Bayesian Dynamic Network Modeling with Censored Flow Data

by

Brian Cozzi

Department of Statistical Science
Duke University

Date: _____

Approved:

_____

Mike West, Advisor

_____

David Banks

_____

Li Ma

A thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2020

ABSTRACT

Bayesian Dynamic Network Modeling with Censored Flow Data

by

Brian Cozzi

Department of Statistical Science
Duke University

Date: _____

Approved:

_____

Mike West, Advisor

_____

David Banks

_____

Li Ma

An abstract of a thesis submitted in partial fulfillment of the
requirements for the degree of Master of Science
in the Department of Statistical Science
in the Graduate School of
Duke University

2020

Abstract

There is an abundance of applied and theoretical statistical research focused on the analysis of network data. However, few applications have the flexibility to account for the inherently constrained flow that results from a limited capacity at destination nodes and, thus, may provide an incomplete picture of the underlying data generating process. This thesis works to address this shortcoming by modifying Bayesian Dynamic Flow Modeling for a context where the capacity at the destination node is limited. To that end, it develops a methodology for updating prior beliefs about flow rates when the flow is censored and the posterior does not have a recognizable form. This method is applied to a publicly available bike sharing dataset that exhibits censored flows during high-volume times of the day. The results compare the network characterization from models with and without the censored updating methodology. This highlights specific circumstances in which the estimates of underlying demand from both models are most at odds with one another and provides a framework for guiding the analysis of datasets that can be similarly represented.

Acknowledgements

I would like to thank my advisor, Mike West, for his outstanding mentorship and guidance on this topic and others, as well as David Banks and Li Ma for serving on my committee. I would also like to acknowledge Haohan Chen for sharing his exceptional code for the Dynamic Count Mixture Model and Dynamic Gravity Model that served as the foundation for the application. Additionally, I would like to thank my friends and family for their enduring support (especially when they have indulged me in my fondness of rental bikes). Lastly, I would like to thank the City of Chicago for curating the dataset used in this application.

Contents

List of Figures

Chapter 1

Introduction

There is an abundance of applied and theoretical statistical research focused on the analysis of data which can be represented as a network. The data in these applications span domains such transportation [13, 1, 12], banking, and social networks [3]. In many cases, the goal is to accurately characterize activity which facilitates both forecasting of future activity within these networks and monitoring for anomalies [16]. While there has been rapid growth in the number of applications for these techniques, the growth of the size and complexity of the networks of interest have grown in kind. This often leads to a necessary trade off between the suitability of a model and the scalability of the computations required to fit them.

One common way to cope with this is to model the behavior of each directed edge individually under the assumption of conditional independence. Then, assuming these components share some common underlying structure, these individual models can be aggregated to make inferences about more general behaviors of the network. Dynamic dependence network models, developed in [16], take advantage of this decouple/recouple strategy and have since been implemented successfully in a variety of settings. These models capitalize on the potential for multivariate models to richly characterize the behavior of a network while maintaining the flexibility and scalability of modeling individual time series.

However, the observed activity in a network may belie the underlying processes that generate it which complicates the interpretation of the model. Consider, for instance, the flow of traffic onto a particularly crowded segment of a highway. This flow is constrained because only a fixed number of vehicles can occupy a road segment at any given time. This can potentially lead to erroneous conclusions about the activity in the network. For instance, a traditional interpretation of a lack of flow would suggest a lack of activity when, in fact, the network is experiencing an abundance of activity. Similarly, streaming services with limited capacity, either explicitly or practically due to bandwidth constraints, may

experience lower usage. Thus, usage data may not represent the underlying usage patterns. In each of these environments, and many others like them, individuals make policy decisions based solely on observed flow. Therefore, it is critical that policymakers be informed of not just the observed (censored) flow, but also the underlying flow of the network.

This thesis addresses this topic of characterizing overall network activity by accounting for the underlying flow when the observed flow is censored. To that end, this thesis is organized as follows: Chapter 2 introduces the decouple/recouple strategy for network flow data. Chapter 3 develops the theory for updating network flow when the destination is at capacity. Chapter 4 describes the network flow and censoring phenomenon in the context of publicly available city bikeshare data. Chapter 5 concludes with a commentary on the results and their practical implications.

Chapter 2

Modeling Dynamic Network Flow

There is a considerable amount of applied and theoretical work that characterizes the flow of elements or agents within a network of interconnected nodes. One particularly active area of research has been in website traffic flows [4].

A very straightforward paradigm for characterizing flow in a network at a node level is through a multinomial model where a J-dimensional probability vector $\boldsymbol{\theta}_t$ determines the flow of elements currently at node i at time t, n_{it}. This model assumes

$$\mathbf{Y}_{it} \sim MN(n_{it}, \boldsymbol{\theta}_t)$$

for a J-dimensional random vector $\mathbf{Y}_{it}$ where each element corresponds to the transition from node i to node $j \in J$. Note that this also accommodates "self flows" when an element at node i remains at node i.

The count of elements flowing between any pair of nodes can be summarized by the occupancy of the origin node and the probability of transitioning to the destination node. [4] and [3] model this count of elements flowing from an origin node i to a destination node j at time t individually. This non-negative integer value suggests the use of a Poisson distribution such that

$$Y_{ijt} \sim Poi(\phi_{ijt}). \tag{2.1}$$

Framing this problem as the composition of many individual flows invites a wide range of highly flexible modeling techniques from the univariate Dynamic Generalized Linear Model literature. Still, the question remains of how to contextualize these independently modeled, but inherently related flows. This notion of modeling many univariate time series independently and aggregating those results to reconstruct the originally sought-after multivariate distribution is referred to in the literature as *decouple-recouple* [14].

This chapter begins by providing the theory for Dynamic Generalized Linear Models

(DGLMs) in Section 2.1. Section 2.2 considers a special case of a mixture model which is of great relevance in modeling network flow. Section 2.3 describes how these univariate models can be combined to define the components of a rich and intuitive hierarchical model called the "Dynamic Gravity Model".

2.1 Decoupled Flow Using Dynamic Generalized Linear Model

This section reviews the basic fitting and update steps for a Dynamic Generalized Linear Model (DGLM) as described in [15] as well as [11]. Consider the standard univariate Dynamic Linear Model with the following observation and state evolution model:

$$\begin{aligned}
\text{(observation model) } Y_t &= \mathbf{F}_t'\boldsymbol{\theta}_t + \nu_t, \quad \nu_t \sim N\left(0, V_t\right), \\
\text{(state model) } \boldsymbol{\theta}_t &= \mathbf{G}_t\boldsymbol{\theta}_{t-1} + \mathbf{w}_t, \quad \mathbf{w}_t \sim N\left(\mathbf{0}, \mathbf{W}_t\right), \\
\text{(prior model) } \boldsymbol{\theta}_0 &\sim N\left(\mathbf{m_0}, \mathbf{W}_0\right).
\end{aligned}$$

Note that Y_t is a random variable distributed normally with a mean defined by dynamically changing regression coefficient vector $\boldsymbol{\theta}_t$, regressor vector $\mathbf{F}_t'$ and variance V_t.

We now consider the basic form of the generalized linear model in which Y_t follows an exponential family distribution with the following form:

$$p\left(Y_t|\eta_t, \phi\right) = \exp\left[\phi\left\{Y_t\eta_t - a\left(\eta_t\right)\right\}\right] b\left(Y_t, \phi\right) \tag{2.2}$$

where η is the natural parameter satisfying $E\left[Y_t|\eta_t, \phi\right] = \mu_t = \dot{a}\left(\eta_t\right)$ and $V\left[Y_t|\eta_t, \phi\right] = \ddot{a}\left(\eta_t\right)/\phi$.

To update the natural parameter η_t, we choose a conjugate prior for $\eta_t \mid \mathcal{D}_{t-1}$ where

$$\left(\eta_t \mid \mathcal{D}_{t-1}\right) \sim CP[\alpha_t, \beta_t]. \tag{2.3}$$

This conjugacy is highly desirable because it yields straightforward updates for the distribution of η_t,

$$\left(\eta_t|\mathcal{D}_t\right) \sim CP\left[a_t + \phi Y_t, \beta_t + \phi\right],$$

as well as a recognizable predictive distribution for $Y_t \mid \mathcal{D}_{t-1}$.

In the typical Dynamic Linear Model (DLM), Y_t follows a normal distribution with the appropriate values substituted into eqn. 2.2. In this case, the function g is just the identity function and the conjugate prior chosen for the natural parameter η_t is simply the normal distribution. Then, the state vector follows a normal distribution and the ordinary Kalman filter recursions follow naturally [9]. However, if the distribution of Y_t restricts the values which η_t may take, for instance, non-negative or bounded between $[0, 1]$, we should then consider using the state vector to model some transformed value of η_t.

Thus, analogous to a "static" GLM, we model

$$g(\eta_t) = \mathbf{F}_t' \boldsymbol{\theta}_t.$$

Defining the model in these terms presents some major challenges. Namely, we want to maintain the conjugacy in the distributions of η_t and Y_t, but assuming that the distribution of $\boldsymbol{\theta}_t$ is normal (as in the DLM) imposes significant restrictions on the form of the prior for η_t. Therefore, analysis is based on summaries of information about $\boldsymbol{\theta}_t$ in terms of its first- and second-moments where

$$(\boldsymbol{\theta}_t \mid \mathcal{D}_{t-1}) \sim [\mathbf{a}_t, \mathbf{R}_t].$$

The challenge is then connecting the transitions state vector $\boldsymbol{\theta}_t$ with the updates to our beliefs about η_t. The steps to do this are outlined below.

1. Provide initial values for $\mathbf{a}_0$ and $\mathbf{R}_0$.

2. Specify the moments of $\boldsymbol{\theta}_t$ and find the parameters of η_t where their distributions are defined as

$$\boldsymbol{\theta}_t \mid \mathcal{D}_{t-1} \sim [\mathbf{a}_t, \mathbf{R}_t],$$
$$\eta_t \mid \mathcal{D}_{t-1} \sim CP[\alpha_t, \beta_t].$$

   Solve α_t and β_t are solved for such that

$$E(g(\eta_t) \mid \mathcal{D}_{t-1}) = \mathbf{F}'\mathbf{a}_t,$$
$$V(g(\eta_t) \mid \mathcal{D}_{t-1}) = \mathbf{F}'\mathbf{R}_t\mathbf{F}.$$

3. Update η_t where

$$\eta_t \mid \mathcal{D}_t \sim CP[\alpha_t + \phi Y_t, \beta_t + \phi].$$

Then, define $g_t = E(\eta_t \mid \mathcal{D}_t)$ and $p_t = V(\eta_t \mid \mathcal{D}_t)$

4. Update $\boldsymbol{\theta}_t$ where

$$\boldsymbol{\theta}_t \mid \mathcal{D}_t \sim [\mathbf{m}_t, \mathbf{W}_t].$$

Using g_t and p_t from the previous step, the first and second moments are defined as

$$\mathbf{m}_t = \mathbf{a}_t + \mathbf{R}_t\mathbf{F}_t(g_t - \mathbf{F}'\mathbf{a}_t)/\mathbf{F}_\mathbf{t}'\mathbf{R}_t\mathbf{F}_t,$$
$$\mathbf{W}_t = \mathbf{R}_t - \mathbf{R}_t\mathbf{F}_t\mathbf{F}_t'\mathbf{R}_t'(1 - p_t/\mathbf{F}_\mathbf{t}'\mathbf{R}_t\mathbf{F}_t)/\mathbf{F}_\mathbf{t}'\mathbf{R}_t\mathbf{F}_t.$$

5. Perform the State Transition by incrementing t and defining the new first and second moments as

$$\mathbf{a}_t = \mathbf{G}_t\mathbf{m}_{t-1},$$
$$\mathbf{R}_t = \mathbf{G}_t\mathbf{W}_{t-1}\mathbf{G}_t'/\delta_t.$$

$\delta_t \in (0, 1]$ is a hyperparameter chosen by the modeler.

6. Repeat these steps 2-5 until $t = T$.

Specific details about the derivation of these equations can be found in [15].

This process is referred to as *sequential updating* which provides parameter estimates based on the data up to the current time point t. Depending on the application, it may also be useful to consider *retrospective updating* which can provide parameter estimates at a time point t based on all data before and after it in the time series. This process is outlined in Appendix 3 of [4].

2.2 Dynamic Count Mixture Model

There has been some theoretical work done to address case-specific nuances that make the application of standard count models (e.g. a Poisson distribution) unrealistic representations of the data generating process. One common challenge is characterizing intermittent counts, or circumstances in which there are many observations with a value of 0. Such a preponderance of zeros leads to likelihood functions that favor small Poisson means. In a Poisson/gamma conjugate context, such as that of a DGLM, this means that posteriors for the Poisson mean will increasingly concentrate on smaller values as data are sequentially processed.

Some models accommodate this intermittent count by modeling both the size of the count and the occurrence of a non-zero count [6]. [2] describes an approach called a Dynamic Count Mixture Model (DCMM) to accommodate many 0-counts along with intermittent bursts of activity using DGLMs. A DCMM essentially builds two DGLMs: a binary DGLM which models whether the count is nonzero and a Poisson DGLM models the counts greater than 0. The model defines the count y_t where

$$z_t \sim \text{Ber}(\pi_t), \tag{2.4}$$

$$y_t \mid z_t = \begin{cases} 0, & \text{if } z_t = 0, \\ 1 + \text{x}_t,\ x_t \sim \text{Po}(\mu_t), & \text{if } z_t = 1. \end{cases} \tag{2.5}$$

Using the format and notation of the DGLMs in the previous section, the natural parameters for the distributions of the random variables are

$$\eta_t^0 = \text{logit}(\pi_t) = \mathbf{F}_t^{0\prime}\boldsymbol{\theta}_t^0,$$
$$\eta_t^+ = \log(\mu_t) = \mathbf{F}_t^{+\prime}\boldsymbol{\theta}_t^+,$$

with the superscript 0 denoting the parameters for the Bernoulli random variable and the superscript + denoting the parameters for the Poisson. The conjugate priors for the transformed values of the natural parameters are the Beta and Gamma distributions respectively.

2.3 Recouple Using Dynamic Gravity Model

With this formulation of univariate DGLMs, there is an opportunity for "recoupling" with multivariate models that more richly characterize the behavior of the network while maintaining the flexibility (and scaleability) of modeling individual ("decoupled") time series. To that end, Dynamic Gravity Models (DGMs) have the capacity to that allow the mean parameter of the Poisson distribution to be decomposed into 4 constituent parts such that

$$\phi_{ijt} = \mu_t \alpha_{it} \beta_{jt} \gamma_{ijt}.$$

This model formulation, used in [4] and [3], is highly similar to the model used in [13] but allows the parameters to change over time. The parameters of this model can be interpreted as follows: μ_t characterizes the baseline activity at time t, α_{it} characterizes the origin effect, β_{jt} characterizes the destination effect, and γ_{ijt} captures the interaction between the particular origin and destination. Models of similar forms have been used in transportation studies (e.g. [13]) where the interaction term is typically a function of physical distance between nodes.

There is a straightforward, bijective mapping of the Poisson parameters to the DGM components. As in [4], consider the zero-sum constraint on the logged values. That is, for any given time point t, $\sum_i \log(\alpha_{it}) = \sum_j \log(\beta_{jt}) = \sum_j \log(\gamma_{ijt}) = \sum_i \log(\gamma_{ijt}) = 0$. For simplicity, we denote $\log(\phi_{ijt}) = f_{ijt}$.

1. Calculate the baseline level of activity

$$h_t = \log(\mu_t) = \left(\frac{\sum_i \sum_j f_{ijt}}{I^2} \right)$$

2. Calculate the origin effect

$$a_{it} = \log(\alpha_{it}) = \left(\frac{\sum_j f_{ijt}}{I} \right) - h_t$$

3. Calculate the destination effect

$$b_{jt} = \log(\beta_{jt}) = \left(\frac{\sum_i f_{ijt}}{I} \right) - h_t$$

4. Calculate the origin-destination interaction

$$g_{ijt} = \log(\gamma_{ijt}) = f_{ijt} - h_t - a_{it} - b_{jt}$$

When using the DCMM of the previous section, instead of using the Poisson parameter, it is also possible to use the predictions $\hat{y}_{ijt}$ in its place to calculate the components of the dynamic gravity model as done in [3].

Chapter 3

Posterior Updates with Censored Counts

Data censoring occurs when the value of an observation is restricted from taking certain values, leaving the "true" value of an observation unknown. Generally, censoring is categorized as right, left, or interval censored when the true observation is greater than some value, less than some value, or outside an interval respectively. It is often of great interest in applied problems to gain insight into these unknown values to better understand the data generating process.

For instance, suppose there is some set of independent observations $(y_1, y_2, \ldots, y_n)$ (according to some function $f(\theta)$) that is right-censored by some known value C. If the goal is to find the parameter θ that maximizes the probability of the observed data $(p(y_1, \ldots, y_n \mid \theta))$, this involves writing the likelihood for both censored and uncensored data and maximizing it with respect to θ. This can be expressed as

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \prod_{i=1}^{n} d_i f(y \mid \theta) \times (1 - d_i)(1 - F(C \mid \theta)),$$

where F refers to the CDF, and the indicator $d_i = 1$ if $y_i < C$ and $d_i = 0$ otherwise. [7] reviews a well-known model that uses this approach called the "Tobit" regression and extends it to a Gaussian process model.

Instead, of maximizing $p(y_1, \ldots, y_n \mid \theta)$, this thesis takes a Bayesian approach which allows for probability statements to be made directly about θ via its posterior distribution $p(\theta \mid y_1, \ldots, y_n)$. In situations in which the prior $p(\theta)$ and likelihood $p(y \mid \theta)$ are conjugate pairs, the posterior distribution has the same form as the prior which facilitates iterative updates to prior beliefs as new data are observed.

This chapter discusses the nuances of updating the posterior distribution when the form of the data likelihood *does not* allow for straightforward posterior updates in the case of right-censored count data. Section 3.1 discusses the form of the posterior distribution when

an observation has been right-censored and Section 3.2 describes a method for estimating this distribution that facilitates straightforward posterior updates.

3.1 Posterior with Right-Censored Count Data

An advantage of using a Bayesian approach is its ability to easily perform iterative updates of prior beliefs with each subsequent observation when the prior and likelihood are conjugate pairs. In the case of count data, the distribution of the data x_t is represented by a Poisson distribution (with parameter λ_t) and the distribution of λ_t is assumed to be Gamma which allows for sequential updates of the distribution of λ_t of the following form:

$$\lambda_t|\mathcal{D}_t \sim \text{Gamma}(\alpha_t^*, \beta_t^*),$$

where the parameters are calculated using the data at time t such that

$$\alpha_t^* = \alpha_t + x_t,$$

$$\beta_t^* = \beta_t + 1.$$

Note that when x_t (the observed count at time t) is censored, α^* may be underestimated which means that the posterior mean of λ_t (α^*/β^*) would also be artificially low. Therefore, when x_t is censored, the posterior density must be specified as

$$\begin{aligned} p(\lambda_t|y \geq x_t) &\propto p(\lambda_t \mid \alpha_t, \beta_t) p(y \geq x_t \mid \lambda_t) \\ &\propto p(\lambda_t \mid \alpha_t, \beta_t) \times \left(\sum_{k=x_t}^{\infty} p(y = k \mid \lambda_t) \right). \end{aligned} \tag{3.1}$$

Leveraging conjugacy, this density can be rewritten as an infinite mixture of gamma distributions such that

$$\begin{aligned} p(\lambda_t|y \geq x_t) &\propto \sum_{y=x_t}^{\infty} \frac{\beta_t^{\alpha_t}\Gamma(\alpha_t + y)}{\Gamma(\alpha_t)(\beta_t + 1)^{\alpha_t + y} y!} Ga(\lambda_t \mid \alpha_t + y, \beta_t + 1) \\ &\propto \sum_{y=x_t}^{\infty} p_y Ga(\lambda_t \mid \alpha_t + y, \beta_t + 1). \end{aligned} \tag{3.2}$$

The weight on each mixture component can then be defined as

$$p_y = c \cdot \frac{\Gamma(\alpha_t + y)}{(\beta_t + 1)^{\alpha_t + y} y!},$$

where c is a constant that forces $\sum_{y=k}^{\infty} p_y = 1$. In practice, the weights for each mixture component, and by extension c, can be approximated by choosing a sufficient number of components.

In practice, it is therefore useful to consider the appropriate number of components. Figure 3.1 provides an example of some exploratory work that may be done to provide insight into this problem. Setting $\alpha = 3$ and $\beta = \alpha/m$, this plot explores the size of normalized weights of the mixture components p_y for different values of the prior mean m and the maximum value which can be observed k. The horizontal blue line shows an arbitrary threshold of $\log(10^{-3})$. The modeler must make the determination of an acceptable threshold such that once a component weight falls below that threshold, the remaining components are disregarded.

This example highlights the challenge of approximating the mixture components under some circumstances. For smaller values of k relative to the value m (the solid lines), the relative weights of the mixture components decrease quickly. However, the weights for larger values of k decay much more slowly and would thus require more mixture components to create a reasonable approximation.

With the mixture components calculated, it is then possible to calculate the density of the infinite mixture of gammas. Figure 3.2 shows the prior gamma density ($p(\lambda_t)$) for the specified α and β parameters along with its associated posterior densities ($p(\lambda_t \mid y > k)$) for different values of k. We see that these posterior distributions appear to resemble a gamma distribution. Therefore, it stands to reason that this infinite mixture can be reasonably approximated by a single gamma distribution to more easily accommodate the posterior updates as in the uncensored case. To that end, we attempt to find the parameters of a gamma distribution that best approximate this mixture.

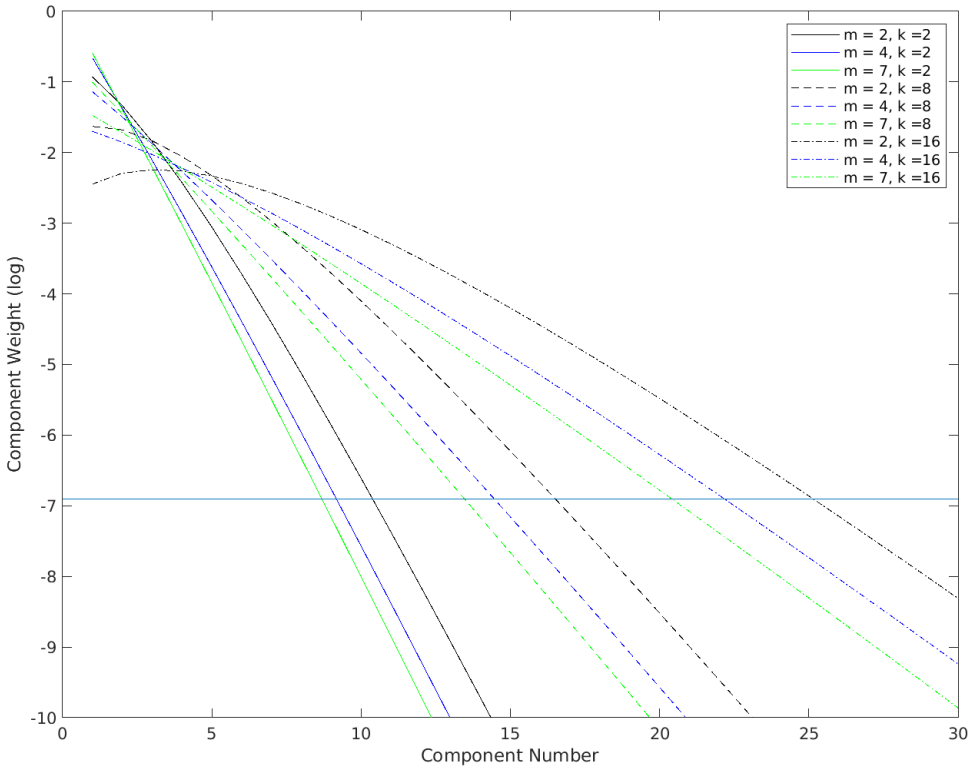


Figure 3.1: Log weights of mixture components for different means and values of k. The horizontal blue line shows the value of $\log(10^{-3})$.

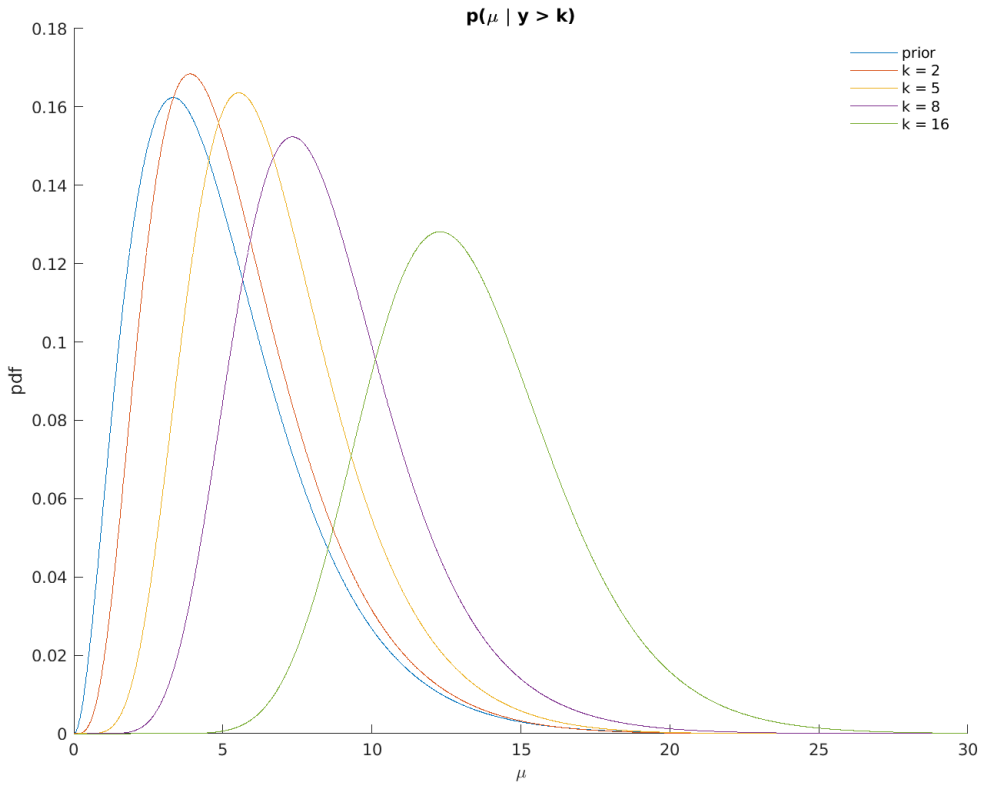


Figure 3.2: Examples of densities of infinite mixtures of gamma distributions with parameters $\alpha = 3$, $m = 5$, $\beta = \alpha/m$

3.2 Approximating the Infinite Mixture of Gammas

Kullback-Leibler (KL) divergence is defined below using $p(y)$ which refers to the complicated infinite mixture of gamma distributions and $g(y)$ which refers to the single gamma distribution whose parameters we wish to set to approximate $p(y)$. The divergence of $g(y)$ from $p(y)$ is

$$\begin{aligned} K_{g|p} &= \int_0^\infty \log\left(\frac{p(y)}{g(y)}\right) p(y)dy \\ &= \int_0^\infty \log(p(y))p(y)dy - \int_0^\infty \log(g(y))p(y)dy \\ &= -H_p + S_{g|p}, \end{aligned}$$

where H_p and S_g represent the integrals above. The objective is to find the parameters of $g(y)$ which minimize this expression. Since H_p does not depend on $g(y)$, we focus instead on $S_{g|p} = -\int_0^\infty \log(g(y))p(y)dy$ which can be simplified to

$$\begin{aligned} S_{g|p} &= -\int_0^\infty \log(g(y))p(y)dy \\ &= E_{p(y)}(-\alpha \log \beta + \log \Gamma(\alpha) + (1-\alpha)\log y + \beta y) \\ &= -\alpha \log \beta + \log \Gamma(\alpha) + (1-\alpha)h + \beta m \end{aligned}$$

where $m = E_{p(y)}(y)$ and $h = E_{p(y)}(\log y)$. These quantities can be approximated using samples from the mixture of gammas. We now find the minimum by differentiating with respect to α and β such that

$$\begin{aligned} \frac{\partial S_{g|p}}{\partial \alpha} &= -\log \beta + \psi(\alpha) - h, \\ \frac{\partial S_{g|p}}{\partial \beta} &= \frac{\alpha}{\beta} - m. \end{aligned}$$

Equating these to 0 and noting that $\beta = \alpha/m$, we can first solve for α using Newton-Raphson method and plug in this value to find β. Finding the determinant of the second derivative matrix shows that this is the minimum as long as $\alpha > 0$.

To illustrate this method's efficacy, Figure 3.3 summarizes the results for the posterior densities $p(\lambda_t \mid y > k)$ for $k \in (2, 8, 16)$, $\alpha = 3$, and $m = 5$. The histograms show the distribution of 10,000 samples from each of the (infinite) mixture of gammas which are used to estimate $E_{p(y)}(y)$ and $E_{p(y)}(\log y)$. The solid black line shows the true density of this function $p(y)$ from the infinite mixture of gammas and the dashed blue line shows the density of the single fitted gamma distribution $g(y)$. In each case, the blue line is virtually indistinguishable from the black suggesting that this single gamma distribution sufficiently approximates the mixture of gamma distributions.

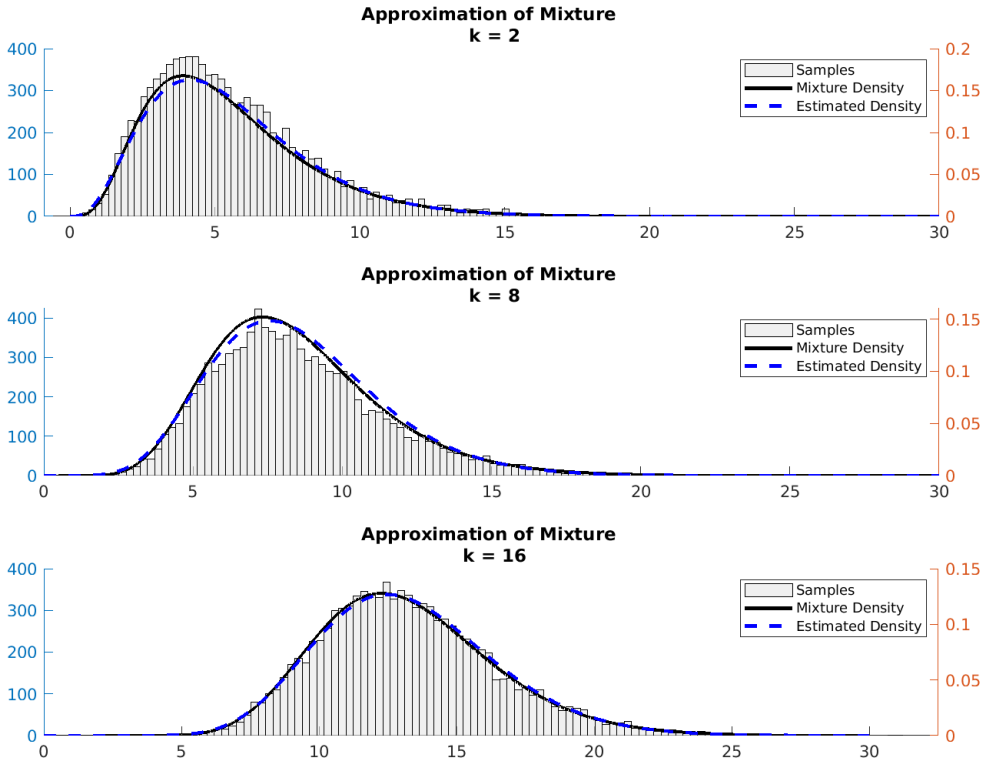


Figure 3.3: Histograms show samples from the infinite mixture of gamma distributions with parameters $\alpha = 3$, $m = 5$, $\beta = \alpha/m$. Solid black lines show the the density of this infinite mixture of gammas. Blue lines show the density of a single gamma distribution with parameters calculated to minimize KL divergence. The left axis shows the counts for each bar of the histogram and the right axis shows the values of the density functions.

Chapter 4

Bike Sharing Network Flow

The City of Chicago operates a bike sharing service with 612 stations where customers can interact with a kiosk to check-out a bike that can be returned to any station in the network (including the one from which it originated). These bikes are often used for recreational activity, but can also serve as the primary mode of transportation to work for some (when the weather is amenable).

However, periods of high activity such as morning and evening rush hour, as well as late afternoon recreational activities, can strain the capacity of the stations leaving some completely empty and others completely full. The case in which stations are filled, which this application focuses on, can be much more challenging to customers as they may not be able to return the bike at their intended destination and must find an alternative destination. This can be problematic for customers for two reasons. First, this service charges by time the bike has been in use meaning that, until the bike has been returned, the customer is charged for time spent finding an available station. Related to this, a customer may be reluctant to use this eco-friendly mode of transportation if their trip could be extended by trying to find an available station.

To address this, the City of Chicago redistributes bikes throughout the network to accommodate stations that experience sudden shortages in the number of available bikes or docks. With this initiative and other demand planning strategies, it is therefore useful to consider the underlying dynamic demand in order to redistribute bikes most efficiently. Moreover, understanding the underlying activity in the network is crucial when planning expansion of services.

To address these objectives, this chapter is organized into several sections. Section 4.1 introduces the dataset used for this activity. Section 4.2 provides insights into the censoring phenomenon in this dataset. Section 4.3 outlines the application of the methodology devel-

oped in the previous two chapters and Section 4.4 reviews the results for these data. The chapter concludes in Section 4.5 with an overview of extensions to generalize this model and more richly characterize the underlying sources of demand.

4.1 Data Overview

These bike stations are straightforwardly represented as a network where each station represents the nodes which are "connected" via trips between the stations. Each of the stations contain only a fixed number of docks where a bike may be returned.

Data describing the bike sharing activity in this network comes in two forms: bike-level trip data and station-level occupancy data.[1] The trip data are comprised of the origin and destination stations as well as the start and end time for each trip. The occupancy data contains, for each station, the count of available docks which are refreshed every 10 minutes on the 5-minute mark (XX:05, XX:15, XX:25, etc.).

To narrow the focus of this analysis, the data are limited to a single 24-hour period exhibiting "typical" weekday behavior in terms of weather and special events. The day of Wednesday, June 26, 2019 fit this criterion, so it follows that the results obtained here could be reasonably generalized to other weekdays with similar conditions.

4.1.1 Station Data

In order to reduce the sparsity of these 612 stations and to manage the computational burden of running the model, the data are constrained to the top most active nodes. "Active nodes" are stations meeting at least one of the following criteria:

1. One of the top 50 highest net change stations ($|\ \text{occupancy}_{06:30} - \text{occupancy}_{10:00}\ |$) in the *morning* rush hour period (between 06:30 and 10:00)

[1] These data are available for download from https://www.divvybikes.com/system-data.

2. One of the top 50 highest net change stations in the *evening* rush hour period (between 16:00 and 18:30)

3. One of the top 50 highest traffic stations (inflows + outflows) in the *morning* rush hour period (between 16:00 and 18:30)

4. One of the top 50 highest traffic stations (inflows + outflows) in the *evening* rush hour period (between 16:00 and 18:30)

Applying this inclusion criteria yields 153 unique stations. All other stations considered insufficiently active are aggregated into an "Other" station category.

It is worth noting that these data, in contrast to other network flow models (e.g. [3], [4]), there is no designation of an "external node". That is, a node which individuals may use to enter and exit the network. In other words, this is a closed network, so elements flowing in the network must stay in the network unless the operators choose to add or remove elements.

4.1.2 Trip Data

To combine the station and trip information, the trip time indices are formatted to correspond with the time intervals created by the 10-minute refresh cycles of the docking stations. The trips are defined using the time the bike arrived at the destination station. The reason for this is to improve the interpretability of flows within the network. In previous work, individuals that stay at a particular node are counted as flows with the same origin and destination node ("self flow"). However, that designation in this context is a bit more complex.

Trips often start and end at the same node; recreational riders may be interested in exploring the vicinity of the station and then returning. Defining a self flow for bikes that remain at the same station obscures this meaning. Therefore, for the purposes of reducing sparsity and improving interpretability, only the trip's ending time point is used to determine flows.

4.2 Data Exploration

The daily activity of this network aligns closely with what one might expect. Figure 4.1 shows the number of trips taken per 10-minute interval as well as the number of stations throughout the network that are full. There are two intervals of high activity that one would expect to correspond to the morning and evening rush hour. Additionally, there is some activity sustained throughout the day suggesting that the evening rush hour may be composed of both commuter traffic and recreational activity. Following these periods of high activity, one also notes the increase in the number of stations that are filled to capacity. It is during this period that we would expect the most observations to be censored.

There is also an opportunity to see this phenomenon at a station level. Figure 4.2 shows the number of available docks for incoming bikes at 4 stations which presumably see a high volume of commuter activity. The slopes of these plots provide insight into the net change in number of bikes available at each station, with higher numbers indicating emptier stations and lower numbers indicating fuller stations. The blue and red line (stations 36 and 49) suggest a high volume of incoming bikes relative to the number of outgoing bikes. During the morning high-traffic period, these stations are filled to capacity (count of available docks shrinks to 0) and in the evening we see approximately an equal and opposite change. The green and purple line (91 and 174) show just the opposite effect. In the morning, there are very few available docks though the number grows rapidly as the morning progresses. These two groups show two examples of morning origin-evening destination nodes and two examples showing the opposite effect. One important nuance of these data is that, though there are apparent high-level changes, the number of trips between any two stations tends to be very low with few exceeding 3 within a given 10-minute interval. Therefore, while this is in theory a fully connected network, there are relatively few station pairs between which flows occur. In the network with the 153 stations (nodes), there are $2,038$ edges between them. This is far less than the 153^2 one would expect in a more dense network.

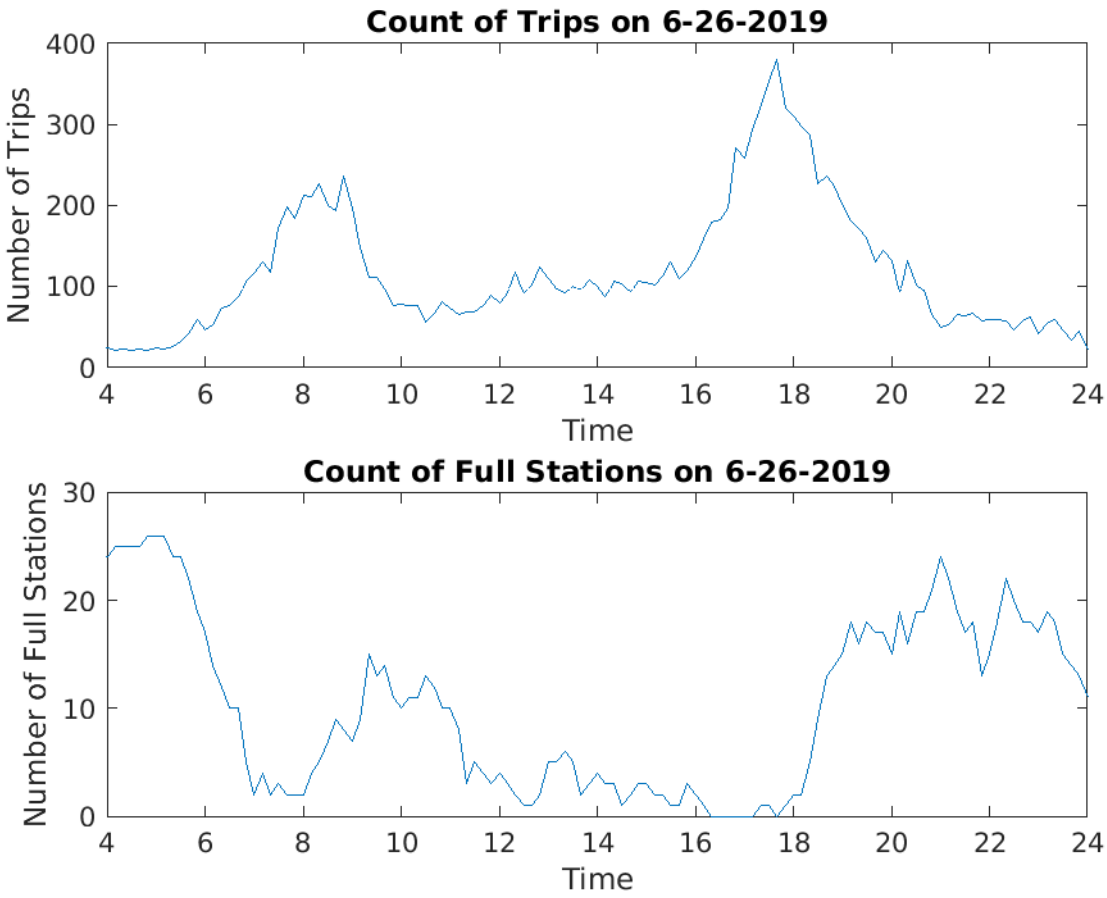


Figure 4.1: The plot on top shows the number of trips taken in each 10-minute interval during the day of 6-26. The plot below shows the count of stations without any available docks (i.e. full)

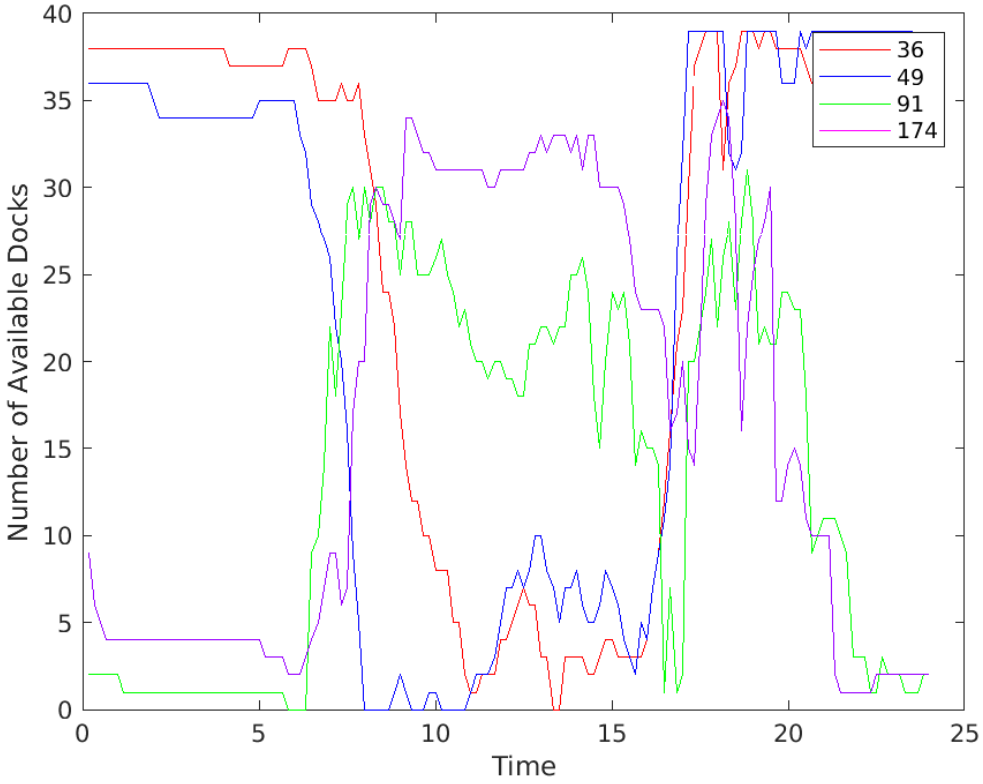


Figure 4.2: Count of available docks for 4 bike sharing stations during June 26, 2019

4.3 Model Fitting

The modeling objective is to characterize the flow of bikes throughout the network of stations, even in circumstances in which stations are filled. To that end, this model leverages the methodology described in chapters 2 and 3 to develop a dynamic count mixture model that has the capacity to update the parameters of the Poisson parameter when the counts are censored. This model is referred to below as the "Censored" model. For contrast, there is another model that assumes no censoring of the data which is referred to as the "Uncensored" Model.

4.3.1 DCMM Characteristics

For the purposes of this example, a second-order polynomial DGLM (see [15], Ch 7 and [11] Ch 4 for more details) was fitted for both the Bernoulli and Poisson DGLMs such that,

$$\mathbf{F}_t = \mathbf{F} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\mathbf{G}_t = \mathbf{G} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

It is worth noting that this is a relatively simple model where the state vector $\boldsymbol{\theta}_t$ has only two elements: θ_t^1 to represent the expected value of the series and the other θ_t^2 representing the change from that level. This can be straightforwardly deduced from the state update equation

$$\begin{aligned} \boldsymbol{\theta}_t = \begin{bmatrix} \theta_t^1 \\ \theta_t^2 \end{bmatrix} &= \mathbf{G} \begin{bmatrix} \theta_{t-1}^1 \\ \theta_{t-1}^2 \end{bmatrix} + \mathbf{w}_t \\ &= \begin{bmatrix} \theta_{t-1}^1 + \theta_{t-1}^2 \\ \theta_{t-1}^2 \end{bmatrix} + \mathbf{w}_t, \end{aligned}$$

where $\mathbf{w}_t$ is the two-dimensional innovation vector. Increasing the dimension of $\mathbf{F}$ and $\mathbf{G}$

can represent higher order polynomials. Other extensions such as including random effects or exogenous variables are discussed as possible extensions.

Referencing the state update equations in Section 2.1, it is also worth considering the relevant values for the discount parameter δ where, in practice, $\delta \geq .9$ [11]. A relatively high value of .95 was used for both the Poisson and Bernoulli DGLM to encourage stability of the state innovations.

4.3.2 Parameter Updates

As noted in Chapter 2, the Bernoulli parameters are updated with every new datapoint and the Poisson parameters are updated only when the observed flow at time t, $x_t > 0$. However, these assumptions change slightly when the data are censored. The table below summarizes the appropriate circumstances under which to consider an observation missing and to not update parameters, and when to assume an observation has been censored which would lead to a censored update for the Poisson distribution. To determine censoring of the observed flows into a station, we look at the number of available docks at the next time point C_{t+1}.

$x_t > 0$	$C_{t+1} = 0$	Skip Bernoulli	Skip Poisson	Censored Poisson Update
F	T	✓	✓	
F	F		✓	
T	T	✓		✓

We conclude that if the count at time $t + 1$ is 0, the Poisson update should be skipped and, if we find that in the following time point the destination is at capacity, the Bernoulli update should be skipped. These update rules are somewhat restrictive for this dataset due to the intermittent nature of flows between any two stations.

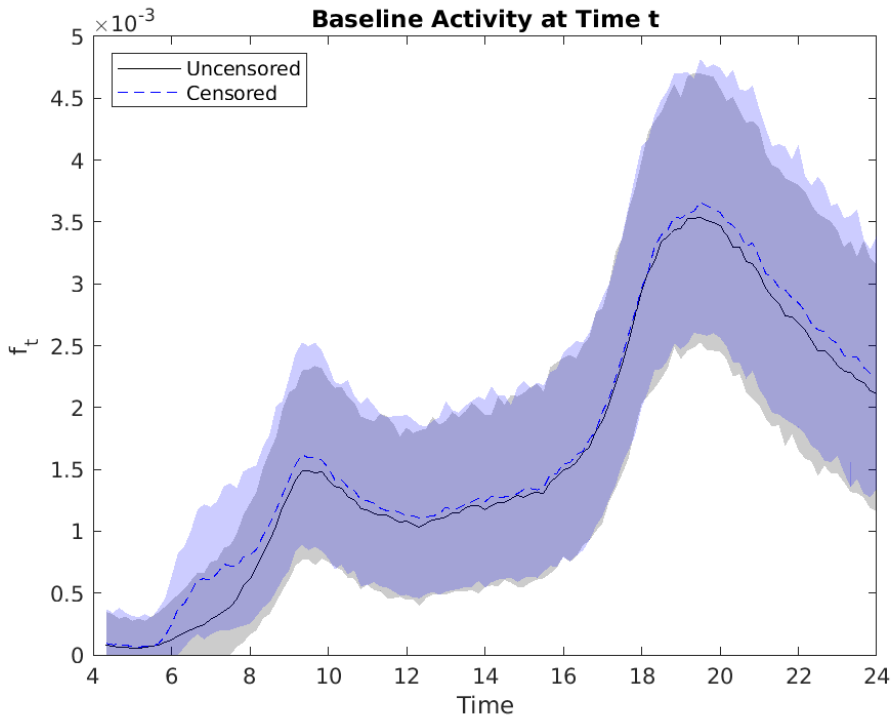


Figure 4.3: Illustration of baseline activity f_t changes with censored Poisson updates and non-censored Poisson updates with $\delta = .95$

4.4 Results

Recall that the objective is to characterize the flow at a given time point t in the day. Therefore, as noted in Chapter 2, the insights should be based on the parameters calculated during the retrospective analysis.

We first consider the estimate of baseline activity $f_t \mid \mathcal{D}_{\mathcal{T}}$. Figure 4.3 shows the baseline activity estimates which, as one might expect, resemble the daily trip data. The dotted line shows the estimated baseline activity when censoring is assumed and the solid black line shows the model that assumes no censoring. The 95% credible intervals suggest that there is a high level of variability around the estimates, though there still appear to be some meaningful and intuitive trends.

First, we see that in the early morning, the censored model shows increased activity relative to the uncensored model. Additionally, though subtle, there is a distinct increase in

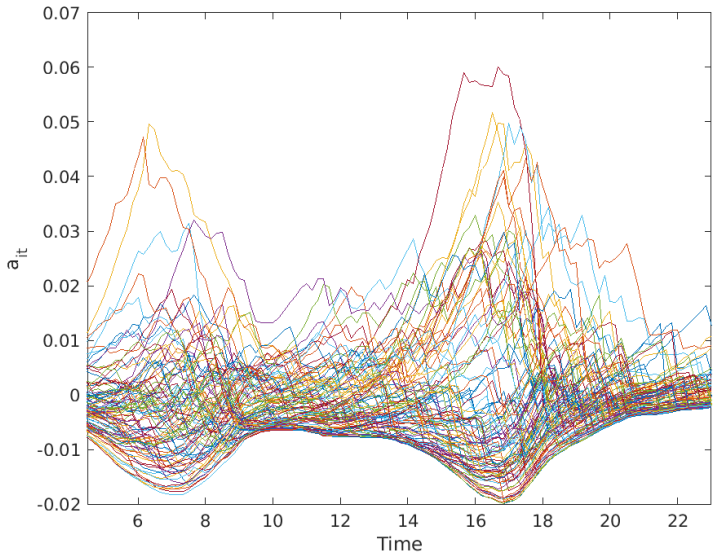
baseline activity levels at peak morning activity (approximately 09:00). Another difference is the sustained activity through the evening. This particular increase is interesting because it is not during traditional weekday rush hour periods. It should be noted that in both of these times, there tend to be a fair number of stations that are filled to capacity. However, since this is simply baseline activity, it may be more useful to look at node-specific behavior through their origin effects $a_{it} \mid \mathcal{D}_{\mathcal{T}}$ and $b_{jt} \mid \mathcal{D}_{\mathcal{T}}$.

We can see the estimates of $a_{it} \mid \mathcal{D}_{\mathcal{T}}$ in Figure 4.4a and $b_{jt} \mid \mathcal{D}_{\mathcal{T}}$ in Figure 4.4b. As observed in the net change in available docks of in Figure 4.2, there appear to be stations with strong "morning-origin" and others with strong "evening-destination" effects. It is worth examining individual stations more closely to view these time-of-day effects. That is, the tendency for a node to function primarily as an origin or destination at different times of the day.

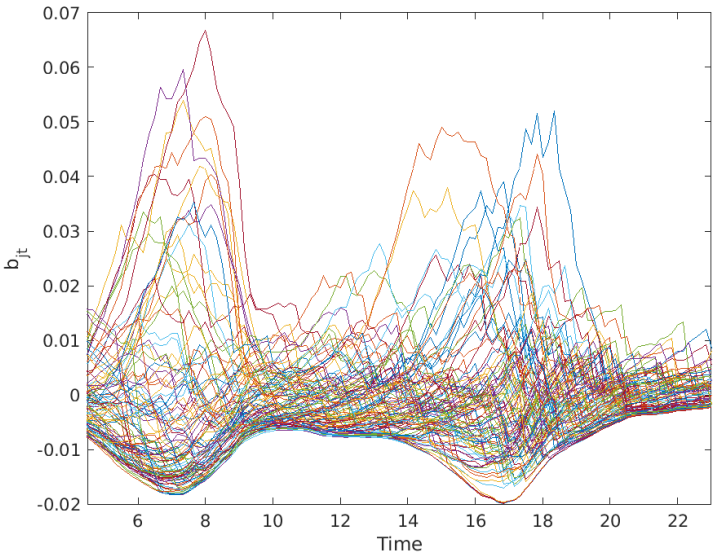
There are very few stations with sustained activity throughout the day. Therefore, as a starting point, it is helpful to consider the "Other" station described above which is just the combination of stations with insufficient activity. Because it is comprised of multiple stations, its activity is not as intermittent as individual stations. This shows a similar trend to the overall baseline activity.

Figure 4.6 shows some destination effects for some selected nodes that fill to capacity at some point during the morning rush hour. As before, the dashed lines represent the estimates for b_{jt} for the model that assumes censoring and the solid lines show the estimates for the model that does not. For the most part, the destination effects look fairly similar at first.

However, we notice that for several stations (40, 49, and 67), the uncensored model assumes that the already falling destination effect becomes negative which suggests it is an unfavorable destination. The censored model, though initially decreasing, abruptly stops its descent and maintains a moderate non-negative destination effect. The effect for the uncensored model appears to recover eventually after nearly two hours. These trends are suggestive of the censored model compensating for the sudden lack of capacity by continuing



(a) Plot of a_{it} for all stations on June 26, 2019



(b) Plot of b_{jt} for all stations on June 26, 2019

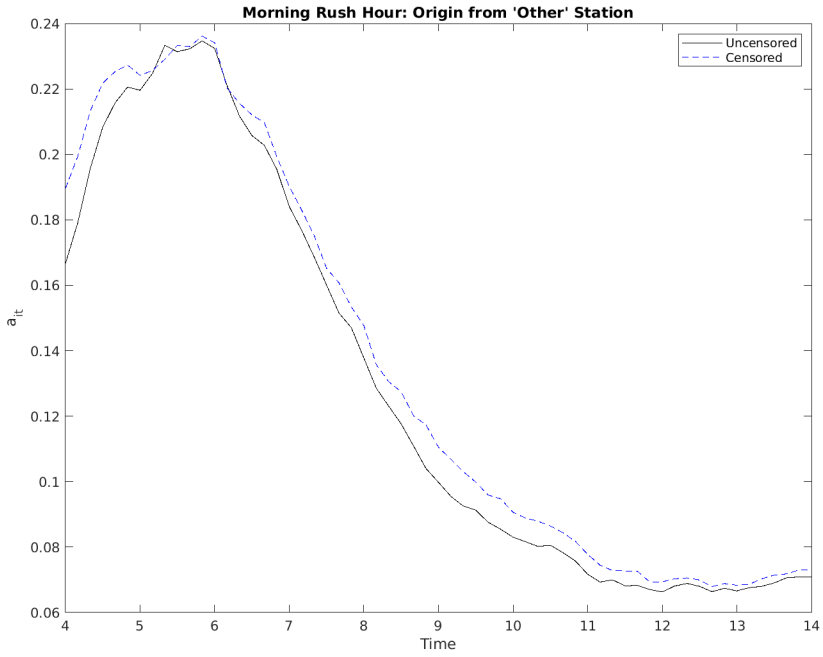


Figure 4.5: Origin effects for "Other" station

the trend of flows into the station. Note the timing that the number of available docks at station 49' falls to 0 in Figure 4.2.

To examine this trend further, we look at the predicted flows into station 49 during the morning hours in Figure 4.7. The blue line represents the true sum of flow into station 49 ($\sum_i y_{ijt}, j = 49$) and the red and yellow lines correspond to the predictions of the uncensored and censored model respectively.

4.5 Discussion and Extensions

This application showcases the major differences between a model that assumes censoring and another that does not. We first contrast the overall baseline activity effects f_t of the two models where the most notable differences occur in the early morning hours and in the times of peak activity. The morning activity stands out because it occurs when, based on Figure 4.1, the number of full stations is relatively low even though the number of trips is

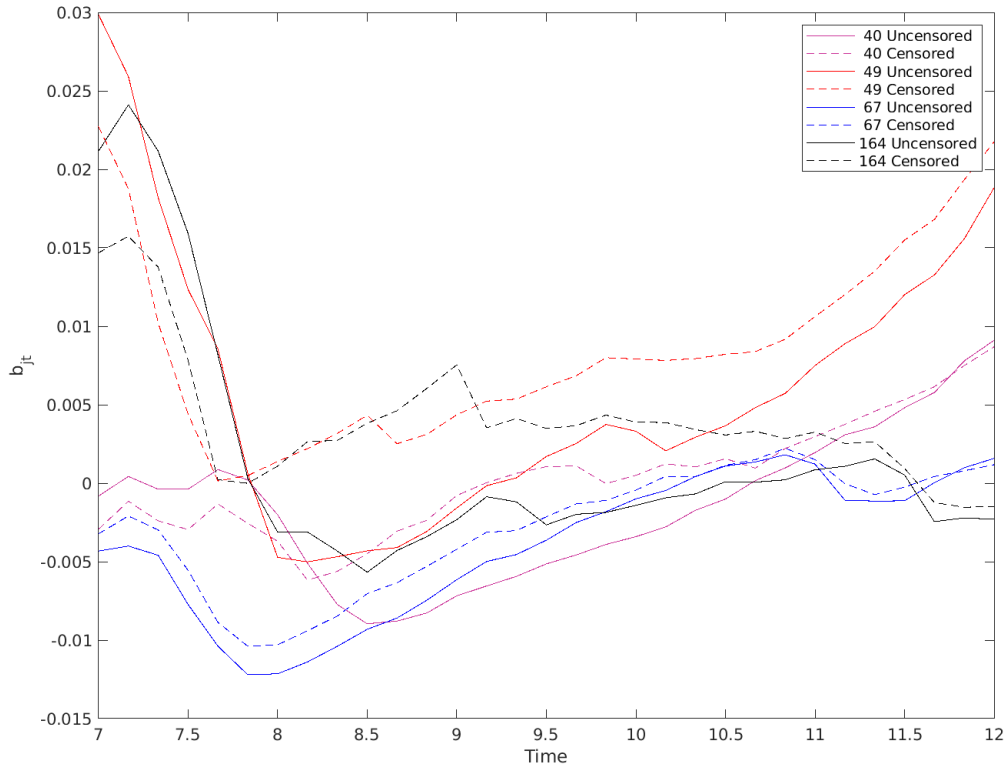


Figure 4.6: Lines show the estimated parameter b_{jt} during the hours of 07:00-10:00 for 5 stations. Dashed lines show parameters estimates using censored updates and solid lines show estimates when all data are assumed to be uncensored.

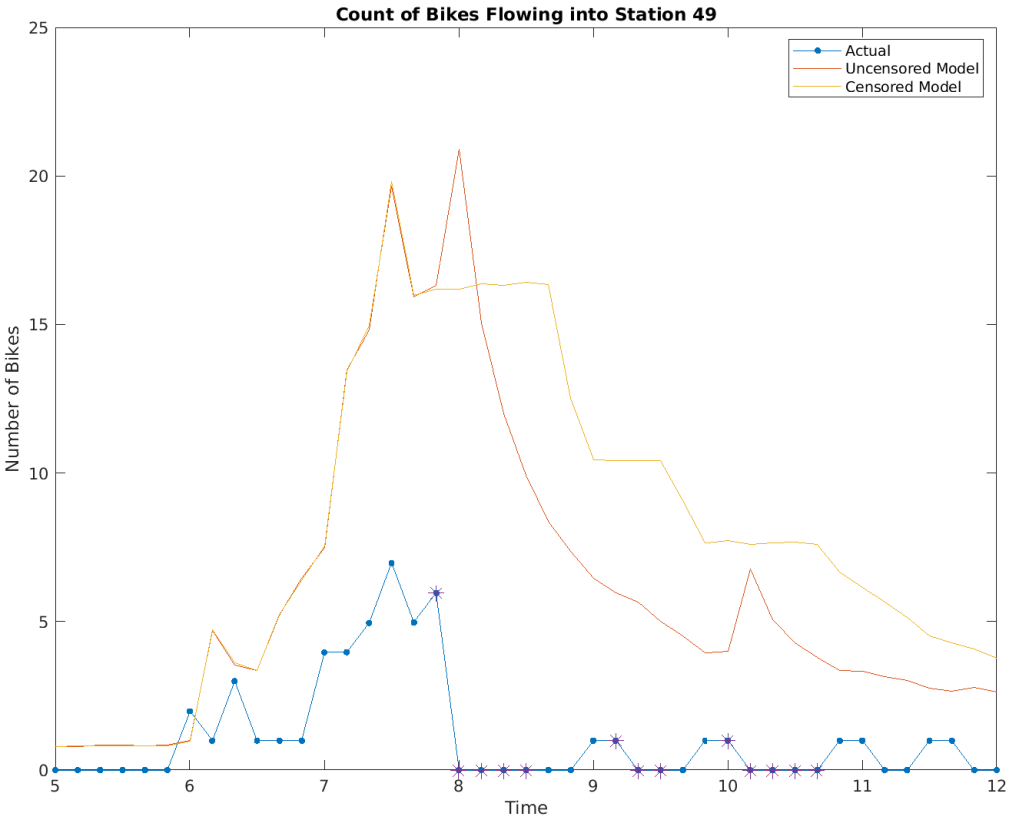


Figure 4.7: The blue line shows the true number of arrivals at station 49. The red and yellow lines show the predicted flow into station 49 under the "Uncensored" and "Censored" models respectively. The red asterisks show when the number of arrivals was censored.

relatively high. An explanation for this can be supported, in part, by examining the plot of available docks for station 91 in Figure 4.2. This particular station begins with very few available docks which is most likely in preparation for the large number of bikes that will be removed during the morning rush hour. Still, note that station 91 serves as the destination for several trips in the early morning hours causing the number of available docks to fall to 0 and the corresponding flows to be censored. This is, in fact, a relatively common trend among stations with strong morning-origin effects; 28 other stations showed a similar censoring pattern on this day which could explain the increase in activity in the early morning hours under the censored model.

The other major difference, the increased activity during times of peak activity, is an intuitive trend. As stations begin filling during times of high activity, the censored model assumes that this high activity continues which results in higher baseline activity. This can be seen in Figure 4.6 where stations that exhibit a strong destination effect in the morning continue on this trajectory in the censored model while the uncensored model shows it becoming a less desirable destination.

There is another interesting behavior exhibited by both models leading up to the censoring at these selected stations. It appears that as a station began filling, these stations became less desirable destinations (relative to other stations) as exhibited by the decreasing destination effect b_{jt}. This may suggest that some individuals planned their trip to avoid this station based on outside knowledge that, with high probability, they would not be able to dock at these stations. In other words, the trend of the station filling is itself predictive of the decreasing demand for docks at that station.

In addition to this phenomenon, it is possible to add other components to this relatively simple model to better describe the underlying network flow processes. Based on the exploratory data analysis, there appear to be two daily "seasonal" trends in this dataset: one for the steadily increasing recreational activity starting in the late morning and another for the abrupt commuter activity. Thus, the demand for each time interval could be described as a mixture of recreational activity and commuter activity. Future models may

be able to account for both trends which could provide further insight to the demand and, presumably, improve the characterization of the underlying demand. Additionally, future models may also want to add exogenous variables that allow this model to generalize over days with different conditions.

Chapter 5

Conclusion

While there has been a proliferation of studies leveraging network flow data, few take into account the change in network behavior due to constrained occupancy at a destination node. This thesis addresses this shortcoming in the literature by outlining a general methodology that estimates the underlying flow when these constraints are present. Insights into the underlying flow, which are highlighted in the application to bike sharing data and can be straightforwardly be extended to similar data, have practical relevance for policymakers and researchers.

The application in this thesis develops applies the methodology for characterizing the underlying flow between individual bike stations for the purposes of understanding network dynamics when stations have been filled to capacity. This level of granularity is unusual in the context of other bike sharing literature which typically models the data at a station level. Modeling it using the DGM affords unique insight into the determinants of traffic into and out of individual stations as well as general trends in the network overall. This is especially helpful when considering redistribution of bikes from overcrowded to vacant stations. Identifying times of consistent surges in activity between two stations can inform these efforts about when and where the bikes should be redistributed. Furthermore, characterizing the *underlying* flow highlights activity into or from a particular station that most significantly contributes to the source of a shortage or surplus in bikes.

This methodology is also relevant to other transportation networks such as vehicle traffic flows and public transit networks. In these contexts, it is worth noting that the networks may be much larger in terms of number of nodes as well as number of edges. In the cases of large dense networks, the scalability of the decouple/recouple strategy becomes especially relevant. Since the sequential and retrospective analyses are performed independently for each edge of the network, this naturally encourages a parallel implementation on a GPU

[8]. The application in Chapter 4 used a network with $2,038$ edges which ran in under 10 minutes on a standard laptop that leveraged Matlab's parallel processing functionality on a CPU.

Indeed, this methodology can be extended rather straightforwardly to any network through which discrete elements are exchanged and the recipients of these elements (nodes) have a limited capacity. A notable application is explored in [5] which considers the flow of patients among hospitals. In a similar vein, hospitals themselves often struggle with appropriately assigning patients to different departments internally due to a limited number of beds in each department. The intermittent flow of patients between departments within the hospital often reflects scarcity of resources in some departments and it is not unusual for every bed within a department to be full. Characterizing the flow of patients to departments that are at or near capacity could provide insights for future demand planning and could lead to a reorganization of resources to improve patient care.

Abstracting these methods further, the methodology for characterizing underlying demand can be used straightforwardly in other applications that involve (possibly many) conditionally independent, intermittent counts. For instance, one could also consider an application in product demand planning. Specifically, the demand for multiple related products is, of course, limited by the inventory at a retail location. The demand for each product could be modeled independently where, in the event an item is sold out during one of the time intervals, the demand for that item would be censored and the underlying demand could be estimated. These conditionally independent models can then be recombined to account for covariation between items (see [16]). It is also worth mentioning that in the context of supply chain management, one could construct a network under the assumption that the flows are limited by inventory constraints.

The model described in this thesis builds on the highly flexible and scalable approaches for modeling count time series that have been successfully implemented in a variety of domains. However, in these applications, observed count may not adequately represent the underlying count. Thus, the main theoretical contribution of this work is the development

of a methodology to update prior beliefs about the underlying count when observations are censored. There is great potential to apply this framework to other similarly-structured datasets which can provide modelers and policymakers relevant insights into the problems they seek to address.

Bibliography

[1] O. Anacleto, C. Queen, and C. J. Albers. Forecasting multivariate road traffic flows using Bayesian dynamic graphical models, splines and others traffic variables. *Australian and New Zealand Journal of Statistics*, 55:69–86, 2013.

[2] L. R. Berry and M. West. Bayesian forecasting of many count-valued time series. *Journal of Business and Economic Statistics*, 2019. arXiv:1805.05232. Published online: 25 Jun 2019.

[3] H. Chen. Dynamic network modeling. Masters Thesis, Department of Statistical Science, Duke University, Durham, North Carolina, 2019.

[4] X. Chen, K. Irie, D. Banks, R. Haslinger, J. Thomas, and M. West. Scalable Bayesian modeling, monitoring and analysis of dynamic network flow data. *Journal of the American Statistical Association*, 113:519–533, 2018.

[5] P. Congdon. A Bayesian approach to prediction using the gravity model, with an application to patient flow modeling. *Geographical Analysis*, 32(3):205–224, 2000.

[6] J. D. Croston. Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society volume*, 23:289–303, 1972.

[7] D. Gammelli, I. Peled, F. Rodrigues, D. Pacino, H. A. Kurtaran, and F. C. Pereira. Estimating latent demand of shared mobility through censored gaussian processes, 2020.

[8] L. Gruber and M. West. GPU-accelerated Bayesian learning and forecasting in simultaneous graphical dynamic linear models. *Bayesian Anal.*, 11:125–149, 2016.

[9] R. Kalman. Mathematical description of linear dynamical system. *Journal of the Society for Industrial and Applied Mathematics*, 1(2):152–192, 1963.

[10] S. Minhas, P. Hoff, and M. Ward. A new approach to analyzing coevolving longitudinal networks in international relations. *Journal of Peace Research*, 53, 3 2016.

[11] R. Prado and M. West. *Time Series: Modeling, Computation and Inference*. Chapman & Hall/CRC Press, Boca Raton, Florida, 2010.

[12] C. Tebaldi, M. West, and A. F. Karr. Statistical analyses of freeway traffic flows. *Journal of Forecasting*, 21:39–68, 2002.

[13] M. West. Statistical inference for gravity models in transportation flow forecasting. Discussion Paper 94-20, Duke University, Durham, North Carolina, and Technical Report #60, National Institute of Statistical Sciences, Research Triangle Park, North Carolina, 1994.

[14] M. West. Bayesian forecasting of multivariate time series: scalability, structure uncertainty and decisions (with discussion). *Annals of the Institute of Statistical Mathematics*, 72:1–31, 2020.

[15] M. West and P. J. Harrison. *Bayesian Forecasting and Dynamic Models*. Springer-Verlag, New York, 2nd edition, 1997.

[16] Z. Y. Zhao, M. Xie, and M. West. Dynamic dependence networks: Financial time series forecasting and portfolio decisions (with discussion). *Applied Stochastic Models in Business and Industry*, 32:311–339, 2016.

[17] X. Zhou, J. Nakajima, and M. West. Bayesian forecasting and portfolio decisions using dynamic dependent sparse factor models. *International Journal of Forecasting*, 30:963–980, 2014.