorporates more specific employment characteristics, such as percent of employees under the age of 30, from the LEHD Survey. These variables are included to better identify groups of riders in addition to areas of high ridership. There are certainly other demographic characteristics which could be considered; these 18 variables represent Metro Transit's intuition about bus ridership and could easily be expanded.

4 Models

In this section, we explore several models of Metro Transit's bus ridership, starting with a baseline Poisson regression and building to a spatial model. All models share the same response variable, average number of weekday boardings per Census block group, and the same set of covariates. Additionally, all models are fit in the Bayesian framework using Stan version 2.19.1 via the `rstan` R package.

4.1 Model 1: Baseline Poisson Regression

For each block group $i = 1, 2, \ldots, 1495$, the "baseline" model is a Bayesian Poisson regression. For ridership Y_i and covariates $x_{i1}, ..., x_{i18}$ described above, the Poisson regression is written

$$Y_i \sim \text{Poisson}(E_i \lambda_i)$$
$$log(\lambda_i) = \beta_0 + \sum_{k=1}^{18} x_{ik}\beta_k$$
$$\beta_0, \beta_k \sim \text{Normal}(0, 1)$$

The structure of Model 1 assumes that the mean of the Poisson distribution can be modeled by the latent function $\beta_0 + \sum_{k=1}^{18} x_{ik}\beta_k$ for regression coefficients β_0 and $\beta_1, \beta_2, \ldots, \beta_{18}$.

In this basic Poisson regression, we give β_0 and all β_k standard normal priors which are weakly informative. Because covariates x_k are standardized to have mean zero and standard deviation one, we expect the associated parameters β_k to be relatively close to zero.

The **offset** E_i is often termed "exposure" in applications such as disease risk modeling. In our case, E_i is the number of times a bus stops on an average weekday in block group i. Intuitively, the offset term accounts for "supply" and normalizes the rate λ_i. More technically, the offset term scales the Poisson output to be a rate (number of boardings per stop) rather than a count (total number of boardings). This allows us to make more direct comparisons across block groups with different levels of existing bus service. Mathematically, the offset is a covariate with parameter set equal to 1. Recall, the Poisson regression assumes we can model the expected value of Y with a linear combination of predictors:

$$log(E(Y_i)) = \beta_0 + \sum_{k=1}^{18} x_{ik}\beta_k$$

Therefore, if we wish to model the rate Y/E, the equation is rewritten

$$log(E(Y_i)/E_i) = \beta_0^{'} + \sum_{k=1}^{18} x_{ik}\beta_k^{'}$$
$$log(E(Y_i)) = log(E_i) + \beta_0^{'} + \sum_{k=1}^{18} x_{ik}\beta_k^{'}$$

The baseline Poisson regression, Model 1, was fit on four chains of 10,000 iterations each using the Markov chain Monte Carlo (MCMC) methods implemented in Stan. The Stan code to fit Model 1 and a complete table of coefficient estimates are available in the Appendix. A posterior predictive check reveals the flaws in this model. Figure 4 shows the observed distribution of Y in dark blue with "fitted values" in light blue. These fitted values are simulated sets of data from the posterior of Model 1. If the model is performing well, we expect the simulated data to closely follow the distribution of our observed Y_i. Model 1 clearly underestimates the number of block groups with low ridership and overestimates the number of boardings in block groups with medium to high ridership. This is largely due to extra Poisson variance present in the ridership data which we will address in later models. First, however, we will explore variable selection.

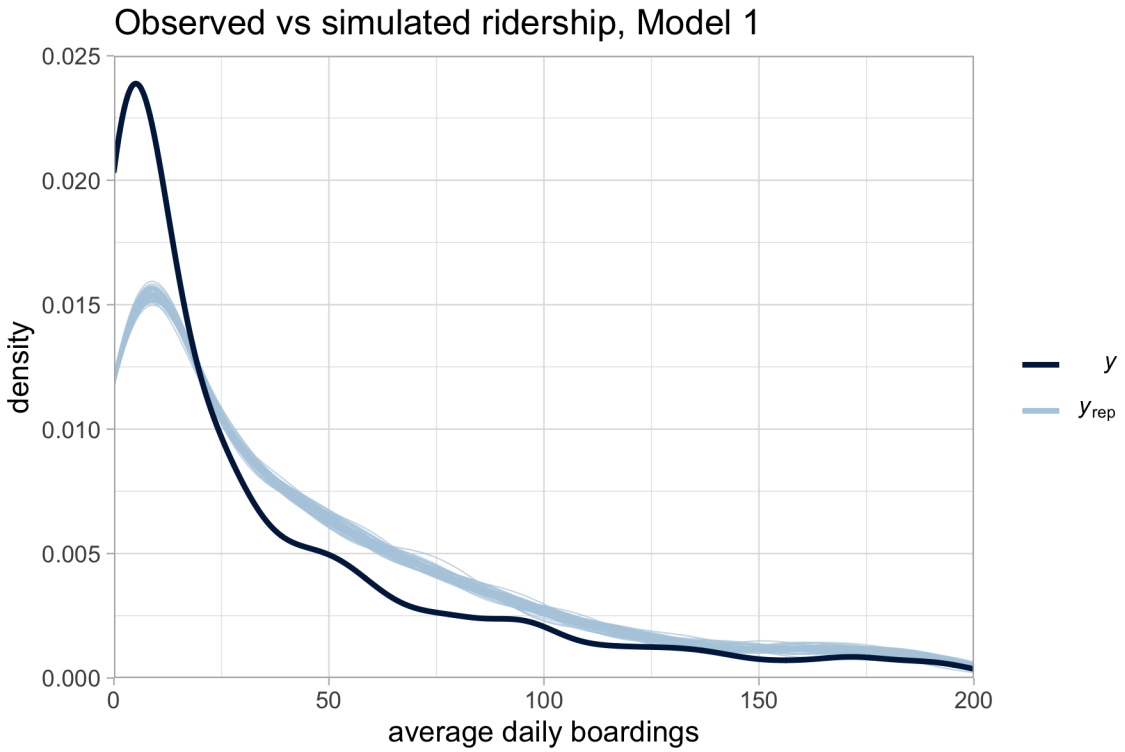

Figure 4: Posterior predictive check for Model 1.

4.2 Model 2: Horseshoe Priors

In Model 1, we give coefficients $\beta_1, \ldots, \beta_{18}$ simple normal priors. With a large set of possible covariates, however, we may reasonably expect some of the coefficients β_k to be equal to zero – in other words, that some of the predictors x_k are not significantly associated with response Y when controlling for other model predictors. In this section, we explore a method of variable selection using shrinkage priors to determine which β_k have the most significant impacts on Y.

In the Bayesian setting, variable selection is conducted by assigning β_k a prior with most of its mass close to zero. Such a prior forces coefficients towards zero unless the associated coefficient is significantly different than zero. Erp et al (2019) provide an introduction to many continuous priors with this feature. Ridge regression and the Bayesian LASSO are two of the most approachable and commonly used shrinkage priors (Erp, Oberski, and Mulder 2019).

The horseshoe prior, proposed by Carvalho et al (2010) is a popular choice in the Bayesian literature, in part because it is similar to the "gold standard" of Bayesian variable selection, the spike-and-slab (Carvalho, Polson, and Scott 2010, Piironen and Vehtari 2017). The spike-and-slab is a discrete mixture model and some recent work, including development of the horseshoe prior, approximates a continuous version of the spike-and-slab. Following the notation of Piironen and Vehtari (2017), the horseshoe prior can be expressed as

$$\begin{aligned} \beta_k | \lambda_k, \tau &\sim \text{Normal}(0, \tau^2 \lambda_k^2) \\ \lambda_k &\sim \text{Half-Cauchy}(0, 1) \end{aligned}$$

for $k = 1, 2, ..., 18$. The horseshoe prior is so-named because for fixed values $\tau = \lambda_k = 1$, the prior resembles a Beta(1/2, 1/2) distribution, or a horseshoe (Figure 5). The shape of this distribution

illustrates how the horseshoe favors values of β_k which are either zero or significantly far from zero. In the case of $\tau = \lambda_k = 1$, "far" from zero is around one (Figure 5).

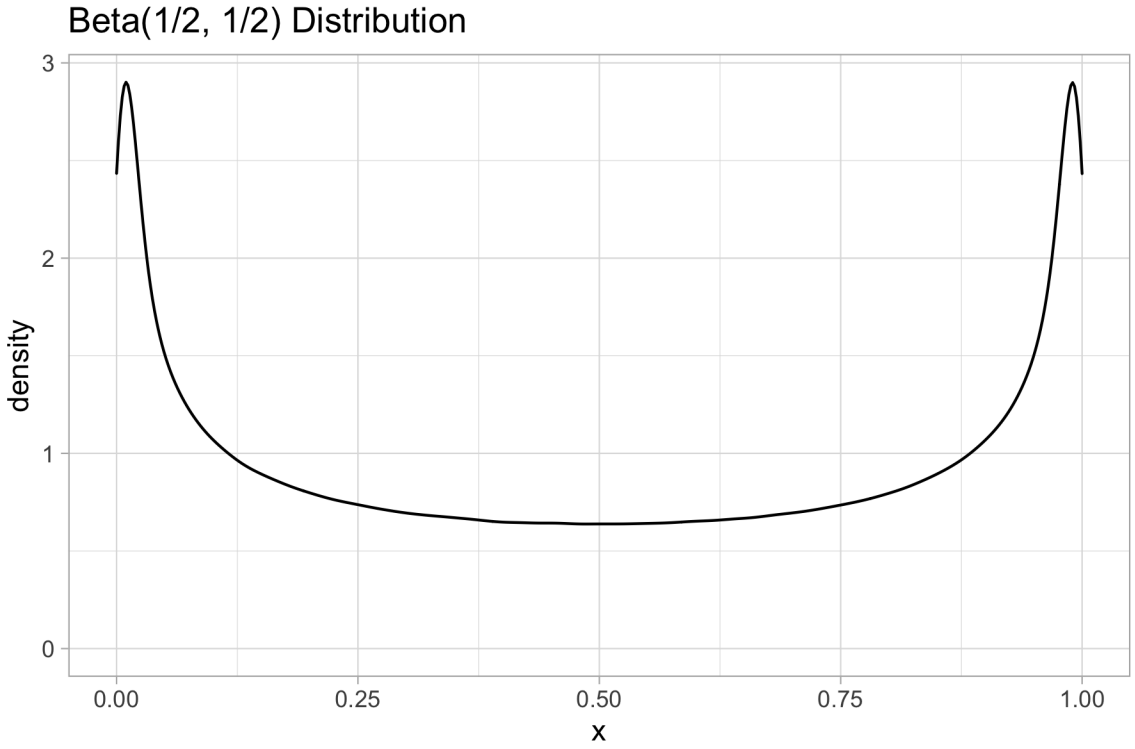

Figure 5: The horseshoe prior resembles a Beta(1/2, 1/2) distribution for fixed values $\tau = \lambda_k = 1$

The horseshoe is often favored because it is a global-local shrinkage prior. This means that τ shrinks *all* parameters towards zero while the local parameter λ_k and its heavy Cauchy tails allow larger coefficients to remain unshrunk. While this is precisely the goal of a sparsity inducing prior, the horseshoe prior can fail to regularize large coefficients *at all*, which can be a problem when parameters are weakly identified by data. With no regularization applied to the largest coefficients, there is a risk of overfitting. Additionally, there is no consensus in the literature for assigning priors to τ. Piironen and Vehtari (2017) introduce the **regularized horseshoe** to address both of these shortcomings. The regularized horseshoe builds upon the standard horseshoe such that

$$\begin{aligned}
\beta_k|\lambda_j, \tau, c &\sim \text{Normal}(0, \tau^2\tilde{\lambda_k^2}) \\
\tilde{\lambda_k^2} &= \frac{c^2\lambda_k^2}{c^2 + \tau^2\lambda_k^2} \\
\lambda_k &\sim \text{Half-Cauchy}(0, 1) \\
c^2 &\sim \text{Inverse-Gamma}(a, b)
\end{aligned}$$

where $c > 0$ helps to regularize β_k far from zero. When c approaches infinity, the regularized horseshoe becomes the standard horseshoe. Additionally, c in the regularized horseshoe corresponds to the slab width in the spike-and-slab prior. Note that when $a = v/s$ and $b = vs^2/2$, the distributuion of c can be rewritten as Student-t$_v(0, s^2)$.

Piironen and Vehtari (2017) provide a set of recommendations for setting priors which we follow in this study. Specifically, they provide guidance for setting a prior for τ. Rather than fixing τ, or

giving it a non-informative prior, they recommend giving τ a Half-Cauchy distribution with location zero and scale equal to τ_0. The scale parameter τ_0 is determined by the number of parameters we expect to be significantly different from zero, or the effective number of nonzero parameters. The scale is calculated as follows:

$$\tau_0 = \frac{p_0}{k - p_0} \frac{\sigma}{\sqrt{n}}$$

where p_0 is an estimate for the number of nonzero parameters, and k is the total number of parameters in the model. Piironen and Vehtari find that even a crude estimate of p_0 improves accuracy and computation time. We follow these recommendations below.

We fit Model 2 using code adapted from Piironen and Vehtari (2017). Like Model 1, we fit Model 2 on four chains of 10,000 iterations each. Because many of our covariates are correlated, we assume a priori that approximately half of the 18 covariates described in Table 1 are relevant to ridership. With $p_0 = 8$, we have $\tau_0 \approx 0.002$. The horseshoe priors increase the computational complexity, and therefore computation time, of Model 2 significantly.

It is crucial to note that Model 2 experiences several divergent transitions after warmup. Stan uses a Hamiltonian Monte Carlo sampling algorithm. Divergent transitions occur when the simulated Hamiltonian departs from the path of the theoretical Hamiltonian. Generally, we should not trust posterior estimates from models with divergent transitions (Stan Development Team 2020a). However "false alarm" divergent transitions can occur when the cutoff for determining a divergent transition is too low (Vehtari, Gelman, and Gabry 2017). Because only eight of 20,000 post-warmup iterations encountered divergent transitions, we can proceed with caution using Model 2. The parameter estimates produced by Model 2 (see Appendix) closely match those produced by Model 1, suggesting that the few divergent transitions did not have a significant adverse effect on posterior estimates.

The posterior predictive check for Model 2 (Figure 6) matches Model 1 almost perfectly - this is to be expected. The addition of horseshoe priors on β_k does not attempt to improve model fit, but to determine which covariates are most influential.

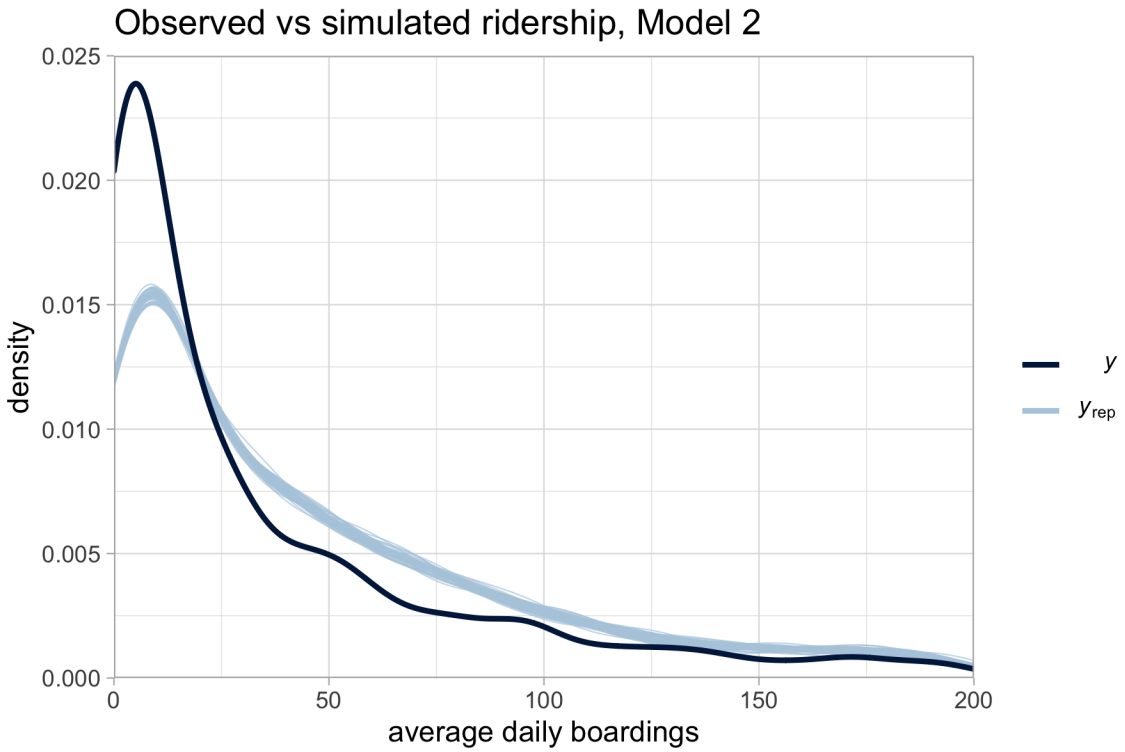

Figure 6: Posterior predictive check for Model 2.

The regularized horseshoe cannot shrink coefficients exactly to zero, so we must choose a "cutoff" point t to determine which covariates to leave out of the model. There are multiple reasonable options for this cutoff. Because the x_k in this setting are normalized, small coefficients are relatively influential. Based on these small magnitudes, we might select $t = 0.025$ or $t = 0.05$ as cutoffs. Table 2 summarizes which variables x_k would be included based on each of those cutoffs.

Covariate	Meaning	t = 0.025	t = 0.05
x_1	median household income	cut	cut
x_2	percent residents identified as white alone	include	include
x_3	area of 10 minute walk isochrone	include	cut
x_4	percent residents with a high school diploma	include	cut
x_5	percent residents with a bachelor's degree	cut	cut
x_6	percent of residents who rent their residence	include	include
x_7	percent of residents without a vehicle	cut	cut
x_8	percent of residents who speak only English	cut	cut
x_9	percent of residents identified as foreign	include	include
x_{10}	employment density	include	include
x_{11}	percent of jobs in block group for white employees	include	include
x_{12}	percent of jobs in block group for male employees	include	include
x_{13}	percent of jobs in block group with no college degree	cut	cut
x_{14}	percent of jobs in block group which make less than $40,000	include	cut
x_{15}	percent of jobs in block group for employees under age 30	include	include
x_{16}	median age	include	include
x_{17}	percent residents who use transit to commute to work	include	cut
x_{18}	population density	include	cut

Table 2: List of covariates "shrunk" by Model 2.

Based on our prior, with $p_0 = 8$, we expect around eight covariates to remain in the model. A cutoff of 0.05 leaves exactly eight variables in the model (a 0.025 cutoff leaves 13). There are five variables shrunk under both cutoff rules: household income, percent of residents with a bachelors degree, percent of residents without a vehicle, percent of residents who speak only English, and percent of jobs for employees with no college degree. Some of these variables are clearly correlated with other x_k which remain in the model. For example, percent of residents identified as foreign (x_9) remains in the model while percent of residents who only speak English (x_8) is shrunk. In this case, the shrinkage prior is somewhat helpful in determining which of two similar predictors is more influential. However, there are likely simpler avenues for making these distinctions.

Notably, we likely would have drawn similar conclusions about which variables to include and exclude based on Model 1. The variables whose 95% credible intervals cross zero in Model 1 (x_1, x_8) are also shrunk by Model 2. Model 2 *does* consistently shrink the coefficients of "irrelevant" variables, such as x_1 and x_8, to be closer to 0.

Although the regularized horseshoe is performing shrinkage, we expect the added computational complexity and sampling problems (specifically divergent transitions) to outweigh the benefits of shrinkage in future models. While the horseshoe may be valuable in other settings, we proceed with simpler priors. We briefly describe uses for shrinkage priors in Section 6.

4.3 Model 3: Overdispersed Poisson Regression

Models 1 and 2 assume that our response variable follows a Poisson distribution. Specifically, this means that we assume $E(Y) = Var(Y)$. However, in the case of Metro Transit's ridership data, $Var(Y) >> E(Y)$. That is, the variance is significantly larger than the expected value, and the assumption in our Poisson regression is not met. In this case, we call the unexpectedly large variance "overdispersion" or "extra Poisson variance."

Data with this feature can be difficult to model directly with a Poisson regression. One option is to use a generalization of the Poisson distribution, such as a Negative-Binomial, to model the data. For the Metro Transit application, we instead modify the latent function to include a separate term to account for overdispersion. As such, the baseline Poisson regression can be rewritten to include

$$log(\lambda_i) = \beta_0 + \sum_{k=1}^{18} x_{ik}\beta_k + \theta_i\sigma$$
$$\theta_i \sim \text{Normal}(0, 1)$$
$$\sigma \sim \text{Normal}(0, 5)$$

where θ_i is a set of heterogeneous random effects, one for each block group. These random effects adjust the rate λ_i to account for nearly all extra Poisson variance. In practice, we scale θ_i by a single parameter σ to allow easier fitting in Stan, following Morris et al (2019).

Model 3 is *significantly* more computationally expensive to fit than Model 1, but less so than Model 2. Slow sampling is largely due to significant autocorrelation between iterations. The autocorrelation made it necessary to run Model 3 on four chains for 20,000 iterations each. Reparameterization could likely reduce autocorrelation and speed up computation time.

The parameter estimates for β_0 and $\beta_1, \beta_2, \ldots, \beta_{18}$ under Model 3 are in the Appendix. Notably, the direction and magnitude of the coefficient estimates are similar to those produced by Models 1 and 2. Specifically, influential variables such as x_6 remain influential in this setting. The consistency of estimates between models suggests that our coefficients *are* well identified by our data and that shrinkage priors are somewhat unnecessary. That said, we experimented with adding horseshoe priors to Model 3 with limited success. The coefficient estimates were similar to those produced by Model 2, computation time was longer still and the model experienced more divergent transitions than Model 2. Because of these substantial sampling issues, we omit discussion of this intermediate model.

A posterior predictive check reveals the improved fit of Model 3. In Figure 7, we see the simulated values following the observed distribution very closely. The addition of an overdispersion parameter is a simple change which greatly improves model fit. Our final model will decompose the random effects θ_i into spatial and non-spatial components.

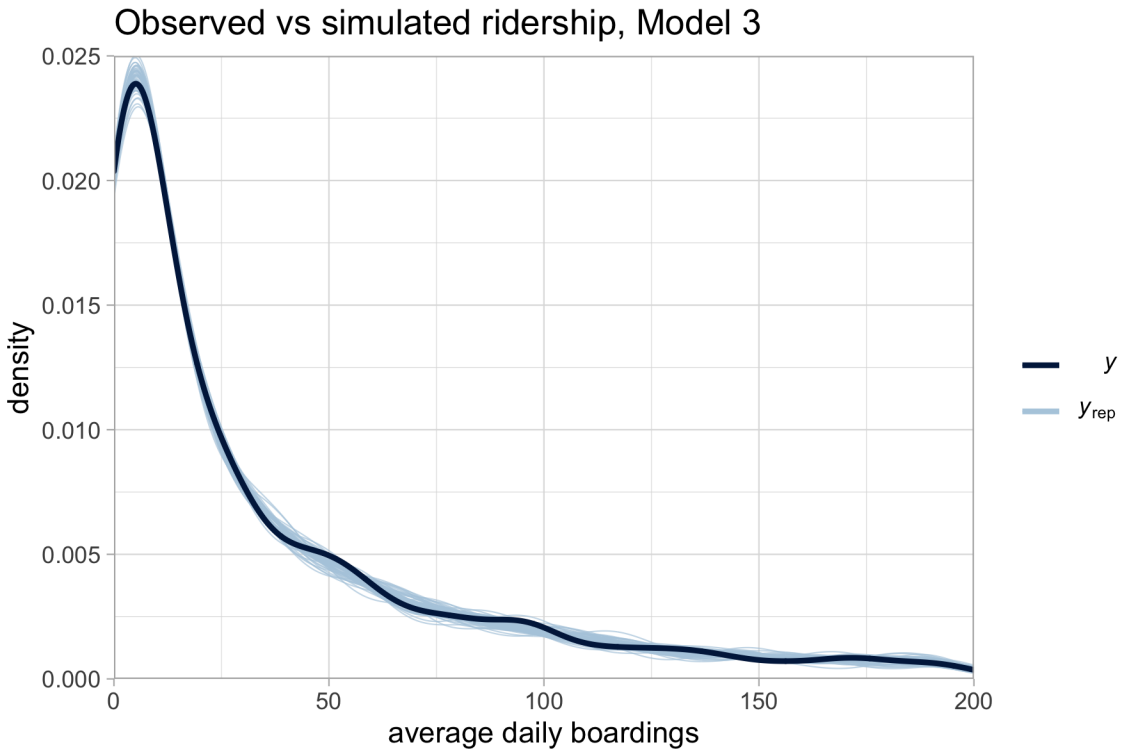

Figure 7: Posterior predictive check for Model 3.

4.4 Model 4: BYM2

The previous three models address the first goal of this study: incorporating additional demographic information. In this section, we address incorporating spatial information into our model of ridership. In short, the goal of this final model is to share information across block groups.

When modeling data with a spatial component, we generally expect adjacent areas to be more similar than areas which are far apart. While this is a straightforward assumption, it can improve model fits in several ways. First, it can encode prior information about response Y not otherwise represented by covariates x_k (ie, that nearby observations are more similar than distant ones). Second, it can provide geographic smoothing when observations are sparse or noisy. This is of particular interest in this study. In the case of Metro Transit's ridership, smoothing can help avoid overfitting areas such as outliers in downtown Minneapolis and better represent block groups with no existing bus stops.

This section will summarize Conditional Autoregressive (CAR) and Intrinsic Conditional Autoregressive (ICAR) priors, common methods for Bayesian areal unit modeling. We specifically detail the Besag-York-Mollié prior and it's recent reparameterization, the BYM2. Finally, we apply the BYM2 to Metro Transit's ridership data, building upon Model 3.

4.4.1 (Intrinsic) Conditional Autoregressive Priors

Conditional Autoregressive (CAR) priors are one of the most widespread Bayesian approaches to spatial modeling. CAR models were introduced in Besag (1974) and remain perhaps the primary method for Bayesian areal data modeling.

Areal data corresponds to finitely many discrete areal units, such as counties or block groups. Areal data tends to be noisy when the counted event is rare, the population of areal unit i is small, or

the physical boundaries of areal units present challenges. CAR priors are often used with small-area data to smooth such noisy counts.

Generally, CAR priors rely on a binary neighborhood structure. Areal units i and j are considered neighbors if they share a boundary. For irregularly shaped areal data, such as our 1,495 Census block groups, we default to a "queen" neighborhood structure wherein areal units sharing *any* points of contact are deemed neighbors. For $n = 1495$ regions, the neighborhood relationships are encoded in a 1495×1495 neighborhood matrix W. Matrix W is defined such that $w_{ij} = 1$ if if regions i and j are neighbors (denoted $i \sim j$) and zero otherwise. Regions are not considered neighbors of themselves. In other words, the diagonal entries $w_{ii} = 0$.

Spatial information enters the model through a set of spatial random effects $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_{1495})$ which are based on the neighborhood relationships described above. We can consider these ϕ_i to be the spatial counterpart to Model 3's θ_i. The distribution of each ϕ_i is determined by the sum of its neighbors values such that

$$\phi_i | \phi_j, j \neq i \sim \text{Normal}\left(\alpha \frac{\sum_{i=1}^{1495} w_{ij}}{d_{ii}} \phi_j, \sigma^2\right)$$

In other words, the spatial effect in block group i is determined by the mean of its neighbors' spatial effects. The parameter $\alpha \in (0, 1)$ indicates the amount of spatial autocorrelation present in Y. Intrinsic Conditional Autoregressive (ICAR) priors simplify CAR priors by setting $\alpha = 1$.

Besag (1974) proved that the joint distribution of $\boldsymbol{\phi}$ is a multivariate normal centered at zero where the variance is given by a symmetric positive definite precision matrix Q. This finding simplifies the above to

$$\phi \sim \text{Normal}(0, Q^{-1})$$

In a classic CAR prior, the precision matrix Q is determined by the neighborhood matrix W and diagonal matrix D in which d_{ii} indicates the number of neighbors region i has. All off-diagonal entries in D are zero. Using these matrices, Q is defined

$$Q = D(I - \alpha W)$$

where I is the identity matrix. Because ICAR priors set $\alpha = 1$, the definition of Q can be simplified as follows:

$$Q = D - W$$

This simplification reduces the computational complexity of fitting a CAR/ICAR prior significantly. The Besag-York-Mollié parameterization used in this study uses an ICAR prior.

The general ICAR prior specifies each ϕ_i to be normally distributed with its mean equal to the mean of its neighbor's values as follows:

$$\text{p}(\phi_i | \phi_{i \sim j}) = \text{Normal}\left(\frac{\sum_{i \sim j} \phi_i}{d_{ii}}, \frac{\sigma_i^2}{d_{ii}}\right)$$

This is how the ICAR prior "borrows strength" from geographic neighbors to smooth noisy estimates. Intuitively, the spatial random effects in block group i will be similar to the spatial random effects in its neighboring block groups. The variance of each ϕ_i decreases as its number of neighbors d_{ii} increases. In other words, we are more confident in estimates for block groups with many neighbors than in block groups with only one or two neighbors, in which case we are not able to borrow as much strength from nearby block groups.

The expression above is helpful for building intuition about how the ICAR prior operates. However, it is not computationally optimal. The joint distribution can be rewritten as a pairwise difference expressed

$$\mathrm{p}(\phi) \propto \exp\left(-\frac{1}{2}\sum_{j:i\sim j}(\phi_i - \phi_j)^2\right)$$

which is significantly more efficient to fit in Stan (Morris et al. 2019). Morris et al (2019) use this pairwise difference formulation and careful parameterization to fit the ICAR prior with fewer computations per iteration.

4.4.2 BYM2

The section above describes ICAR priors in general. In this section, we discuss the specific model used in this analysis, the Besag-York-Mollié (BYM) Poisson model (Besag, York, and Mollié 1991). The BYM is a"classical" spatial Bayesian method (Riebler et al. 2016). The BYM capitalizes on the ICAR prior's ability to smooth noisy estimates and share information across geographic units.

The latent function in the BYM model contains heterogeneous random effects (like Model 3) as well as a spatial ICAR term in order to account for both spatial and non-spatial heterogeneity. The BYM model can be thought of as decomposing Model 3's θ_i into spatial and non-spatial components. Specifically, the BYM model expands the latent function from previous models such that

$$log(\lambda_i) = \beta_0 + \sum_{k=1}^{K} x_{ik}\beta_k + \theta_i + \phi_i$$

where ϕ_i is the addition: an ICAR component.

The BYM model as written above is appealing in its simplicity. However, the lack of informative hyperpriors specified for θ_i and ϕ_i can make the model incredibly challenging to fit. Extra Poisson variance can be explained entirely by θ_i, entirely by ϕ_i, or anywhere in between. In practice, we rarely have information about where that ratio falls. Riebler et al (2016) further explain the sampling issues faced by the BYM model with no hyperpriors. In short, the sampler is forced to explore all possible combinations of θ_i and ϕ_i, no matter how unlikely a particular combination my be. In the context of fitting models with Stan, this is likely to cause unnecessarily long computation times and perhaps to yield biased posterior estimates. There are some suggestions in the literature for setting hyperpriors for θ_i and ϕ_i (Besag, York, and Mollié 1991, Bernardinelli, Clayton, and Montomoli 1995). The existing methods, however, tend to rely on the specific data at hand which makes the selection of hyperparameters unnecessarily challenging.

The BYM2 model described Simpson et al (2017) and Riebler et al (2016) aims to solve the problems with the BYM described above. The latent function of the BYM2 rewrites the BYM as

$$log(\lambda_i) = \beta_0 + \sum_{k=1}^{K} x_{ik}\beta_k + \left(\sqrt{\frac{\rho}{s}}\theta_i^* + \sqrt{1-\rho}\phi_i^*\right)\sigma$$

where the new parameters $\rho, s,$ and σ are included to aid in sampling from the posterior. While the reparameterization appears significantly more complex than the BYM, the additional parameters do not change the structure of the BYM but rather allow Stan to explore the posterior more efficiently. In summary,

- $\rho \in [0,1]$ determines how much variance or overdispersion is caused by spatial versus nonspatial error terms. $\rho = 1$ corresponds to all variance being spatial.
- The term $\left(\sqrt{\frac{\rho}{s}}\theta_i^* + \sqrt{1-\rho}\phi_i^*\right)\sigma$ is a set of convolved random effects. The convolution of θ_i and ϕ_i allows the two sets of random effects to interact more efficiently than in previous parameterizations. This is the primary improvement of the BYM2: the convolution significantly improves computational efficiency and decreases computation time.

- σ is the standard deviation of the entire convolved random effect
- s is a scaling factor such that $\text{Var}(\theta_i) \approx \text{Var}(\phi_i) \approx 1$, an assumption which is necessary for σ to truly be the standard deviation of the error terms.

In addition to improving sampling efficiency, the BYM2 takes a more "fully Bayesian" approach to hyperprior and hyperparameter selection. We follow the recommendations of Riebler et al (2016) and Morris et al (2019) in setting the following priors:

- $\theta_i \sim \text{Normal}(0, n)$ where n is the number of connected subgraphs in the neighborhood graph. In many cases, such as modeling all counties in a state, $n = 1$ and θ_i gets the same $\text{Normal}(0, 1)$ prior it did under Model 3. It is possible to fit the BYM2 with $n > 1$, although the computation time is significantly longer when σ and s must be calculated for each subgraph.
- $\sigma \sim \text{Normal}(0, 1)$
- $\rho \sim \text{Beta}(1/2, 1/2)$. Recall this prior is shaped like a horseshoe (Figure 5). This prior suggests that variance will be mostly spatial or mostly nonspatial and allows the sampler to avoid spending unnecessary time exploring unlikely combinations of θ_i and ϕ_i.

Model 4 builds upon Model 3 by replacing simple heterogenous random effects θ_i with the convolved random effets described above. Model 4 uses all 18 covariates and no horseshoe priors. This model converges much more efficiently than Model 3 due to careful parameterization. Model 4 was run on four chains for 10,000 iterations each. The parameter estimates for Model 4 are in the Appendix.

Once again, the direction and magnitude of coefficient estimates are similar to those produced by previous models. There are several key differences in Model 4's parameter estimates discussed in Section 5. The mixing parameter ρ has a posterior mean of 0.893, indicating that a majority of the variance in Metro Transit's ridership can be explained by spatial autocorrelation. Like Model 3, the BYM2 model performs well in the posterior predictive check (Figure 9).

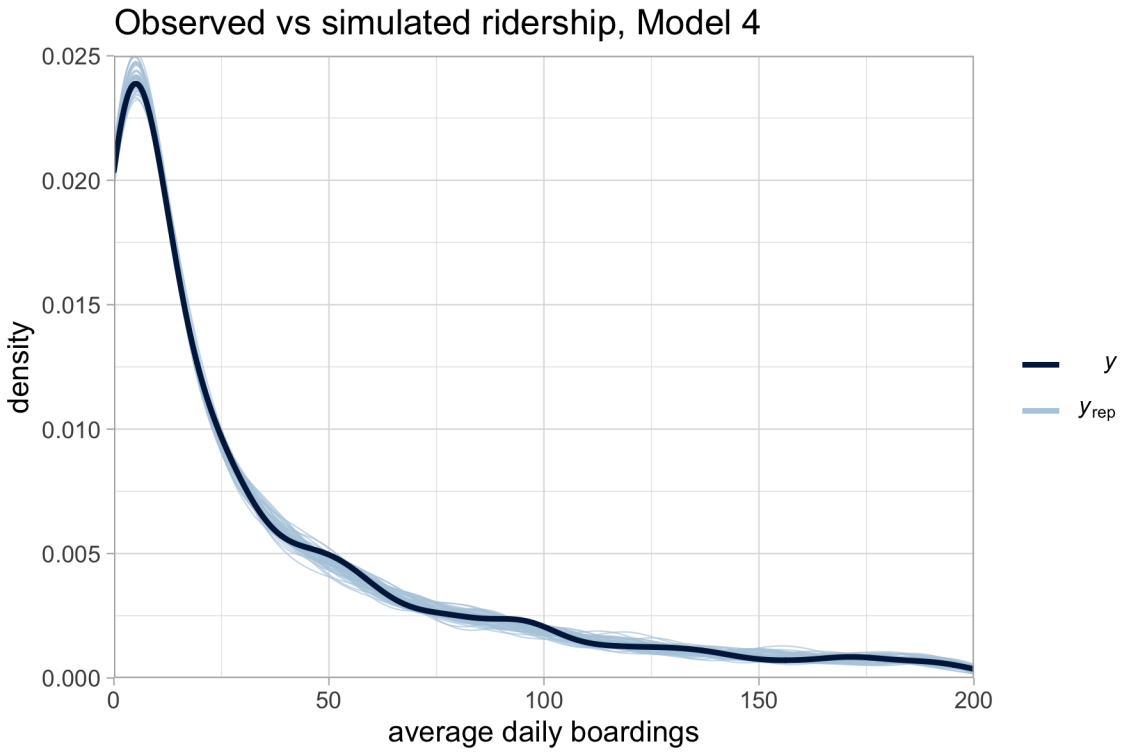

Figure 8: Posterior predictive check for Model 4.

Note that it is possible to fit the BYM2 with horseshoe priors, and the code to do so is included in the Appendix. In the setting of this study, fitting the BYM2 + horseshoe model took more than twice as long as Models 2 or 4 to fit and performed similarly to Model 2 in terms of shrinkage. In a different modeling context, with more "irrelevant" variables, the BYM2 + horseshoe model may be informative enough to be worth the additional computation time. For this study, we omit discussion of this model.

5 Results

In this analysis, we present four models of Metro Transit's bus ridership. The baseline Poisson regression, Model 1, is greatly limited by overdispersion in Metro Transit's ridership data. Model 2 introduces the regularized horseshoe, a shrinkage prior with only moderate success. In the setting of this analysis, the horseshoe has limited utility because the covariates were selected specifically for their relevance to ridership. Model 3, which includes random effects θ_i to account for overdispersion, captures the distribution of the data much more accurately than Models 1 or 2. In this sense, Model 3 achieves the goal of a "good" fit. Model 4, the BYM2, performs similarly to Model 3 in terms of fit. The primary difference between Models 3 and 4 is that Model 4 decomposes overdispersion into a combination of spatial (ϕ_i) and non-spatial (θ_i) errors.

Table 3 summarizes the results of Model 4. These results are generally similar across all models (detailed coeffcient tables are found in the Appendix). Many of these relationships match intuition. For example, consider negative coefficients β_1 and β_{16}, corresponding to median household income and median age: we generally expect older, more affluent residents to own their own vehicles, live in less transit-rich suburbs, and ride the bus less often.

Correlation	Variables
Positive	Percent residents white alone, 10-minute walk isochrone, percent renters, percent speak English only, percent foreign, employment density, percent male employees, percent employees under age 30, percent who use transit to commute percent residents who don't own vehicles
Negative	median household income, percent residents with high school diploma, percent residents with bachelors degree, percent employees white, percent employees no college degree, percent employees earning less than $40,000, median age, population density

Table 3: List of positively and negatively correlated covariates, Model 4

One peculiarity is the consistently negative coefficient for x_{18}, population density. All previous work indicates that population density should have a positive (and significant) effect on ridership. In this case, we suspect that x_6, percent of residents who rent their home, may be capturing something population density alone does not. Areas with high concentrations of renters are largely in the downtowns and around the University of Minnesota campus. A priori, we expect ridership to be higher in all of these areas because of density *and* the presence of more and more frequent transit service. In short, percent renters appears to capture places where population is dense *and* transit exists.

Notably, some β_k change dramatically once spatial autocorrelation is taken into consideration. For example, β_1, associated with median household income, was barely distinguishable from zero under Models 1 and 2. The same coefficient has a posterior mean of -0.134 under Model 4. This tells us a few things about x_1 as a predictor. First, it is likely spatially autocorrelated itself. Second, we know now that median household income is not a meaningful predictor across the entire 7-county metropolitan area. Because income and transit availability vary so much across the region, income alone is not a strong predictor of ridership. Low income areas far away from the urban core may have lower ridership than expected and vice versa. Once we control for location, however, we can see the effects of x_1 more clearly. In other words, median household income is only a meaningful predictor in small areas with similar levels of existing transit service.

Some parameter estimates also change significantly between Models 3 and 4. For example, β_{17}, representing percent of workers who commute via transit, has a posterior mean of 0.114 under Model 3. It's posterior mean is only 0.067 under Model 4. This makes some intuitive sense: residents only have the option to commute via transit in places that *already* provide transit. In this case, the addition of spatial information decreases the importance of predictors which can be explained in part by physical location. The coefficient for x_6, percent of residents who rent their home, follows a similar pattern. The posterior mean of β_6 is 0.241 under Model 3 and only 0.151 under Model 4. This suggests that x_6 as a predictor was largely capturing a spatial pattern which the spatial term ϕ is able to account.

Model 4 performs well in terms of prediction. Figure 11 shows the observed versus fitted boardings within Saint Paul city limits for Model 4. Figure 12 similarly shows the residuals in the city limits. The fit and residuals are similar across the entire 7-county region, Saint Paul is shown as an example. The largest residual in Saint Paul is in the highest ridership block group in the downtown core. Model 4 underestimates ridership in this block group by approximately 130 boardings per day, a relative error of 1.5%. Model 3, on the other hand, overestimates boardings by roughly the same margin. Outliers such as this block group in Saint Paul illuminate the smoothing capabilities of the BYM2.

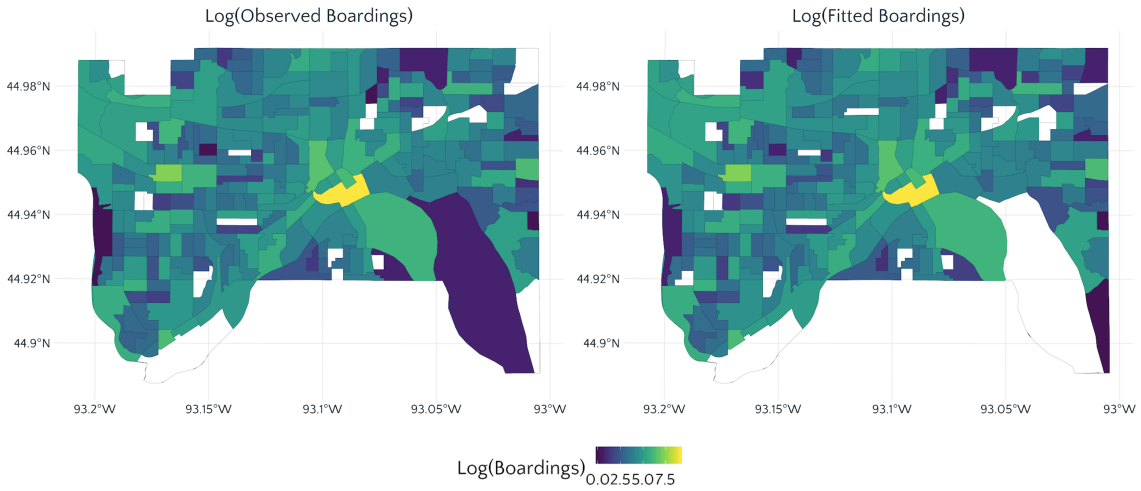

Figure 9: Observed versus Fitted boardings in Saint Paul, Model 4.

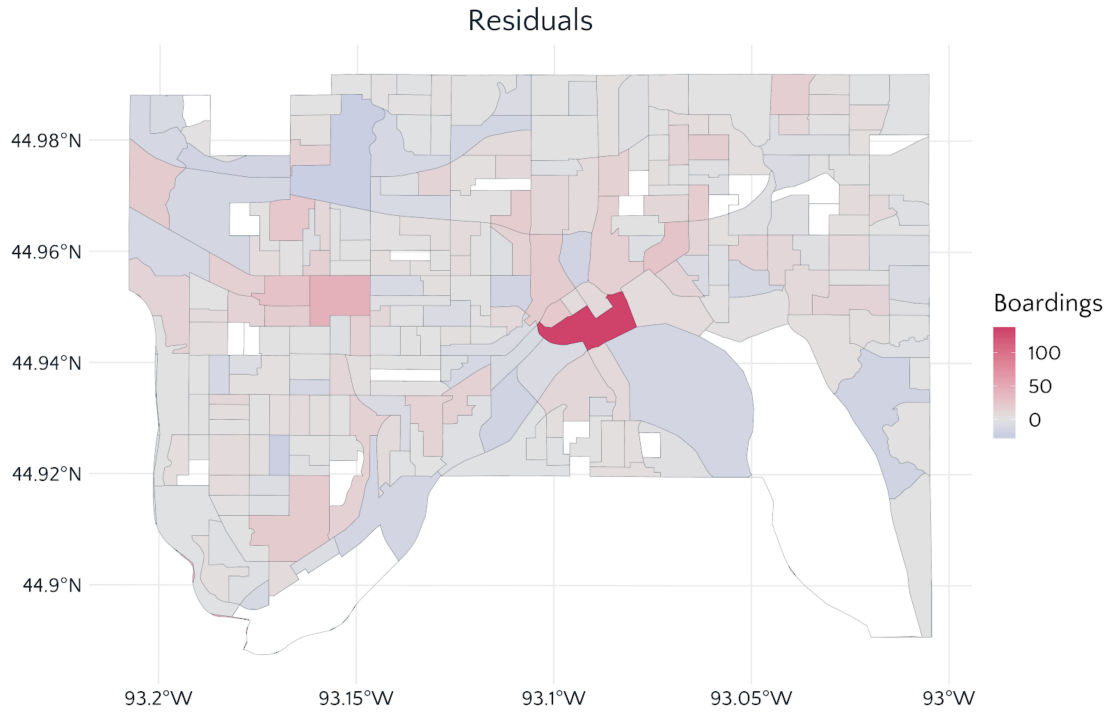

Figure 10: Residuals for Model 4.

6 Conclusions

Transit agencies can benefit from more rigorous modeling in several ways. Simple models such as the TMA model presented in Section 2 are often sufficient for "big picture" planning. In the case

of Metro Transit, the TMA model and maps are often presented to cities within Metro Transit's coverage area to explain how service is allocated across the metro area. However, when it comes to understanding more specific trends, spatial patterns, or demographics, more complex methods are appropriate.

The overdispersed Poisson regression (Model 3) is likely sufficient for many agency-facing projects. It retains some of the simplicity of the TMA model while adding significant explanatory value. Incorporating demographic predictors such as percent of residents without personal vehicles can test the agency's intuition about ridership in the metro area. Vehicle access is a prime example of an intuitive assumption which is challenged by data and modeling. While the percent of residents without personal vehicles is positively correlated with ridership, the magnitude of the correlation is fairly small (posterior mean 0.05 under Model 4). Such findings may encourage agencies to reconsider metrics often used in planning tasks.

When geographic smoothing is desired, spatial models such as the BYM2 (Model 4) are appropriate. Such models may help address known issues with modeling transit ridership across space, including the fixed locations of bus stops, bounding issues with Census geography, and outliers.

The horseshoe prior we explored proved to be somewhat ineffective in this setting. Shrinkage priors may still be a valuable tool for transit modeling, specifically when there are more covariates with less clear relationships to ridership. The prior may be more useful in analyzing survey data or "big data" where there are many variables of unknown importance. The effects of shrinkage will also be more apparent in settings with β_k further from zero.

Future work may include spatial models of transit ridership which treat individual bus stops as point data, rather than aggregating stops to areal units. This question becomes significantly more computationally expensive as we have approximately 13,000 active bus stops in the region versus 1,495 bock groups with ridership. Similarly, alternate models could explore transit-specific geographies or other methods of spatial aggregation. Additionally, the models presented in this paper represent a single point in time. Temporal or spatiotemporal models with similar structures built upon Poisson regressions are possible. Temporal modeling is the next step in helping transit agencies understand the decline in ridership over the past decade. Finally, incorporation of additional spatial, temporal, and demographic predictors is a simple way to continue furthering our understanding of ridership.

7 Appendix

This Appendix contains Stan code and complete parameter estimates for each model described above.

7.1 Model 1

Stan code

```
data {
  // number obs
  int<lower=0> N;
  // response
  int<lower=0> y[N];
  // "offset" (number of stops)
  vector<lower=0>[N] E;
  // number of covariates
  int<lower=1> K;
  // covariates
  matrix[N, K] x;
}
transformed data {
  vector[N] log_E = log(E);
}
parameters {
  // intercept
  real beta_0;
  // covariates
  vector[K] beta;
}
transformed parameters{
  // latent function variables
  vector[N] f;
  f = log_E + beta_0 + x*beta;
}
model {
  /// model
  y ~ poisson_log(f);
  // prior on betas
  beta_0 ~ normal(0.0, 3);
  beta ~ normal(0.0, 1);
}
generated quantities {
  vector[N] log_lik;
    for(i in 1:N) log_lik[i] = poisson_log_lpmf(y[i] | log_E[i] + beta_0 + x[i,]*beta
        );
}
```

Coefficient Table

Parameter	Rhat	n_eff	mean	sd	2.5%	97.5%
β_0	1.000	19365	-1.376	0.005	-1.385	-1.366
β_1	1.000	18085	0.006	0.007	-0.007	0.019
β_2	1.000	19330	0.063	0.005	0.054	0.072
β_3	1.000	18752	-0.027	0.004	-0.035	-0.019
β_4	1.000	18026	-0.029	0.004	-0.038	-0.021
β_5	1.000	18380	-0.013	0.005	-0.022	-0.004
β_6	1.000	15809	0.241	0.005	0.231	0.251
β_7	1.000	18930	0.014	0.003	0.007	0.020
β_8	1.000	17331	-0.003	0.005	-0.012	0.006
β_9	1.000	19360	0.074	0.004	0.067	0.082
β_{10}	1.000	21329	0.066	0.001	0.065	0.068
β_{11}	1.000	20224	-0.121	0.004	-0.128	-0.114
β_{12}	1.000	22539	-0.069	0.004	-0.076	-0.061
β_{13}	1.000	18372	-0.011	0.005	-0.021	-0.001
β_{14}	1.000	15602	-0.026	0.005	-0.035	-0.017
β_{15}	1.000	15480	0.113	0.005	0.104	0.122
β_{16}	1.000	19904	-0.085	0.004	-0.093	-0.078
β_{17}	1.000	21169	0.047	0.003	0.042	0.053
β_{18}	1.000	22032	-0.025	0.003	-0.031	-0.020

Table 4: Stan output for Model 1

7.2 Model 2

Stan Code

This Stan code is adapted from Piironen and Vehtari (2017). We set `scale_global` equal to our estimate of $\tau_0 \approx 0.002$. Setting `nu_global` $= 1$ gives τ a half-Cauchy prior and `nu_local` $= 1$ corresponds to the horseshoe. The parameters `slab_scale` and `slab_df` allow us to modify the Student-t distribution of c which determines the amount of regularization on large coefficients. This code can easily be modified to fit the overdispersed Poisson regression with horseshoe priors by adding to the latent function.

```
data {
  // number obs
  int<lower=0> N;
  // response
  int<lower=0> y[N];
  // "offset" (number of stops)
  vector<lower=0>[N] E;
  // number of covariates
  int<lower=1> K;
  // covariates
  matrix[N, K] x;
  // prior sd for intercept
  real<lower=0> scale_icept;
  // scale for half t for tau
  real<lower=0> scale_global;
  // df for tau, lambda
  real<lower=1> nu_global;
  real<lower=1> nu_local;
  // slab scale, df for reg horseshoe
  real<lower=0> slab_scale;
```

```
21    real<lower=0> slab_df;
22  }
23  transformed data {
24    vector[N] log_E = log(E);
25  }
26  parameters {
27    // intercept
28    real beta_0;
29    // parameterization from vehtari
30    vector[K] z;
31    real<lower=0> aux1_global;
32    real<lower=0> aux2_global;
33    vector<lower=0>[K] aux1_local;
34    vector<lower=0>[K] aux2_local;
35    real<lower=0> caux;
36  }
37  transformed parameters {
38    // global shrinkage
39    real<lower=0> tau;
40    // local shrinkage
41    vector<lower=0>[K] lambda;
42    vector<lower=0>[K] lambda_tilde;
43    // slab scale
44    real<lower=0> c;
45    // regression coefficients
46    vector[K] beta;
47    // latent function variables
48    vector[N] f;
49
50    lambda = aux1_local .* sqrt(aux2_local);
51    tau = aux1_global * sqrt(aux2_global) * scale_global;
52    c = slab_scale * sqrt(caux);
53    lambda_tilde = sqrt(c^2 * square(lambda) ./ (c^2 + tau^2 * square(lambda)));
54    beta = z .* lambda_tilde*tau;
55    f = log_E + beta_0 + x*beta;
56  }
57  model {
58    // half t and inverse gamma priors
59    z ~ normal(0, 1);
60    aux1_local ~ normal(0, 1);
61    aux2_local ~ inv_gamma(0.5*nu_local, 0.5*nu_local);
62    aux1_global ~ normal(0, 1);
63    aux2_global ~ inv_gamma(0.5*nu_global, 0.5*nu_global);
64    caux ~ inv_gamma(0.5*slab_df, 0.5*slab_df);
65    beta_0 ~ normal(0, scale_icept);
66    // and the model
67    y ~ poisson_log(f);
68  }
69  generated quantities {
70    vector[N] log_lik;
71      for(i in 1:N) log_lik[i] = poisson_log_lpmf(y[i] | log_E[i] + beta_0 + x[i,]*beta
          );
72  }
```

Coefficient Table

Parameter	Rhat	n_eff	mean	sd	2.5%	97.5%
β_0	1.000	47507	-1.376	0.005	-1.385	-1.367
β_1	1.000	17963	0.004	0.006	-0.006	0.016
β_2	1.000	19711	0.062	0.004	0.053	0.071
β_3	1.000	21601	-0.027	0.004	-0.035	-0.019
β_4	1.000	18799	-0.029	0.004	-0.037	-0.020
β_5	1.000	16561	-0.011	0.005	-0.020	-0.002
β_6	1.000	20168	0.241	0.005	0.231	0.251
β_7	1.000	18710	0.013	0.003	0.006	0.020
β_8	1.000	18517	-0.002	0.004	-0.011	0.006
β_9	1.000	19590	0.075	0.004	0.067	0.082
β_{10}	1.000	19535	0.066	0.001	0.065	0.068
β_{11}	1.000	20005	-0.120	0.004	-0.127	-0.113
β_{12}	1.000	19085	-0.069	0.004	-0.076	-0.061
β_{13}	1.000	14806	-0.010	0.005	-0.021	-0.000
β_{14}	1.000	20035	-0.026	0.005	-0.035	-0.017
β_{15}	1.000	17978	0.113	0.005	0.103	0.122
β_{16}	1.000	19840	-0.085	0.004	-0.092	-0.077
β_{17}	1.000	19550	0.047	0.003	0.042	0.053
β_{18}	1.000	20124	-0.025	0.003	-0.031	-0.020

Table 5: Stan output for Model 2

7.3 Model 3

Stan Code

This code adds the overdisperion parameter by modifying the latent function from Model 1.

```
data {
  // number obs
  int<lower=0> N;
  // response
  int<lower=0> y[N];
  // "offset" (number of stops)
  vector<lower=0>[N] E;
  // number of covariates
  int<lower=1> K;
  // covariates
  matrix[N, K] x;
}
transformed data {
  vector[N] log_E = log(E);
}
parameters {
  // intercept
  real beta_0;
  // covariates
  vector[K] beta;
  // overdispersion
  vector[N] theta;
  // random effects variance
  real<lower=0> sigma;
}
transformed parameters{
```

```
27    // latent function variables
28    vector[N] re;
29    vector[N] f;
30    re = theta * sigma;
31    f = log_E + beta_0 + x*beta + re;
32  }
33  model {
34    /// model
35    y ~ poisson_log(f);
36    // prior on betas
37    beta_0 ~ normal(0.0, 1);
38    beta ~ normal(0.0, 0.2);
39    // overdispersion
40    theta ~ normal(0, 1);
41    sigma ~ normal(0, 5);
42  }
43  generated quantities {
44    vector[N] log_lik;
45      for(i in 1:N) log_lik[i] = poisson_log_lpmf(y[i] | log_E[i] + beta_0 + x[i,]*beta
            + re[i]);
46  }
```

Coefficient Table

Parameter	Rhat	n_eff	mean	sd	2.5%	97.5%
$beta_0$	1.004	1354	-1.904	0.031	-1.966	-1.843
θ_5	1.000	6986	3.225	0.295	2.624	3.786
σ	1.003	1800	0.985	0.022	0.944	1.029
β_1	1.002	1546	-0.111	0.049	-0.208	-0.016
β_2	1.002	1095	0.004	0.050	-0.094	0.101
β_3	1.003	1322	0.042	0.034	-0.025	0.108
β_4	1.001	1208	-0.029	0.040	-0.106	0.049
β_5	1.002	1282	-0.017	0.042	-0.098	0.067
β_6	1.002	988	0.149	0.047	0.055	0.241
β_7	1.004	1019	0.061	0.039	-0.015	0.139
β_8	1.002	1178	0.048	0.047	-0.046	0.138
β_9	1.004	1125	0.018	0.041	-0.063	0.097
β_{10}	1.001	1796	0.144	0.032	0.080	0.207
β_{11}	1.004	1152	-0.083	0.032	-0.145	-0.020
β_{12}	1.001	1629	0.018	0.030	-0.040	0.077
β_{13}	1.004	2101	-0.020	0.032	-0.082	0.043
β_{14}	1.005	1415	-0.071	0.035	-0.139	-0.001
β_{15}	1.002	1581	0.099	0.033	0.035	0.166
β_{16}	1.004	1160	-0.099	0.036	-0.170	-0.029
β_{17}	1.004	997	0.114	0.034	0.047	0.182
β_{18}	1.002	1589	-0.011	0.030	-0.071	0.048

Table 6: Stan output for Model 3.

7.4 Model 4

Stan Code

This code is adopted from Morris et al (2019). Detailed explanation of the parameterization shown here can be found in their case study.

```
data {
  int<lower=0> N;
  int<lower=0> N_edges;
  int<lower=1, upper=N> node1[N_edges];  // node1[i] adjacent to node2[i]
  int<lower=1, upper=N> node2[N_edges];  // and node1[i] < node2[i]

  int<lower=0> y[N];              // count outcomes
  vector<lower=0>[N] E;           // exposure
  int<lower=1> K;                 // num covariates
  matrix[N, K] x;                 // design matrix

  real<lower=0> scaling_factor; // scales the variance of the spatial effects
}
transformed data {
  vector[N] log_E = log(E);
}
parameters {
  real beta0;            // intercept
  vector[K] beta;        // covariates

  real<lower=0> sigma;        // overall standard deviation
  real<lower=0, upper=1> rho; // proportion unstructured vs. spatially structured
      variance

  vector[N] theta;       // heterogeneous effects
  vector[N] phi;         // spatial effects
}
transformed parameters {
  vector[N] convolved_re;
  vector[N] f;
  // variance of each component should be approximately equal to 1
  convolved_re =  sqrt(1 - rho) * theta + sqrt(rho / scaling_factor) * phi;

  f = log_E + beta0 + x * beta + convolved_re * sigma;
}
model {
  y ~ poisson_log(f);  // co-variates

  // This is the prior for phi! (up to proportionality)
  target += -0.5 * dot_self(phi[node1] - phi[node2]);

  beta0 ~ normal(0.0, 1.0);
  beta ~ normal(0.0, 1.0);
  theta ~ normal(0.0, 1.0);
  sigma ~ normal(0, 1.0);
  rho ~ beta(0.5, 0.5);
  // soft sum-to-zero constraint on phi)
  sum(phi) ~ normal(0, 0.001 * N);  // equivalent to mean(phi) ~ normal(0,0.001)
}
generated quantities {
  vector[N] log_lik;
    for(i in 1:N) log_lik[i] = poisson_log_lpmf(y[i] | log_E[i] + beta0 + x[i, ] *
        beta + convolved_re * sigma);
}
```

Coefficient Table

Parameter	Rhat	n_eff	mean	sd	2.5%	97.5%
β_0	1.000	5803	-1.893	0.029	-1.950	-1.836
ρ	1.004	556	0.893	0.037	0.806	0.943
σ	1.004	552	2.755	0.373	2.050	3.495
θ_5	1.000	20232	-0.195	0.781	-1.743	1.334
ϕ_5	1.002	1150	6.340	1.699	3.526	10.018
β_1	1.000	5929	-0.134	0.051	-0.234	-0.034
β_2	1.001	4670	0.009	0.052	-0.093	0.112
β_3	1.000	4406	0.054	0.036	-0.017	0.125
β_4	1.000	4538	-0.032	0.040	-0.112	0.047
β_5	1.000	4585	-0.044	0.044	-0.130	0.043
β_6	1.000	5071	0.151	0.048	0.057	0.244
β_7	1.001	4621	0.050	0.038	-0.025	0.125
β_8	1.001	5089	0.057	0.047	-0.033	0.148
β_9	1.000	4603	0.049	0.042	-0.034	0.131
β_{10}	1.000	6915	0.146	0.031	0.085	0.208
β_{11}	1.001	6255	-0.084	0.031	-0.145	-0.023
β_{12}	1.000	6891	0.014	0.029	-0.043	0.071
β_{13}	1.000	8267	-0.027	0.031	-0.088	0.035
β_{14}	1.001	6420	-0.069	0.034	-0.135	-0.001
β_{15}	1.000	6277	0.090	0.033	0.027	0.155
β_{16}	1.000	5411	-0.093	0.035	-0.162	-0.025
β_{17}	1.001	2951	0.067	0.040	-0.012	0.146
β_{18}	1.001	6516	-0.027	0.031	-0.088	0.034

Table 7: Stan output for Model 4.

7.5 Model 5

Stan Code

A fifth model was considered, the code for which is shown below. This model combines the horseshoe prior and BYM2.

```
data {
  int<lower=0> N;
  int<lower=0> N_edges;
  int<lower=1, upper=N> node1[N_edges];  // node1[i] adjacent to node2[i]
  int<lower=1, upper=N> node2[N_edges];  // and node1[i] < node2[i]

  int<lower=0> y[N];                // count outcomes
  vector<lower=0>[N] E;             // exposure
  int<lower=1> K;                   // num covariates
  matrix[N, K] x;                   // design matrix

  real<lower=0> scaling_factor; // scales the variance of the spatial effects

  //horseshoe
  // prior sd for intercept
  real<lower=0> scale_icept;
  // scale for half t for tau
  real<lower=0> scale_global;
  // df for tau, lambda
  real<lower=1> nu_global;
```

```
  real<lower=1> nu_local;
  // slab scale, df for reg horseshoe
  real<lower=0> slab_scale;
  real<lower=0> slab_df;
}
transformed data {
  vector[N] log_E = log(E);
}
parameters {
  real beta0;            // intercept

  real<lower=0> sigma;        // overall standard deviation
  real<lower=0, upper=1> rho; // proportion unstructured vs. spatially structured
      variance

  vector[N] theta;       // heterogeneous effects
  vector[N] phi;         // spatial effects

  // horseshoe
  vector[K] z;
  real<lower=0> aux1_global;
  real<lower=0> aux2_global;
  vector<lower=0>[K] aux1_local;
  vector<lower=0>[K] aux2_local;
  real<lower=0> caux;
}
transformed parameters {
  vector[N] convolved_re;

  // horseshoe
    // global shrinkage
  real<lower=0> tau;
  // local shrinkage
  vector<lower=0>[K] lambda;
  vector<lower=0>[K] lambda_tilde;
  // slab scale
  real<lower=0> c;
  // regression coefficients
  vector[K] beta;
  // latent function variables
  vector[N] f;

  lambda = aux1_local .* sqrt(aux2_local);
  tau = aux1_global * sqrt(aux2_global) * scale_global;
  c = slab_scale * sqrt(caux);
  lambda_tilde = sqrt(c^2 * square(lambda) ./ (c^2 + tau^2 * square(lambda)));
  beta = z .* lambda_tilde*tau;

  // variance of each component should be approximately equal to 1
  convolved_re =  sqrt(1 - rho) * theta + sqrt(rho / scaling_factor) * phi;

  f = log_E + beta0 + x * beta + convolved_re * sigma;
}
model {
  y ~ poisson_log(f);  // co-variates

  // This is the prior for phi! (up to proportionality)
  target += -0.5 * dot_self(phi[node1] - phi[node2]);

  beta0 ~ normal(0.0, scale_icept);
  theta ~ normal(0.0, 1.0);
  sigma ~ normal(0, 1.0);
  rho ~ beta(0.5, 0.5);
  // soft sum-to-zero constraint on phi)
  sum(phi) ~ normal(0, 0.001 * N);  // equivalent to mean(phi) ~ normal(0,0.001)
```

```
85
86    //horseshoe
87    // half t and inverse gamma priors
88    z ~ normal(0, 1);
89    aux1_local ~ normal(0, 1);
90    aux2_local ~ inv_gamma(0.5*nu_local, 0.5*nu_local);
91    aux1_global ~ normal(0, 1);
92    aux2_global ~ inv_gamma(0.5*nu_global, 0.5*nu_global);
93    caux ~ inv_gamma(0.5*slab_df, 0.5*slab_df);
94    beta0 ~ normal(0, scale_icept);
95  }
96  generated quantities {
97    vector[N] log_lik;
98      for(i in 1:N) log_lik[i] = poisson_log_lpmf(y[i] | log_E[i] + beta0 + x[i, ] *
          beta + convolved_re * sigma);
99  }
```

Coefficient Table

Parameter	Rhat	n_eff	mean	sd	2.5%	97.5%
β_0	1.000	5837	-1.878	0.028	-1.934	-1.822
σ	1.010	721	2.826	0.359	2.138	3.561
θ_5	1.000	19566	-0.188	0.787	-1.722	1.349
ρ	1.012	713	0.900	0.031	0.824	0.946
ϕ_5	1.004	1652	5.974	1.573	3.264	9.527
β_1	1.001	3637	-0.119	0.059	-0.233	-0.002
β_2	1.001	8170	-0.005	0.027	-0.069	0.049
β_3	1.000	3987	0.024	0.030	-0.019	0.096
β_4	1.000	8531	-0.004	0.022	-0.057	0.040
β_5	1.000	7619	-0.012	0.026	-0.076	0.030
β_6	1.001	2999	0.164	0.055	0.051	0.268
β_7	1.000	3726	0.032	0.035	-0.016	0.114
β_8	1.000	7956	0.010	0.025	-0.032	0.074
β_9	1.000	5622	0.019	0.028	-0.023	0.089
β_{10}	1.000	7530	0.144	0.033	0.079	0.208
β_{11}	1.000	4558	-0.057	0.034	-0.123	0.001
β_{12}	1.000	11150	0.004	0.018	-0.031	0.046
β_{13}	1.000	8289	-0.017	0.024	-0.073	0.020
β_{14}	1.000	5469	-0.028	0.031	-0.100	0.015
β_{15}	1.000	4822	0.060	0.035	-0.001	0.130
β_{16}	1.001	3863	-0.060	0.038	-0.134	0.003
β_{17}	1.000	2094	0.051	0.041	-0.008	0.137
β_{18}	1.000	6949	-0.010	0.021	-0.061	0.026

Table 8: Stan output for Model 5

References

2040 Transportation Policy Plan. "2040 Transportation Policy Plan". https://metrocouncil.org/Transportation/Planning-2/Key-Transportation-Planning-Documents/Transportation-Policy-Plan/The-Adopted-2040-TPP-(1)/Final-2040-Transportation-Policy-Plan/2040-TPP-Appendix-G-Transit-Design-and-Perf-Standa.aspx.

Banerjee, Sudipto, Bradley P. Carlin, and Alan E. Gelfand. 2014. *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press. ISBN: 978-1-4398-1918-0.

Bernardinelli, L, D Clayton, and C Montomoli. 1995. "Bayesian estimates of disease maps: how important are priors?" *Statistics in medicine* 14 (21-22): 2411–2431.

Besag, Julian, Jeremy York, and Annie Mollié. 1991. "Bayesian image restoration, with two applications in spatial statistics". *Annals of the institute of statistical mathematics* 43 (1): 1–20.

Carvalho, Carlos M., Nicholas G. Polson, and James G. Scott. 2010. "The horseshoe estimator for sparse signals". *Biometrika* 97 (2): 465–480. ISSN: 00063444. http://www.jstor.org/stable/25734098.

Erp, Sara van, Daniel L. Oberski, and Joris Mulder. 2019. "Shrinkage priors for Bayesian penalized regression". *Journal of Mathematical Psychology* 89 (): 31–50. ISSN: 0022-2496, visited on 04/03/2020. doi:10.1016/j.jmp.2018.12.004. http://www.sciencedirect.com/science/article/pii/S0022249618300567.

Metro Transit. 2019. "Transit System Performance Evaluation". https://metrocouncil.org/Transportation/Publications-And-Resources/Transit/2019-Transit-System-Performance-Evaluation.aspx.

Morris, Mitzi, et al. 2019. "Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan". *Spatial and Spatio-temporal Epidemiology* 31 (). ISSN: 1877-5845, visited on 04/03/2020. doi:10.1016/j.sste.2019.100301. http://www.sciencedirect.com/science/article/pii/S1877584518301175.

Piironen, Juho, and Aki Vehtari. 2017. "Sparsity information and regularization in the horseshoe and other shrinkage priors". Publisher: The Institute of Mathematical Statistics and the Bernoulli Society, *Electronic Journal of Statistics* 11 (2): 5018–5051. ISSN: 1935-7524, visited on 04/03/2020. doi:10.1214/17-EJS1337SI. https://projecteuclid.org/euclid.ejs/1513306866.

"Public transport is in decline in many wealthy cities". 2018. *The Economist* (). https://www.economist.com/international/2018/06/21/public-transport-is-in-decline-in-many-wealthy-cities.

Rider's Almanac Blog. 2018. "Light rail, Bus Rapid Transit lines set annual ridership records". *Rider's Almanac Blog*. Visited on 04/03/2020. https://www.metrotransit.org/light-rail-bus-rapid-transit-lines-set-annual-ridership-records.

— . 2016. "Metro Transit ridership tops 82.6 million in 2016". *Rider's Almanac Blog*. Visited on 04/03/2020. https://www.metrotransit.org/metro-transit-ridership-tops-826-million-in-2016.

— . 2015. "Metro Transit ridership tops 85.8 million in 2015". *Rider's Almanac Blog*. Visited on 04/03/2020. https://www.metrotransit.org/metro-transit-ridership-tops-858-million-in-2015.

— . 2017. "Rail lines set records as Metro Transit ridership tops 81.9 million in 2017". *Rider's Almanac Blog*. Visited on 04/03/2020. https://www.metrotransit.org/rail-lines-set-records-as-metro-transit-ridership-tops-819-million-in-2017.

— . 2019. "Ridership growing in corridors with fast frequent service". *Rider's Almanac Blog*. Visited on 04/03/2020. https://www.metrotransit.org/ridership-growing-in-corridors-with-fast-frequent-service.

Riebler, Andrea, et al. 2016. "An intuitive Bayesian spatial model for disease mapping that accounts for scaling". *Statistical methods in medical research* 25 (4): 1145–1165.

Stan Development Team. 2020a. "Brief Guide to Stan's Warnings". https://mc-stan.org/article/warnings.html.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". *Statistics and computing* 27 (5): 1413–1432.

Viggiano, Cecilia A. (Cecilia Ann). 2017. "Bus network sketch planning with origin-destination travel data". Thesis, Massachusetts Institute of Technology. Visited on 04/03/2020. https://dspace.mit.edu/handle/1721.1/111441.

R Packages

Bivand, Roger, Tim Keitt, and Barry Rowlingson. 2019. *rgdal: Bindings for the 'Geospatial' Data Abstraction Library*. R package version 1.4-8. https://CRAN.R-project.org/package=rgdal.

Bivand, Roger, and David W. S. Wong. 2018. *Comparing implementations of global and local indicators of spatial association*. 3. https://doi.org/10.1007/s11749-018-0599-x.

Dowle, Matt, and Arun Srinivasan. 2019. *data.table: Extension of 'data.frame'*. R package version 1.12.8. https://CRAN.R-project.org/package=data.table.

Gabry, Jonah, and Tristan Mahr. 2019. *bayesplot: Plotting for Bayesian Models*. R package version 1.7.1. https://mc-stan.org/bayesplot.

Grolemund, Garrett, and Hadley Wickham. 2011. *Dates and Times Made Easy with lubridate*. 3. http://www.jstatsoft.org/v40/i03/.

Henry, Lionel, and Hadley Wickham. 2019. *purrr: Functional Programming Tools*. R package version 0.3.3. https://CRAN.R-project.org/package=purrr.

Hester, Jim, and Hadley Wickham. 2020. *odbc: Connect to ODBC Compatible Databases (using the DBI Interface)*. R package version 1.2.2. https://CRAN.R-project.org/package=odbc.

Pebesma, Edzer. 2018. *Simple Features for R: Standardized Support for Spatial Vector Data*. 1. doi:10.32614/RJ-2018-009. https://doi.org/10.32614/RJ-2018-009.

R Special Interest Group on Databases (R-SIG-DB), Hadley Wickham, and Kirill Müller. 2019. *DBI: R Database Interface*. R package version 1.1.0. https://CRAN.R-project.org/package=DBI.

Stan Development Team. 2020b. *RStan: the R interface to Stan*. R package version 2.19.3. http://mc-stan.org/.

Walker, Kyle. 2020a. *tidycensus: Load US Census Boundary and Attribute Data as 'tidyverse' and 'sf'-Ready Data Frames*. R package version 0.9.6. https://CRAN.R-project.org/package=tidycensus.

— . 2020b. *tigris: Load Census TIGER/Line Shapefiles*. R package version 0.9.2. https://CRAN.R-project.org/package=tigris.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. https://ggplot2.tidyverse.org.

Wickham, Hadley, et al. 2019. *Welcome to the tidyverse*. 43. doi:10.21105/joss.01686.