Title
Bicyclist Exposure Estimation Using Heterogeneous Demand Data Sources

Author
Proulx, Frank Roland

Publication Date
2016

Peer reviewed|Thesis/dissertation

Bicyclist Exposure Estimation Using Heterogeneous Demand Data Sources

by

Frank Roland Proulx

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Civil and Environmental Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alexey Pozdnukhov, Chair
Offer Grembek, PhD, Co-chair
Professor Joan Walker
Professor Michael Anderson

Fall 2016

Bicyclist Exposure Estimation Using Heterogeneous Demand Data Sources

Abstract

Bicyclist Exposure Estimation Using Heterogeneous Demand Data Sources

by

Frank Roland Proulx

Doctor of Philosophy in Civil and Environmental Engineering

University of California, Berkeley

Professor Alexey Pozdnukhov, Chair

Offer Grembek, PhD, Co-chair

Quantifying risks and the effects of risk factors requires controlling for exposure, or the number of opportunities for the adverse outcome in question to occur. In the context of traffic crashes, traffic volumes are frequently used as an exposure measure. Efforts to study bicyclist crash risk have historically been hindered by the lack of widespread exposure data. This study presents methods to estimate bicycle traffic volumes across an entire urban network.

The first major chapter of the dissertation presents a data schema for classifying bicycle demand datasets. There is an ever-growing abundance of transportation data, with some of the fastest growth seen in realm of non-motorized demand. However, all of the available datasets provide incomplete information about the system. For example, some only represent a time series of observations at a single location in space (automated counters), while others cover all space and time but only represent a small subset of the population of people and trips (crowdsourced data). In order to understand how these heterogeneous sources of information correspond to one another, it was deemed necessary to first identify their differences. Six metadata characteristics were defined, which are termed the population scope, trip aggregation, temporal scope, temporal resolution, spatial scale, and demographics. Levels are defined for each dimension, and examples of generic datasets are discussed in terms of their metadata dimension.

The second major chapter of the dissertation presents a method of fusing multiple link-level demand estimates to infer peak-hour bicycle traffic volumes. While the method is agnostic to the specific sources being used, it is presented with a case study of San Francisco, CA using data from regional travel demand models, a smartphone crowdsourcing application, and bikeshare system ridership. The de-

fined process entails first converting the datasets to a common format in terms of their metadata dimensions, and then fitting these homogenized link-level estimates to observed counts using a weighted regression technique modeled after Geographically Weighted Regression. The fitting parameters associated with each dataset are hypothesized to vary geospatially, and the means by which this variation occurs is controlled by the specified weighting scheme. A distance decay weighting, where observations further from a given location contribute less to the parameter estimates, is found to produce the best results. Cross-validation is employed for model comparison and the selection of features and hyperparameter values. It is shown that, on the basis of cross-validated Root-Mean Square Deviation, that fusing data sources provides greater predictive accuracy than can be achieved using any individual source, and that utilizing localized regression is more predictive than using a single global parameter for each data set.

The final chapter is about inferring the temporal distribution of traffic based on continuous automated count data. Latent Dirichlet Allocation is applied as a signal decomposition model to identify latent spatio-temporal patterns in the observed count data, which appear to correspond to coherent activity patterns such as AM commuting, PM commuting, and midday cycling. Each link's temporal distribution can thus be expressed in terms of the extent to which each latent pattern is observed on it. The mixture of these patterns on unobserved links is interpolated using a purely autoregressive model, in contrast to the historically ad hoc methods used to determine the temporal characteristics of bicycle traffic on unobserved links.

The primary conclusion of this work is that the lack of exposure data should no longer be considered an insurmountable problem for studying bicycle crashes. Using advanced analytical methods, such as those presented here, in conjunction with the abundance of new datasets provides a means of generating defensible retrospective volume estimates for the entire network. This dissertation paves the way for many future lines of inquiry, including both refinements upon the methods presented here and application of the volume estimates developed here to problems requiring exposure quantities, such as the evaluation of crash risk.

To Jaya, for inspiring me through thick and thin.

Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to thank my advisors Professor Alexey Pozdnukhov and Dr. Offer Grembek for being perpetual sources of encouragement, advise, and ideas. Second, my committee members Professors Joan Walker and Michael Anderson for pushing me to make this dissertation the best that it can be.

I would like to also thank everybody in the SafeTREC family. You all do such inspiring work. In particular, Bob Schneider for being an early and continuing source of inspiration. Thanks for keeping me employed while in graduate school.

Further, all of my Cal friends, especially Timothy Brathwaite for always being a step ahead on stats and ready to help, Darren Reger for showing me how to pick up heavy things and keep a balanced life, and Teddy Forscher for helping me to actually ride bicycles rather than just study abstractions of them. Thanks to Siyu Chen for all of the help with data management - at this rate, you're going to have finished 5 dissertations by the time you finish undergrad.

And, my family. Dad, for passing on the bicycle AND computer science interests to me from a young age, and for all your help and encouragement along the way. Mom and Lucy, I wish you were still here to see what I've accomplished. Momma and Baba, for loving me as a son. Vipul Bhaiya, for putting me up for TRB every year, and the warm welcome into the family.

Above all, thank you to my wonderful wife Jaya. This PhD is an accomplishment of both of us, and would not be possible without your unending support and patience. We made it!

This research was funded in part by the University of California Center on Economic Competitiveness in Transportation and the Federal Highway Administration's Dwight D. Eisenhower Fellowship program.

Chapter 1

Introduction

Cities, states, and countries around the world are increasingly turning to encouraging bicycling as a space and energy efficient form of transportation, a means of promoting public health through increased activity, and simultaneously are increasingly prioritizing traffic safety. For example, the recently released Caltrans Strategic Management Plan identifies simultaneous goals of a 10% reduction of bicyclist fatalities per year, an as-yet-unspecified reduction in the number of bicyclist injuries, and a tripling (percentage-wise) of bicycle mode share by 2020 [Brown Jr. et al., 2015]. While well-intentioned, these goals are fundamentally at odds with each other unless risk to bicyclists is reduced, where "risk" can be understood in a Bernoulli trial sense as the number of expected crashes for a given number of "trial" events. In the context of epidemiological studies, such as those looking at traffic safety, these trials are often referred to as "exposure" [Hauer, 1982].

1 Motivation

The factors underlying risk, particularly to bicyclists, are not thoroughly understood.

When traffic crash prediction models are developed using geographic entities (such as intersections) as the unit of study, traffic volumes are typically used as a proxy for exposure. In the case of multi-class crash models (e.g. bicycle-motor vehicle; pedestrian-motor vehicle), the expected number of crashes is typically expressed as a function of the volumes of both road user classes in an attempt to control for exposure. However, bicycle crash models have historically been hindered by the lack of extensive collection of bicycle volume data. Furthermore, the ideal relationship between the two volumes to proxy for exposures has not been well-established, and is further complicated by the fact that the relevant exposure quantity might

vary depending on the type of crash under consideration.

While many cities in the United States are quickly implementing the collection of bicycle count data as a routine activity, we can expect that there will continue to be poor coverage of this data for years to come. However, many other sources of bicycle demand data are available that, while not providing direct measurement of link-level demand, can be used to help infer volumes across the network. These additional demand data sources include travel demand model estimates, bikeshare usage data, and crowdsourced trip data. Each of these demand data sources are subject to various strengths and limitations.

2 Dissertation Outline

The remainder of the dissertation is structured as follows. Chapter 2 discusses literature relevant to the questions of bicycle risk evaluation and, as a natural extension, exposure estimation. Following that, the bicycle demand datasets under consideration are presented in the context of a novel metadata schema in Chapter 3. In Chapter 4, a method for estimating demand during a single time period (in this case, the PM peak) by fusing together demand estimates from multiple sources using weighted regression. The initial focus is on the PM peak due to the abundance of "ground-truth" counts conducted during this period, as many communities rely upon short-duration manual counts for bicycle volume data collection. To get from the PM peak to total bicycle traffic volumes, some knowledge of the temporal distribution of traffic on each link must be assumed. In Chapter 5, automated bicycle count data is decomposed using Latent Dirichlet Allocation (LDA). This analysis presumes that there are latent temporal travel patterns that have some degree of spatial/directional order to them, which appears to be true. Finally, in Chapter 6 some concluding thoughts are presented on the analysis contained herein, including acknowledgment of limitations and suggestions for future work.

Chapter 2

Literature Review

This chapter reviews the literature as relevant to bicyclist risk evaluation. As has been suggested in the Introduction, a critical component of risk evaluation at the site level and the primary focus of this dissertation is the estimation of bicycle volumes as a measure of exposure. The literature review therefore covers both bicycle crash models in general and approaches to bicycle volume estimation.

1 Crash Prediction/Safety Performance Functions

Analyzing traffic crashes is frequently broken into multiple conditional probabilities, with separate models estimated for crash severity (conditional on a crash having occurred) and crash frequency Lord and Mannering [2010], Savolainen et al. [2011]. When analyzing infrastructural contributors to crash frequency, the location at which crashes occurs must be recorded. The most commonly used model for predicting crash occurrence is the so-called "Safety Performance Function (SPF)", which expresses the expected number of crashes within a given spatial extent (e.g. intersections, road segments) as a function of exposure [Hauer, 1982]. *Exposure* here refers to a quantification of events that could potentially result in a crash. When working with spatially disaggregate crash data, such as using intersections or segments as the unit of measure, annualized traffic volumes are typically used as a measure of exposure.

In the context of motor vehicle (single- or multi-vehicle) crashes, an exposure measure such as Annual Average Daily Traffic (AADT) is most commonly used. However, when considering multi-class crashes such as between bicycles and motor

vehicles, volumes of both vehicle classes need to be taken into account, i.e. controlling for both AADT and Annual Average Daily Bicyclists (AADB).

Using the number of crashes as an outcome measure suggests a count-regression model, and therefore most studies consider a Poisson-family regression model. While many modifications upon the basic Poisson model have been considered in the traffic safety literature (e.g. Negative-Binomial, Zero-Inflated Poisson, Zero-Inflated Negative-Binomial), in all cases the exposure and other risk factors are used to predict the rate parameter of crash occurrence, i.e.:

$$C_i \sim \text{Poisson}(\mu_i) \tag{2.1}$$

$$\log(\mu_i) = \beta_0 + \beta_1 \log(AADB_i) + \beta_2 \log(AADT_i) + \sum_3^n \beta_j X_{ij} \tag{2.2}$$

where

$$\begin{aligned} C_i &= \text{Number of crashes at location (segment or intersection) } i \\ \beta_j &= \text{Estimated parameters} \\ X_i j &= \text{Additional risk factors associated with observation } i \end{aligned}$$

There have been relatively few bicycle SPFs developed, due to both the paucity of available exposure data for bicyclists (AADB), and to do the relative rarity of crashes. Most cities do not have observed bicycle traffic volumes at a sufficient number of locations colocated with crashes to estimate a robust model.

Elvik [2013] identifies three such studies, with two considering intersections and one considering road sections. All three studies document the so-called "Safety-in-Numbers" effect, where the coefficient corresponding to bicycle volumes is less than 1. This suggests that as the number of cyclists increases, the expected number of crashes per cyclist (and therefore the risk posed to any individual cyclist) decreases.

In addition to simply considering exposure in the SPF, additional risk factors can be included via the X_j in equation 2.1. For example, Jonsson [2005] considers the effects of surrounding land uses, visibility, and road class on crash frequency, finding a higher crash risk associated with business districts than residential areas, which is attributed to differences in the relative temporal patterns in bicycle and motor vehicle traffic volumes. Turner et al. [2009] separately consider various types of bicycle crashes (e.g. mid-block, mid-block turning, intersection) and find an increased risk associated with bicycle lanes for mid-block bicycle-motor vehicle crashes and signalized intersections. However, it is suggested that this counter-intuitive finding is attributable to a dataset biased towards high crash locations.

Strauss et al. [2013] simultaneously estimate bicycle activity levels and bicycle crash risk at signalized intersections in an attempt to overcome the lack of observed bicyclist exposure and endogeneity between these quantities, finding a positive relationship between the presence of bus stops and total crosswalk length and crashes, and a negative relationship with the presence of raised medians. Notably, the presence of bicycle facilities was not found to have an effect on injury crash frequency.

In summary, existing bicycle crash frequency models have in general not shown strong evidence of the effect of bicycle facilities on bicycle crash risk. While it is possible that there is truly no relationship (or a positive relationship), it appears premature to draw any strong conclusions on the matter due to the severely limited extent of available data, particularly on bicyclist exposure.

2 Volume Estimation

In order to inform bicycle crash frequency models, estimating bicycle volumes at unobserved locations is an important matter. The question of bicycle volume estimation can be subdivided into two dominant paradigms: choice-based, i.e. through the assignment of trips to the network using a route choice model based on estimated demand, and facility-based ("direct-demand models") Kuzmyak et al. [2014].

Choice-based models

The behavioral approach is a component of larger-scale travel demand models, such as the fourth step in the "four-step" trip-based modeling framework, or based on the trips predicted by an activity-based model [McNally, 2008, Bhat and Koppelman, 2003]. In either case, trips by various modes of transportation are predicted based on travel or activity-travel diaries, either for the current "base case" scenario or for forecast conditions. In the activity-based framework, times of day for trips are also predicted. These trips can be summarized by an Origin-Destination Matrix (ODM), which includes the estimated number of trips between each Origin-Destination Pair by mode, time of day, trip purpose, and conceivably any other means by which trips could be subclassified.

One particular difficulty in estimating bicycling demand (especially compared with driving and public transit ridership) is the relatively high occurrence of "pure recreation" trips, or trips where there is no "destination" per-say, the trip origin and destination are both typically at home, and utility is derived from the act of travel. For example, a recent mode share survey of San Francisco identified that out of the 20% of respondents who ever ride a bicycle, approximately $\frac{1}{4}$ report rid-

ing exclusively for recreation [Corey, Canapary & Galanis, 2011]. Various studies have developed models predicting participation in pure recreational travel, primarily within the activity-based framework (which takes into account temporal constraints) [Sener and Bhat, 2012, Bhat and Gossen, 2004]. However, these models often only predict the type of recreation to be participated in (i.e. in-home, out-of-home, pure recreation), and do not specifically predict pure recreational bicycle trips, let alone the characteristics of such trips like distance and route.

Given an ODM, a route choice model is needed to determine link or intersection-level volumes in the trip assignment process. The estimation and application of route-choice models is a robust area of research unto itself, with two of the main difficulties being accounting for correlated errors between routes that use the same links of the network, and developing an analytical choice set given that the universal set of routes between two points on a network is virtually uncountable. Ben-Akiva et al. developed one of the most widely used solutions to the first problem, the path-size logit model, in which route utilities are penalized based on the length-wise degree of overlap with other routes in the choice set [Ben-Akiva et al., 1984]. Bovy provides a thorough overview of the choice set development problem [Bovy, 2009].

Numerous bicycle route choice models have been developed using stated-preference data or recall interviews (see e.g. [Winters et al., 2010, Kang and Fricker, 2013]). However, with the development of smartphones and affordable GPS transponders, some studies have moved to estimating route choice models based on observed routes, which is not subject to imposing a choice set nor to problems of recall [Broach et al., 2012, Hood et al., 2011].

Broach et al. [2012] distributed GPS transponders to 164 cyclists in Portland, OR and estimated a route choice model based on utilitarian trips, using the Path-Size Logit adjustment to a Multinomial Logit model. Broach et al. find intuitive effects of distance, topography, turns, motor vehicle traffic volumes, and bicycle facility types (as well as some interactions) on route attractiveness to cyclists. Similarly, Hood et al. [2011] estimate a bicycle route choice model for San Francisco, CA based on data collected using the San Francisco County Transportation Authority's "Cycle-tracks" cellphone application and find intuitive effects associated with route length, turn density, wrong way riding, topography, bicycle facilities, including separate parameters for up-slope by gender and trip type (commute vs. non-commute).

The choice-based modeling technique has the strong advantage that it is based in econometric decision theory, and therefore is formulated around explicit models of human behavior. This lends a behavioral interpretation to any model parameter estimates (e.g. one can infer monetary value of time based on the marginal rate of substitution between monetary costs and time in a mode choice model), and is better suited to forecasting because of this. However, when the goal of demand

estimation is volumes across links or through intersections, direct observations of these quantities should be taken into account, which is not directly accommodated in the choice-based framework.

Facility-based models

As opposed to choice-based models which utilize travel surveys as the primary data input to infer choice rules for travelers, facility-based models primarily rely on observed volumes at discrete locations in space (i.e. along links or through intersections) to estimate activity levels. Facility-based models can further be decomposed into the tasks of temporal extrapolation and spatial interpolation. Specifically, given that the desired exposure measure is Annual Average Daily Bicyclists but manual traffic counts are frequently taken during a short duration (such as the PM peak), an understanding of how traffic is distributed across time at each location is needed to determine what a particular PM peak traffic volume implies about AADB. The spatial interpolation question, on the other hand, has to do with inferring how consistently measured traffic volumes differ in space.

Temporal Extrapolation

In order to understand the temporal variation in traffic volumes, continuous automated counters are installed at a subset of locations. The Federal Highway Administration's Traffic Monitoring Guide recommends an approach to temporal extrapolation known as "factoring," where the continuous count sites are classified into "factor groups" based on similarities in their temporal profiles Federal Highway Administration [2013]. The means of the proportions of overall traffic occurring within each sub-interval (e.g. hour of the day) are found for each factor group, and these extrapolation factors are used to normalize any short-duration observed volumes.

Sites falling within a common factor group is frequently understood to not be a function of the magnitude of volumes at the sites, but simply of how those volumes are distributed across time. In practice, multiple continuous count stations with similar traffic distribution patterns often have their expansion factors averaged to create group averages. Assigning even the continuous count sites to groups, then, despite complete exposition of the patterns at these sites requires some thought. For motorized traffic, the TMG recommends defining factor groups based on federal functional road classification. However, this approach is more or less ignorant of the behavioral basis of travel. For instance, a given city might have two multi-use trails, one through the middle of the city and one on the outskirts. We can imagine that these would have very different traffic patterns, with the inner city path carrying a

much higher share of commute traffic and therefore having both stronger AM/PM peaks on weekdays and a higher ratio of weekday:weekend traffic, despite being the same facility type.

The development of factor groups for bicycle traffic has been previously considered. For example, Miranda-Moreno et al. [2013] consider

, based on data from 37 continuous counters installed in five North American cities identifies four factor groups present in the data, and terms them "Recreational", "Mixed Recreational", "Mixed Utilitarian", and "Utilitarian" ascribing a pseudo-behavioral interpretation to the observed patterns. The assignment method in this study involved an iterative manual grouping, where sites were grouped based on hour of day weekday, weekend, and day of week distributions, 95 % Confidence Intervals were calculated for each group's distribution, and sites were reassigned based on whether they fell within the CI or not. Unfortunately, no insight is provided on what the underlying cause of these patterns may be nor is any guidance provided on how to assign an additional site to one of the defined classes. Roll [2013] follows a similar methodology using short-term automated counts of 1-2 weeks in Eugene, OR to identify an additional factor group, which he attributes to the presence of a major university.

Nordback et al. [2013] uses continuous count data from 12 stations in Boulder, CO to explore the extrapolation error of calculating AADB using the TMG Factoring Method based on the amount of "short-duration" count data available, assuming simply a commute/non-commute grouping, and finds a point of diminishing returns around one week of data, with an average absolute percent deviation of 20-30%. Further, extrapolation error is shown to be lowest when the short duration counts are collected when volume variability is lowest, which in their case was between May and October.

El Esawey et al. [2013] investigate various methods for calculating daily adjustment factors for bicycle traffic volumes based on data from Vancouver, BC, and recommend using factors specific to each month of the year. Additionally, they find similar estimation accuracy when using simple weekday vs. weekend factors compared with using a separate factor for each day of the week.

There are a number of problems with the factoring approach for extrapolation. First, it relies upon knowing what group a given site belongs to without having long term count data, which is the basis for the groupings. Second, it ignores any deviations from the systematic time effects due to other factors, namely weather, which has been shown to have a substantial effect on bicycling activity levels.

As mentioned earlier, there are also a number of papers that have explored the temporal variation in bicycle volumes based on long-term patterns. In the earliest identified longitudinal analysis of bicycle volumes, Niemeier [1996] explores AM and

PM peak period counts from 5 sites in Washington using Poisson regression, and finds greater variability in the PM peak period than in the AM peak period, negative associations between volume and temperature and volume and precipitation, and that temperature appears to be a stronger predictor of volumes than precipitation. Tin et al. [2012] consider continuous count data from a single counter in Auckland, NZ using ANOVA and OLS regression, and find negative associations between wind and rain and bike volumes, and a positive association between maximum temperature and the presence of sunshine. Gallop et al. [2012] and Nosal and Miranda-Moreno [2014] both extend the multivariate longitudinal analysis of bicycle volumes based on weather conditions to account for serial autocorrelation, using a Seasonal ARIMA model in the case of Gallop et al. and a simpler ARMA model in the case of Nosal and Miranda-Moreno.

Spatial Interpolation

Beyond the task of expanding volume estimates across time, we face the problem of expanding across space. Whereas the problem of accounting for temporal variation in volumes is a function of the time and weather, spatial variation in volumes is typically explained (in the facility-based perspective) in terms of characteristics of the surrounding environment, such as land use, transportation network characteristics and structure, and sociodemographics. This approach is known as "direct-demand modeling", where volumes at a location are regressed on characteristics of the surroundings. For instance, a direct-demand model may predict bicycle volumes based on population density, number of transit stops, and presence of bicycle facilities within a 500-foot radius around the intersection. This approach is tractable and fairly easy to estimate, but lacks in behavioral realism when considering bicyclists. While the immediate surroundings of an intersection may be a good predictor of pedestrian volumes, bicyclists have greater ability to deviate paths (e.g. to preferred facilities) and travel greater differences (e.g. trips may be generated by and attracted to locations far from the intersection, yet still require routing through the intersection). These mechanisms are not captured well by direct demand models. Additionally, "temporal activation" of land uses is not typically considered (e.g. schools should only have a substantial effect on volume during school hours) - rather, all existing models either consider a temporally aggregate volume estimate (e.g. AADB), or only use a short-duration count period such as the 2-hour PM peak period.

The majority of published direct demand models for estimating bicycle volumes utilize either short-duration counts or aggregate counts. Regressors are considered within buffers of varying radii, such as 0.1mi, 0.25 mi, and 1.0mi, and are typically entered in either an ad hoc, forward stepwise, or backward elimination manner. The

regression models used are frequently either an Ordinary Least Squares (linear or log-linear), or else a count regression (Poisson/Negative Binomial). Basic example of bicycle volume direct-demand models include Haynes and Andrzejewski [2010], Griswold et al. [2011] and Lindsey et al. [2006].

One substantial problem with the majority of bicycle volume direct demand models is that they do not account for spatial autocorrelation - that is, they ignore the spatial quality of the data under consideration. Strauss and Miranda-Moreno [2013] attempts to account for this using an inverse distance weighted spatial lag term. Count data in this paper are extrapolated to annualized values, accounting for weather in multiple ways. Additionally, varying impedance values are tested for the spatial lag term.

3 Crowdsourcing

Crowdsourcing of bicycle demand data entails asking the public to report their trips, typically using GPS-enabled smartphones. This is frequently either done by transportation agencies with the primary goal of understanding bicycle demand, as in the case of the San Francisco County Transportation Authority's "CycleTracks" application Hood et al. [2011], or by private companies whose primary product is analytics on exercise patterns for app users, but who have developed a secondary product of selling anonymized demand data to transportation planning agencies as in the case of Strava Metro [Strava, Inc., 2014].

Applications of and analytical techniques for crowdsourced bike demand data are a nascent field. The are two dominant applications as pertaining to the estimation of link or intersection-level volumes. First, if crowdsourced data are available in a disaggregate, trip-based format, it can be used to estimate a route-choice model, which can then be utilized to assign trip estimates for which only origins and destinations are available to the network, as previously discussed [Hood et al., 2011, Broach et al., 2012]. In some cases, such as Strava Metro, trips are not represented by individual traces, but are instead aggregated to link counts which can be taken as a proxy for volume as in Griffin and Jiao [2015]. However, it is possible and indeed quite likely that this crowdsourced data is biased demographically, temporally, and spatially compared to the true aggregate volumes. To make rigorous evaluations of total bicycle demand using crowdsourced data requires a thorough understanding of what these biases might be, which is to date not well understood in a general sense.

4 Summary

The overall literature gap that this dissertation seeks to make progress towards is the lack of a thorough understanding of bicycle crash risk factors. One of the biggest reasons for this current gap is the lack of quality exposure data, which in turn represents its own gap in the literature. This is where the attention for the remainder of this work will be, and in particular as it pertains to finding a happy median between the current modeling techniques of utility-based travel demand modeling and facility focused "direct-demand" modeling, while also taking into account the rapid growth of crowdsourced datasets that are currently ill-understood.

Chapter 3

Bicycle Demand Data Sources

In this chapter, I will discuss the variety of spatially referenced bicycle demand data sources available. All forms of bicycle demand data ostensibly measure the same underlying phenomenon (people traveling via bicycle across spacetime). However, there are no perfect sources of spatiotemporally resolved information on travel behavior, with all available datasources suffering censorship temporally or spatially, or biases in the populations of people or trip types represented. First, I present a metadata schema intended to assist in identifying how the various available demand datasets relate to one another and how they might need to be homogenized to enter into a common predictive framework. Then, I present various demand datasets. For each, I describe the type of data source generically, including classification according to the proposed metadata schema. Then, I discuss the particulars of the source as collected in my case study city of San Francisco, CA, including some exploratory analysis.

1 Metadata Schema

Demand data can be classified based on its collection format and on its representational format. The collection format describes the raw underlying data, while the representational format describes the data as available to the modeler. Representational formats are not fixed- that is, demand data can be changed between formats, which is the crux of the method proposed herein. Implicitly, then, some data sources might be considered by the analyst until after processing has been performed (e.g. by the data provider).

The proposed schema categorizes datasets on six metadata dimensions, termed the population scope, trip aggregation, temporal scope, temporal resolution, spatial scale, and demographics. Each of these will now be discussed in more detail, including

presentation of possible values that they might take.

Population Scope The population scope defines whether a given dataset is intended to represent the *full population* or a *subset of the population*, where the "population" here is the population of both people and trips. Generally speaking, we are primarily interested in full population estimates; however, there are relatively few methods that are able to capture this. Naturally, when only a subset of the population is represented we must wonder how this sample relates to the full population, and frequently do not have a strong understanding of this connection.

Trip Aggregation Trips can either be represented as aggregate, cross-tabulated quantities or as discrete records.

Temporal Scope The temporal scope denotes the extent of data representation, temporally. For example, a given dataset might provide a full historical time-series or a partial/truncated historical time-series, where observations are attributed to specific date-times. Alternatively, one might encounter an average or "typical" time-series, where either a full/partial time-series was aggregated to an average representation or else an estimate was generated to describe the expectations for a recurrent time period, such as a "normal weekday". The typical time-series definition could be extended beyond the "normal weekday" to account for other factors that might affect travel patterns, such as by generating separate estimates for rainy weekdays and non-rainy weekdays.

Temporal Resolution Whereas the temporal scope details the temporal extent of data, the temporal resolution describes how finely observations have been sliced. There are three primary options for handling this dimension: disaggregate timestamped observations, aggregation to a fixed binning interval (e.g. 15-minutes), or aggregation to a "semantic" binning pattern where bins are defined according to a behavioral intuition. For example, delineating the AM Peak, Midday Off-Peak, PM Peak, and Evening/Night Off-Peak.

Spatial Scale Travel occurs in space as well as time. Accordingly, the spatial dimension must be captured, which can be done with varying degrees of acuity:

- *Site*-based observations are taken at discrete locations in space, such as along links or through intersections.

- *Traces* are empirical records of routes taken, such as those taken using GPS transponders. These can either be in free-space or "map-matched" to the network.
- *Origin-Destination point* datasets provide details of the specific start and end points of trips, but no indication as to the route taken.
- *Origin-Destination zone* datasets only expose the approximate start and end location of trips, again with no explicit routing. These origins and destinations are at the scale of the Travel Analysis Zone (TAZ) system used.

Demographics Any additional details about the nature of trips represented are captured under demographics. This might include details on the trip-makers (e.g. age, gender, income), helmet-usage status, trip purpose, whether trips were made in a group, or any other characteristics of interest that could be collected.

2 Data Sources

Given that we now have a unifying framework for categorizing bicycle demand datasets, I will now present some commonly available types of data, as well as specific details of these datasets as they are used throughout this dissertation in a case study of San Francisco, CA.

Counts

Traffic counts are collected to evaluate demand at fixed locations in space (i.e. are a site-based measure). They can be collected either on a short-duration or continuous basis, with automated technologies or by hand, and either at intersections or along segments. Counts are intended to be a full-population measure, and any deviations from this are assumed to be random due to counter limitations. In San Francisco, both automated and manual counts have been collected at the locations shown in Figure 3.1. The manual count locations shown here are for the 2014 data collection effort, which is used in this dissertation. The automated counters have been installed gradually over time.

Automated

Bicycle counts can be collected with an automated traffic recorder either as a partial time-series or as a full time-series, depending in part on the particular technology

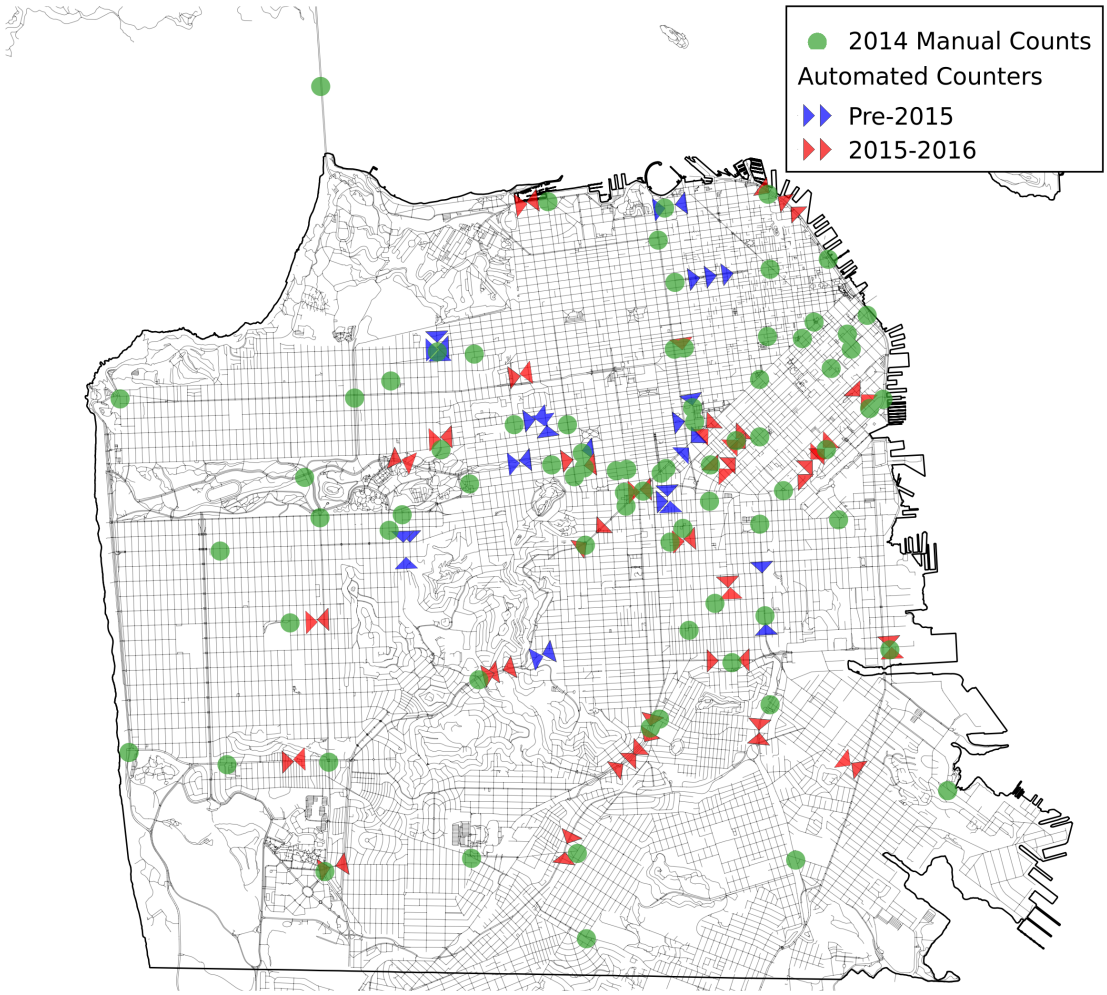


Figure 3.1: Locations of automated and manual counts in San Francisco, CA.

being used [Ryus et al., 2015]. The technologies currently on the market are similar to those used for counting motorized traffic. Continuous (full time-series) counts are collected using induction loops, piezoelectric strips, magnetometers, and radar, with automated video and automated infrared video analysis on the cusp of market penetration and adoption. Limited duration counts are most commonly collected using pneumatic tubes, although induction loops can also be used in this way. All of these technologies (with the exception of the two video-based methods) are suited only for collecting data across a screenline, meaning that they cannot be used to collect turning movements. However, they are all capable of collecting directional counts.

San Francisco has one of the highest densities of automated bicycle counters within a single city. At this time, there are approximately 80 counters installed, although a large proportion of these were installed in the course of the completion of this research. These counters are all "induction loops," similar to those commonly used for signal actuation at intersections. Induction loop counters are known to operate very accurately, but are subject to "bypass errors" when their coverage does not subtend the entire width of a facility [Ryus et al., 2015]. The counters used in this study have not been validated for accuracy with respect to bypass errors.

Manual

Manual bicycle counts can either be collected in the field or by reviewing video footage after the fact. In either case, the primary cost is time spent which scales with the extent (temporally and spatially) of data collected. This is in contrast with automated counters, which effectively have a fixed cost irrespective of the temporal extent of data collected. Accordingly, manual counts are best suited to collecting a very short partial time-series of data, which operationally for many cities amounts to a two-hour count in the AM- or PM-peak.

In San Francisco, an annual PM-peak bicycle count is conducted by volunteers. Counts are collected as turning-movements at a 15-minute temporal resolution. In 2014, the year that the analysis conducted in Chapter 4 of this study is based upon, manual counts were conducted at 77 intersections around the city.

Travel Demand Models

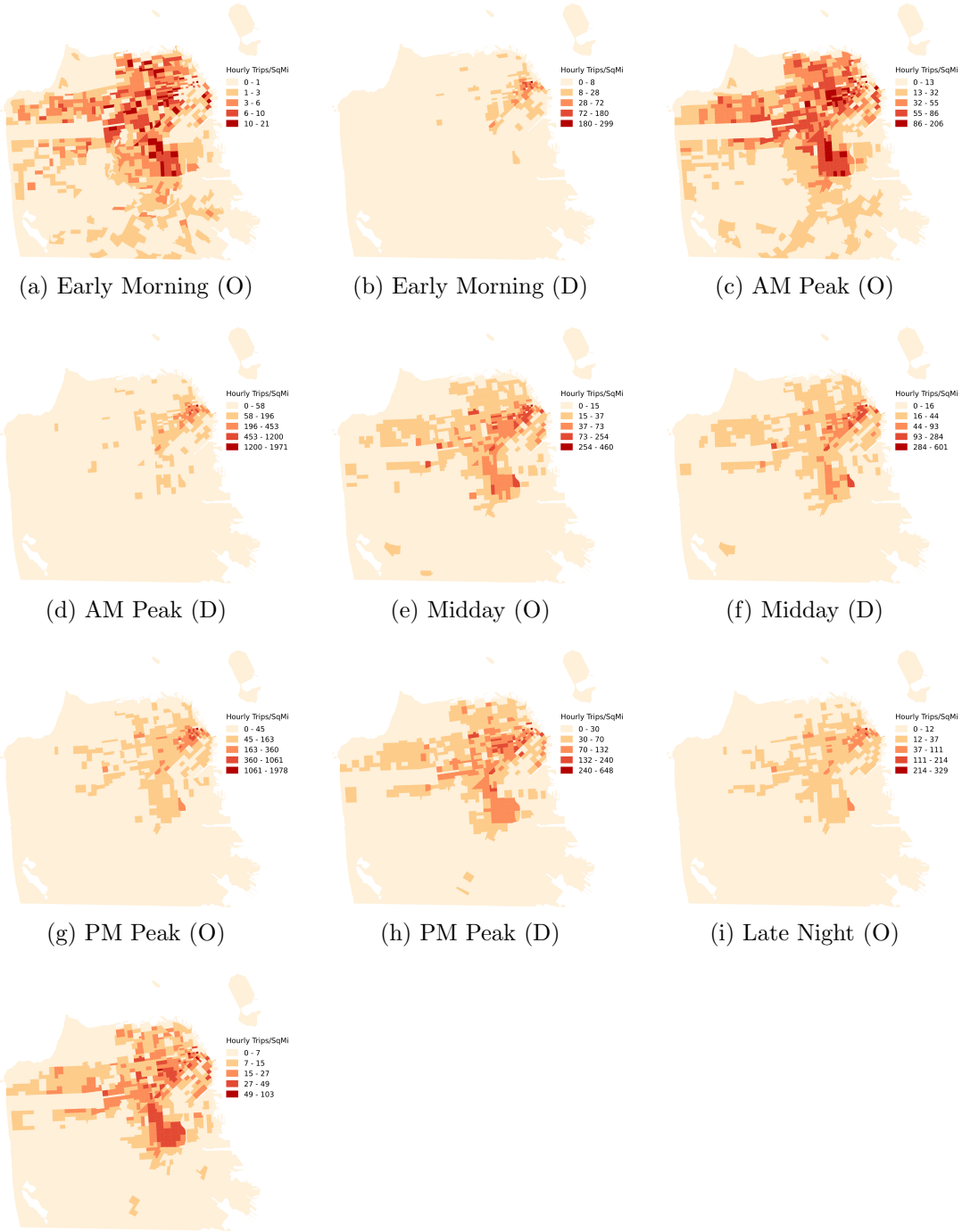
Many regional planning organizations and some large cities maintain extensive utility-based travel demand models. These models make use of detailed activity-travel surveys, typically taken on 5-10 year cycle, where respondents report for a short time-period (e.g. 24-48 hours) all of the activities that they engage in, where and

when these activities took place, and what mode of transportation was taken to travel between activity locations, as well socio-demographic details of the respondent and other members of the respondents' household. Models used in practice involve varying degrees of complexity to approximate human behavior, from the traditional four-step models to modern activity-based models. These travel demand models are used to evaluate the anticipated effects of various policies and infrastructure modifications, such as the widening of highways, raising of tolls, or addition of bicycle lanes, on overall demand as well as mode splits. "Synthetic populations" of potential trip-makers are generated for the region, and for a given policy/network scenario their travel decisions are probabilistically determined according to the demand model.

Modeling of bicycle mode choice in the U.S. context has traditionally been difficult in the utility-based travel demand model framework for a variety of reasons, perhaps most importantly low response rates of bicyclists in travel surveys due to the relatively low bicycle mode share. However, where models do include bicycling as a possible choice in decision-makers choice sets, judgments can be made about bicycle demand for a given scenario. These judgments are often made for a "typical weekday," which limits the ability to forecast demand for weekends when travel patterns are likely substantially different than on weekdays. Furthermore, the effects of factors such as weather and shifting sunrise/sunset times are not often accommodated, which are likely to have profound impacts on peoples' decisions to bicycle. Finally, recreational bicycle trips do not fit well into a model based on utility, particularly if spatial details are desired, and are often not included in these models.

For San Francisco, there are two travel demand models available. Both of these are activity-based travel demand models, which attempt to model individuals' entire day schedules and decision processes such that dependencies between individual trips are maintained. The Metropolitan Transportation Commission (MTC) maintains a region-wide model on a spatial scale using 1,454 Travel Analysis Zones. The San Francisco County Transportation Authority (SFCTA) hosts a more detailed model of travel patterns within the county of San Francisco using a more fine-grained 2,454 TAZ system.

Figure 3.2 depicts the SFCTA's zonal bicycle demand estimates for both trip origins ("O") and destinations ("D") for each of 5 time bins for a "typical weekday". Note that these panels are internally scaled, meaning that they are not directly comparable across time periods. The vast majority of the modeled trip destinations during the early morning and AM peak time periods are located in downtown, largely along the Market Street corridor, but with origins dispersed around the city. For the PM peak period, and late night, the Mission district stands out as a particularly popular destination.



(a) Early Morning (O) (b) Early Morning (D) (c) AM Peak (O)

(d) AM Peak (D) (e) Midday (O) (f) Midday (D)

(g) PM Peak (O) (h) PM Peak (D) (i) Late Night (O)

(j) Late Night (D)

Figure 3.2: SF-CHAMP estimated trips within each time bin by TAZ

Crowdsourced Data

We can consider two different types of crowdsourced bicycle demand data being collected, based on the purpose of data collection. The first is "quantified self" smartphone applications where users record bicycle trips, runs, and other workouts to track their athletic performance. These user-focused applications have had widespread uptake due to the benefits they offer to the user. Some companies have begun selling anonymized usage data, as users have typically not opted in to sharing of their disaggregate data. In terms of transportation planning, this aggregation can be advantageous, as raw GPS data is burdensome to work with. However, information is lost in the process on trip-level details, including origins/destinations and user characteristics, which makes it difficult to assess how representative a given dataset might be. The second type of crowdsourced bicycle demand data is research-focused data, where the primary purpose of data collection is to support research or planning functions.

Strava Metro

Strava is a smartphone application on which users record running and bicycling trips that they make, fitting into the "quantified self" realm. Strava Metro is an anonymized data product that Strava sells for planning purposes, where map-matched link-flows of Strava users are recorded on a minute-by-minute basis across the data collection period. These matched flows are directional, and include separate counts of "athletes" and "activities," where activities are groups of users who appear to be traveling together. In addition, separate aggregates are generated for "commute" trips, which are identified based on user tagging of records, having a different trip origin and destination (as compared with an "origin-origin" trip), and other similar contextual clues. However, in the extract provided for this case study, the commute counts were not directional, so they could not be separately considered.

There are limited ways that the Strava Metro dataset can be visualized. Similar to in Jestico et al. [2016], the main technique available is to simply filter the dataset to gain insight into how the spatial variation in reported traffic patterns varies based on the conditions applied. For instance, Figure 3.3 depicts the average daily volume on all links of the network comparing weekends and weekdays. The dominant features appear to be similar, although the Weekend traffic does appear to have great representation on the west half of the city, especially through Golden Gate Park and along the Great Highway.

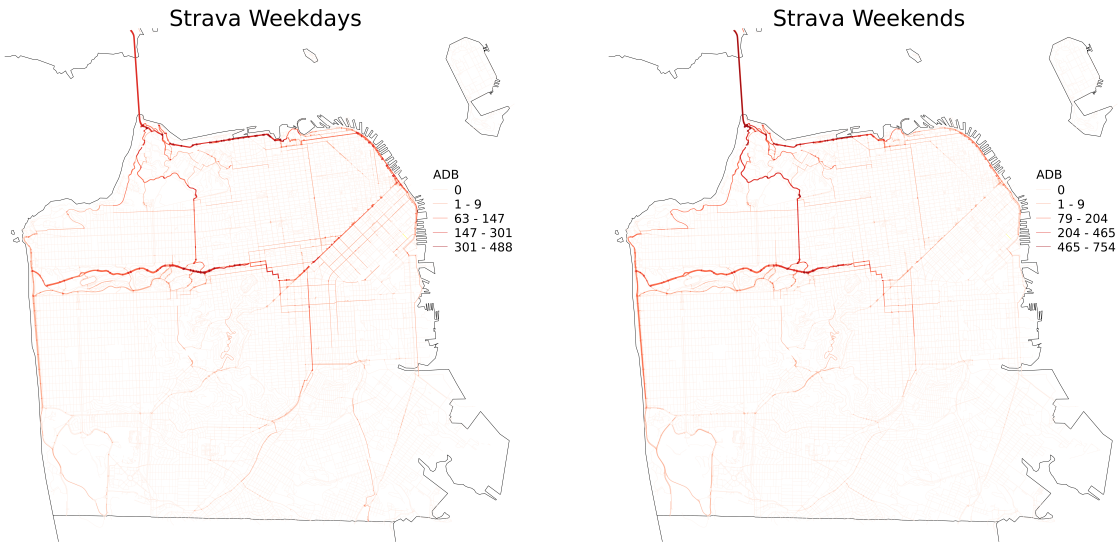


Figure 3.3: Strava aggregates for weekdays and weekends.

Cycletracks

Cycletracks is a smartphone application generated by the San Francisco County Transportation Authority to collect cycling trip GPS data for planning purposes [Hood et al., 2011]. Launched in 2009, the platform has been active continuously, with two incentivized data collection efforts in 2011 and 2013. Many other communities have adopted Cycletracks to collect their own similar data. Users selectively record bicycle trips that they make, and are prompted to provide the purpose of their trip from a drop-down menu. In addition, users are asked (but not required) to provide their age, gender, cycling frequency, and zip codes of their home, work, and school locations. Trips are recorded as GPS traces with 1 second resolution. While this dataset was considered for inclusion in the studies here, the sample size was deemed too small to be useful.

Bikeshare Data

Many cities have recently installed public bikesharing systems. Bikes are stored at a set of fixed-location docking stations, and users are free to check out and return bicycles to any station with an available dock. Users buy a daily or annual membership, and with this membership trips are priced per 30 minute at an increasing rate,

with the first 30 minutes often being free. This pricing scheme is used to increase turnover by encouraging users to return their bicycle if it is no longer being ridden. Most systems make their trip records publicly available, and have held competitions for creative analysis of their datasets. This is a full time-series of timestamped records, with an origin-destination point spatial scale.

Bay Area Bikeshare was launched in 2013, with most of its stations installed around the financial district and South of Market neighborhoods. This is a fairly constrained spatial extent, which limits the inference that can be drawn from this datasource.

Chapter 4

Data Fusion: Geographically Weighted Regression

1 Introduction

As has been argued in Chapter 3, the various available bicycle demand datasets all have shortcomings in terms of their ability to fully characterize bicycle travel patterns across the network. While details of travel patterns such as trip purposes do not necessarily need to be understood in detail for modeling risk, spatial biases in traffic volumes are a critical problem for exposure estimation. These spatial biases can manifest in many ways, either due to modeling inaccuracies or due to differences in the types of trips or people being represented by a particular information source.

The goal of this chapter is to fuse together the available demand datasets to overcome these spatial biases and achieve a higher accuracy estimate of link-level exposure than can be achieved using any individual source. A two step method is proposed. First, datasets are homogenized to a common representation following a specified series of transformations. This homogenization process is necessary to account for differences in how space and time are handled by different sources of data.

In the second step, the datasets' volume estimates are fused together by fitting to observed "ground-truth" counts with a least squares fit. In addition to a standard least squares fit where the parameter weights are the same everywhere in space, spatially varying coefficients are considered in a method similar to Geographically Weighted Regression. In the spatially varying coefficients model, separate parameters are estimated for each link of the network. The variation in these estimates is driven by applying a weighting scheme to the observations such that observations on links

that are expected to be similar to the link where estimation is being conducted contribute more strongly to the loss function than those that are dissimilar. This "similarity" is on the basis of the relative proportion of traffic explained by each dataset. Although we do not have a means of knowing how similar two links are *a priori*, various weighting schemes are considered both on the basis of spatial proximity (i.e. weights decay with distance) and on the basis of link characteristics (e.g. road geometry).

This chapter focuses on exposure estimation for a single time period (in particular, 16h00-19h00 on weekdays) due to the limited availability of ground truth counts outside of this window. In San Francisco, and in most cities, the majority of available bicycle count data comes from manual intersection counts performed during the PM peak. The larger sample size that arises from focusing on this time period allows for increased spatial resolution and hence estimation precision. The disadvantage of focusing on this small time window is that it does not tell us anything about the remainder of the week.

2 Methodology

Data Homogenization

The first step in fusing demand data to achieve volume estimates is homogenization in terms of trip aggregation, temporal scale, temporal resolution, and spatial scale. The population scope and demographic variables are typically left untouched, with the exception that one might aggregate across demographics. For a dataset representing a sub-population, we rarely have a thorough understanding of the particular details of the sub-population represented, and therefore no basis for extending to a full population representation. This lack of understanding of the sub-populations is one piece of the motivation in fusing demand data based on counts, where even "full population" representing datasets, namely utility-based travel demand models, may not adequately capture a subset of the types of trips made, such as recreational or first/last-mile trips to transit. Generally speaking, homogenization to a lower level of granularity requires fewer assumptions to be made. I will now detail the procedure for homogenizing across each of these dimensions.

Trip aggregation

In converting from a disaggregate trip representation to an aggregate trip representation, trips are simply cross-tabulated according to their characteristics on the other

metadata dimensions (e.g. spatial coordinate, time of day, and mode). To generate disaggregate trips from an aggregate representation, on the other hand, would require simulation based on our knowledge of the other metadata dimensions. For example, suppose that we have a dataset with a 1-hour temporal resolution. We might presume a uniform distribution of trip departure times within each hour, and synthesize individual trips based on this distribution. As noted, converting to this more granular representation would introduce noise.

Temporal Scope

Three temporal scope representations have been specified, including the full time-series, partial time-series, and average/typical time-series.

Converting from a full time-series to a partial time-series involves truncating the series to the specified time interval. For example, if we have one dataset which only covers a small subset of time, such as a short-duration count taken during the PM Peak period on a given day, we would truncate other dataset time-series' to match this period. To convert from a partial to a full time-series, an assumption would need to be made about how that partial time-series maps to the full unobserved time-series.

Some sources of demand data refer to a typical or "normal" day. This is particularly common in the case of travel demand models, where the typical conditions being modeled are dependent on the time period in which the underlying travel survey was collected. It is not immediately clear how these average days relate to any given day,

Spatial Scale

Converting between differing spatial scales is the most computation intensive component of this process. Specifically, to convert from an Origin-Destination format (either points or zones) to a site or trace format requires routing trips to the network based on a route choice model. Again, this includes making assumptions (ideally informed by observations), in this case about how bicyclists make their route choice decisions. Conversion from a trace format to a site format is commonly termed "map-matching," as observed GPS traces in raw format must be matched to the network geometry.

Fitting Procedure

Once a common, site-level volume format is achieved for all datasets, they are fit to observed counts according to weighted non-negative least squares. Directional links are the specified unit of analysis. Overall bicycle traffic volumes on each link are predicted as a linear combination of the volume estimates implied by each of the various datasets. That is,

$$y_q = \hat{\boldsymbol{\beta}}_q \boldsymbol{v}_q \tag{4.1}$$

where

$$\begin{aligned} y_q &= \text{ the total bicycle traffic volume on link } q \\ \boldsymbol{v}_q &= \text{ vector of estimated volumes on link } q \text{ from each dataset} \\ \hat{\boldsymbol{\beta}}_q &= \text{ vector of weights dataset weights associated with link } q \end{aligned}$$

In other words, the goal is to estimate the weights $\boldsymbol{\beta}_q$ by which each dataset's traffic volume estimates on each link should be scaled to yield the total traffic volume estimates. It bears noting that these weights have been indexed by the link q, which will be discussed shortly.

The idea behind fusing the demand estimates $\boldsymbol{v}_q$ is that each dataset yields different insight into the aggregate travel patterns, as suggested by the exploratory analysis in the previous chapter. For instance, travel demand models are primarily focused on predicting utilitarian travel such as commute trips. Crowdsourced smartphone applications, on the other hand, appear to capture recreational trips which are not easily predicted by a model.

In using a linear combination of predictions, each data source (i.e. each value in the vector $\boldsymbol{v}_q$) gets a weight β associated with it. These weights can be interpreted as the component of the aggregate travel that a single cyclist in each data source represents. For "full-population" datasets such as travel demand models, the ideal is that the associated weight is 1, as this would mean that each link count predicted by that dataset corresponds to a single count in the real world.

We hypothesize that the dataset weights could exhibit spatial non-stationarity. That is, the relationship between each dataset's predicted values and the overall traffic volume might vary between links. There are a variety of reasons that this non-stationarity could occur, such as:

- Lack of coverage for a given dataset on a subset of links (e.g. bikeshare in San Francisco only describes travel within a small area of the city where stations lie).

- Differences in reporting rates based on location (e.g. Smartphone-based reporting could be biased towards higher income neighborhoods, or towards "recreational routes").
- Biases in route preference for the sub-population being represented (e.g. different GPS datasets might have differences within their samples in terms of degree of preference for bicycle facilities, as evidenced by Watkins et al. [2016]).

Allowing the weights to vary arbitrarily on all links would not work, as we would encounter an identification problem. To get around this, we draw on a technique known as Geographically Weighted Regression (GWR), following after Fotheringham et al. [2002]. In general, GWR is specified as:

$$\underset{\boldsymbol{\beta_i}}{\text{minimize}} \qquad \sum_q (\boldsymbol{\beta_i} \boldsymbol{v}_q - y_q)^2 w_{iq} \tag{4.2}$$

$$\text{s.t.} \qquad \hat{\boldsymbol{\beta}}_{\boldsymbol{i}} \geq \mathbf{0} \tag{4.3}$$

where

$\boldsymbol{\beta}_i =$ vector of parameters estimated at location i
$\boldsymbol{v}_q =$ vector of directional link-volume estimates from set of datasets for link q
$y_q =$ observed volume on link q
$w_{iq}(\boldsymbol{\alpha}) =$ weighting factor for regression point i and observation q, controlled by hyperparameters $\boldsymbol{\alpha}$

As suggested by the spatial variation in the parameters, this regression procedure must be conducted at each location in space (i.e. link) i where volume estimates are desired. To implement this, a method for weighting observations has to be specified, which is encapsulated in the term w_{iq} above. The previously suggested identification issue is avoided here by the weighting scheme. While the weights must vary in space to yield different results than would be achieved with a standard regression procedure, the means by which they vary is controlled by the hyperparameters $\boldsymbol{\alpha}$, which are set using cross-validation. Once optimal values have been found for $\boldsymbol{\alpha}$ the weights for fitting at each location i are effectively fixed, such that the number of free parameters for fitting at this particular location, the size of $\boldsymbol{\beta}_i$, is less than the number of observations used in their determination.

All that the weights affect, then, is how much each observation on link q contributes to the loss function for estimating parameters $\boldsymbol{\beta}_i$ on link i. The specification of this weighting scheme is generally determined by the analyst, but its specification

should reflect our beliefs about the nature of the non-stationarity in the data. Links that are believed to have similar reporting rates to the regression location ought to be weighted higher, as the observations on these links can more accurately predict the dataset weights on the link in question.

Two primary classes of weighting schemes are considered here. The first is the traditional Geographically Weighted Regression, which expresses weights as a function of the distance between between link i and q. The specific functional form used is often referred to as the kernel, and here we consider the following:

$$\text{Gaussian} \qquad w_{iq} = \exp(-\frac{1}{2}(\frac{d_{iq}}{h})^2)$$

$$\text{Bisquare} \qquad w_{iq} = \begin{cases} (1-(\frac{d_{i,q}}{h})^2)^2 & \text{for } d_{i,q} \leq h \\ 0 & \text{for } d_{i,q} > h \end{cases}$$

where in both cases the bandwidth h is a model hyperparameter that affects the attenuation rate of the distance decay function. In previous work on Geographically Weighted Regression, it has been suggested that the specific kernel being used is far less important than the bandwidth value in terms of predictive accuracy [**?**]. Using a distance-decay weighting scheme supposes that the weights $\boldsymbol{\beta}_i$ are spatially continuous, or in other words that the non-stationarities in their estimates are a function of location.

The second class of weighting schemes considered here depends on defining measures of "link similarity" between the links i and q. This form conceives that multiple features $\boldsymbol{x}_{iq}$ which relate link i to q can be combined to form an overall similarity score. Again, two possibilities for this combination are hypothesized here:

- Product: If all features P of $\boldsymbol{x}_{iq}$ are constrained to the range $[0, 1)$ or are binary, $w_{i,q} = \prod_{p \in P} x_{i,q,p}$
- Logistic: $w_{i,q} = \frac{\exp(\alpha + \boldsymbol{x}'_{iq}\boldsymbol{\theta})}{\exp(\alpha + \boldsymbol{x}'_{i,q}\boldsymbol{\theta})+1}$

where α and $\boldsymbol{\theta}$ is are model hyperparameters. One difficulty that arises with using the product formulation with multiple binary features is that the set of observations affecting the fit on any particular link decreases (as an increasing number of observations receive weights of 0).

The features $\boldsymbol{x}_{iq}$ can be defined using anything that enables comparison between links i and q, and can be conceived to affect the relative reporting rates of each dataset. As one example, we could consider an indicator variable for whether or not links i and q have the same bicycle facility present. If this is included using

the "product" formulation above, we would get a "hard weighting" whereby the parameter on link i would only be informed by links with the same bicycle facility present. Using the same measure in the "logistic" formulation would instead yield a "soft" weighting where the weights are higher for links with the same facility present, but non-zero for all.

This "hard weighting" approach can also be formulated for cases where there is a known systematic non-response rate in a particular area. For example, bikeshare system data that only provides origin and destination information cannot tell us anything about travel outside the convex hull of the observed stations. This is an opportune place to apply a hard weight defined by the link lying within the convex hull or not, as without this weighting our estimates for $\boldsymbol{\beta}_i$ would be biased upward outside the zone and downward inside the zone.

In addition to specifying these weighting schemes separately, we can also easily combine them by applying both a link similarity measure and a distance-decay weighting. In this approach, the two weighting factors would be interacted. Taking the bike facility hard weight example again, this would imply that bike lanes near to each other are more similar than bike lanes further away from each other.

Model Selection and Evaluation

In order to select between model forms and to identify optimal hyperparameter values, leave one label out cross-validation is employed and fit is evaluated on the basis of root mean square deviation (RMSD). Cross-validation is the main criterion used here (as opposed to likelihood-ratio tests, for example) because our primary interest is in out-of-sample predictive accuracy [Hastie et al., 2009]. Additionally, because the emphasis here is on prediction, not explanation, tests of statistical significance have not been reported. For more detail than can possibly be given here, see Shmueli [2010] for a comparison of explanatory and predictive modeling.

The "labels" in the cross-validation method refer to the intersections at which observations were taken. On each validation fold, all of the observations from a given intersection are held out as the test set, and the remainder of observations are used for training. This is done to avoid training and testing on observations taken from links on the same intersection, which in many cases can be expected to be highly correlated with each other as all trips through an intersection cross two of the intersection's links.

3 Case Study Data

Now that the proposed model has been formulated in abstract, we will discuss the specific data to be used in the case study. The case study focuses on San Francisco, CA during the 4-7 PM period on weekdays in September, 2014. San Francisco was selected for having a relatively high commute bicycle mode share for American cities (3.5% in 2013, compared with 0.6% nationwide), as well as a wide variety of available demand data sources [U.S. Census Bureau]. The PM peak hour is the focus of this study because the majority of available ground truth data comes from manual, peak period intersection counts.

Demand Datasets

The demand datasets under consideration have been discussed in detail in Chapter 3. To briefly summarize, we are considering the results of two travel demand models (SFCTA and MTC), crowdsourced GPS data (Strava Metro), and usage data from Bay Area Bikeshare (BABS). To homogenize the spatial scale in the case study to a site-based representation, both travel demand models and the bikeshare system data must be routed to the network. Routing is performed here using a bicycle route choice model, generated using GPS traces from San Francisco, and presented in Hood et al. [2011]. The resulting PM peak volume estimates from each dataset are shown in Figure 4.1.

There are some notable patterns in these volume estimates. First, both the SF-Champ and MTC models are deficient in predicting trips along the northern waterfront/across the Golden Gate bridge. This has been traced back to the travel skims (zone-zone distance, time, and cost estimates) used in developing these population level demand forecasts, which have trips for zone pairs that would require crossing the bridge encoded as infeasible. This is hypothesized to be because there are relatively low population densities at close proximities on the opposite end of the bridge, so most trips on these zone pairs would have high costs and low populations, and hence a low expected number of utility-based bicycle trips. That is, even if these trips were not deemed infeasible prior to model application, the number of estimated trips may still be very low. It is also worth noting that the SF-CHAMP and MTC models, while both "full-population" estimates, produce vastly different maximum link volumes - the greatest value on any link suggested by SFCTA's model is nearly six times that predicted by the MTC model.

However, it is very well known that bicycling across the Golden Gate Bridge is a popular activity, as documented both by the observed traffic volumes at the bridge and by Strava Metro. Whereas both travel demand models predict a high densities of

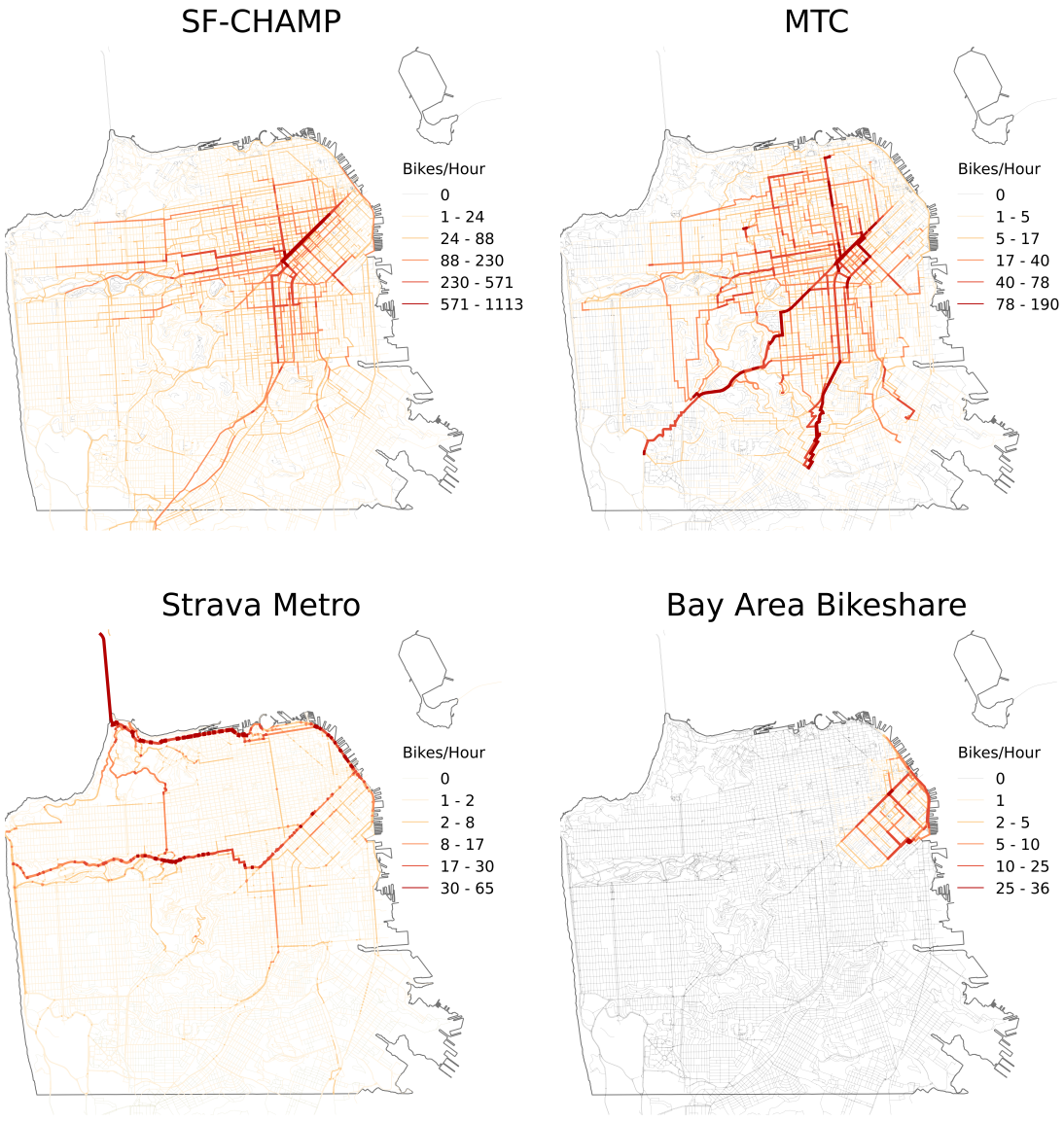


Figure 4.1: September 2014 PM Peak volume estimates for each dataset.

bicycle traffic in downtown San Francisco, Strava Metro primarily picks up on travel along the northern waterfront, Market Street, The Wiggle (a popular East-West bicycle route), and through Golden Gate Park. This difference could be attributable

to the fact that the volumes in the demand models are a result of routing, and thus are limited to the accuracy of the route choice model, whereas Strava data is observed and thus indicates the actual routes taken by users. Alternatively, this could support the common hypothesis that Strava data is disproportionately representative of recreational travel Jestico et al. [2016].

Finally, it is worth noting that the Bikeshare data is limited in spatial scope. The Bikeshare stations in San Francisco are currently limited to a small area focused around Market Street, the South of Market neighborhood, and along the Embarcadero (Eastern waterfront). Because trips are represented on an origin-destination point scale, and thus must be routed to the network to generate volumes, it is unlikely that the actual volumes exactly match those shown here. That is, these volumes assume that travel is direct between the origin and destination. This assumption is a necessity given the nature of the data, and is somewhat justified given that the pricing scheme of the bikeshare system encourages short trips to increase turnover, particularly as this analysis focuses on the the PM peak on weekdays where less than 15% of trips are made by non-subscribers.

Table 4.1: Coefficient of determination matrix for PM Peak volume estimates on observed links.

	Observed Volume	SFCTA	BABS	Strava Metro	MTC
Observed Volume	1.000	0.348	0.148	0.508	0.011
SFCTA	0.348	1.000	0.081	0.045	0.130
BABS	0.148	0.081	1.000	0.009	0.000
Strava Metro	0.508	0.045	0.009	1.000	0.000
MTC	0.011	0.130	0.000	0.000	1.000

In addition to geospatially mapping the volumes predicted by each dataset, we can consider the pairwise coefficients of determination (R^2) between each of the datasets, and especially with the observed volumes, shown in Table 4.1. This tells us, if we were to simply linearly scale the estimates from one dataset (and add an estimated intercept term) how well we would predict the other dataset. In the case of comparing against the ground-truth volumes, this provides some idea as to how well the dataset matches reality.

We see that the most predictive datasets against the observed volumes are Strava Metro ($R^2 = 0.508$) and the SFCTA data ($R^2 = 0.347$). Interestingly, these datasets are not very predictive of each other ($R^2 = 0.045$), further supporting the hypothesis that they are representing different travel patterns, at least for the weekday PM peak period under consideration here.

Weighting Variables

In addition to the demand datasets discussed above, the local model specified here requires additional variables to inform the weighting matrix component based on link similarity, $w_{i,q}^{sim}$. Generally speaking, the features in $\boldsymbol{x}_{i,q}$ can be any categorical or continuous variables on which links i and q can be compared and are expected to be related to representation rates of the various datasets. For this case study, the following variables have been considered:

- Bicycle Facility (Categorical): Facility types considered here are "bike path", "bike lane", "bike route", and "None/Unmarked shared lane".
- Bearing (Continuous): Orientation of the link, based on relative position of link start and end points.
- Bikeshare Zone (Categorical): Link intersects with the convex hull of the bikeshare system stations.
- Street Type (Categorical): Road classification, according to OpenStreetMap scheme: "Primary", "Secondary", "Tertiary", "Residential", "Cycleway", "Path", and "Footway".

4 Results

The data fusion results section is broken into two main sections. First, tests were conducted to compare all possible subsets of datasets in terms of predictive accuracy, using Ordinary Least Squares, a Gaussian distance-decay weighting scheme, and a bisquare distance-decay weighting scheme. For the distance weighted models, searches are performed over a range of possible bandwidth values to find the optimal bandwidth for each subset of data sources. The best prediction is found using the SFCTA model, Strava Metro data, and Bay Area Bikeshare data, weighting observations with a 2500 ft. bandwidth Gaussian kernel. Second, results are shown for experiments performed on the optimal data sources using various additional weighting terms. No substantial improvements to prediction are found here. Finally, the residuals from the OLS and GWR models are compared to help understand why prediction is improved with the spatially-varying coefficients model.

Feature Selection

The feature selection results are shown in Table 4.2. The best model, on the basis of Cross-Validation, is a geographically-weighted model with a Gaussian kernel drawing

on the SF-CHAMP travel demand model, Bikeshare data, and Strava Metro data. This feature set has an optimal bandwidth of 2500 ft. In general, the Gaussian spatial kernel appears to fit slightly better than the bisquare, although the differences are not substantial. Interestingly, including the MTC travel demand model worsens fit. It is also notable that predictive accuracy only improves substantially by using the geographically weighted model when the feature set includes the crowdsourced data. This fits with the idea that the primary errors in the Strava Metro data are spatially continuous, having to do with variation in reporting rates between different regions of the city, whereas the errors in the other datasets might have more to do with problems in applying the route choice model, which does not seem to be corrected by considering nearby observations more heavily.

Table 4.2: Comparison of model predictive accuracy for global and local models using Leave One Label Out Cross-Validation for various combinations of data sources.

Variable Subset	$RMSD_{OLS}$	$RMSD_{Gaussian}$	$RMSD_{Bisquare}$
mtc	74.8	74.1	74.1
babs	70.8	66.7	66.9
sfcta	56.7	56.6	56.7
strava	56.7	34.7	35.4
babs, mtc	67.7	65.6	64.9
sfcta, mtc	56.7	56.6	56.7
sfcta, babs	54.6	54.9	55.3
strava, mtc	55.0	34.2	35.2
strava, babs	50.5	30.9	32.5
strava, sfcta	42.6	25.5	26.5
sfcta, babs, mtc	54.6	54.9	55.5
strava, babs, mtc	49.2	31.1	32.7
strava, sfcta, mtc	42.7	25.6	26.5
strava, sfcta, babs	40.0	24.3	24.8
strava, sfcta, babs, mtc	40.0	24.4	24.9

Additional Weighting Factors

In addition to experimenting with varying the set of features used in prediction, we explore a variety of additional weighting variables for the "link similarity" weights. This experimentation is conducted using the optimal feature set identified in the

strictly distance-weighted component. The results have been summarized in Table 4.3.

Table 4.3: Comparison of model predictive accuracy for various link similarity measures in a local model, using SF-CHAMP, Strava Metro, and BABS datasets.

Kernel	Weighting Link	Variables	Optimal Parameters	RMSD
Gaussian	Identity	BABS Hull	h: 2500	25.4
Gaussian	Identity	Bike Facility	h: 3000	28.0
Gaussian	Identity	Bearing	h: 2750	24.2
Gaussian	Identity	Highway Type	h: 3000	29.3
Gaussian	Logit	Bearing	h: 2328 α: -2.43 θ_{angle}: 1.67	24.0
Gaussian	Logit	Bearing Bike Facility	h: 2533 α: -2.75 θ_{angle}: 1.16 $\theta_{facility}$: 2.01	23.7
—	Logit	Highway Type Bike Facility	α: 0.17 θ_{hwy}: 7.32 $\theta_{facility}$: 4.27	39.7

First, we consider simple indicator weights in conjunction with the Gaussian distance-decay function. Using indicator weights with categorical variables has the effect that any observations on links that are in the same category as the link where the fit is being performed are included, and other sites are not. For example, weighting on the "bike facility" variable means that coefficients fit on links with bike lanes only draw upon observations on links with bike lanes. This supposes that the data weights $\boldsymbol{\beta}_i$ on link i should only be informed by observations on other links that share the specified characteristic with link i. As can be seen in Table 4.3, none of the tested variables substantially improve fit. Part of this could be that by applying this inclusion/exclusion weighting scheme effectively reduces the sample size for each fit location, as dissimilar observations are simply omitted (by applying a weight of 0).

Next, we tested additional weighting formulations that apply a "softer" weighting scheme by utilizing a logistic link function. For example, one such model considers the relative bearings of the two links on a range of 0 for anti-parallel to 1 for parallel orientations, as well as a separate distance-decay function. To help with interpreting

this result, consider some examples. Two links that are located at the same location in space but oriented in opposite directions (i.e. the two sides of a street) would contribute to each other as:

$$w_{ij} = e^{-\frac{1}{2}(0/2328)^2} \frac{exp(-2.43)}{1 + exp(-2.43)} = 0.08 \tag{4.4}$$

while two links 500 ft. from each other oriented in the same direction would contribute to each others' data weights as:

$$w_{ij} = e^{-\frac{1}{2}(\frac{500}{2328})^2} \frac{exp(-2.43 + 1.67)}{1 + exp(-2.43 + 1.67)} = 0.977 \times 0.319 = 0.312 \tag{4.5}$$

As can be seen in the table, the added complexity of these additional models does not seem to improve fit. There are a few possible explanations for this. First, we might not have identified the best weighting variables or weighting formulation to correct for small-scale spatial variation in the input data. Second, while extensive cleaning and checking was performed on the input datasets, it is possible that there are errors in the data that are resulting in a lower limit on the achievable accuracy of the model. That said, these results are of course specific to the datasets used in this study. In cases where substantial traffic is predicted by a source with a clear censoring issue (e.g. a more complete bikeshare network), we would expect that applying this weighting scheme would work better.

The volume estimates from the optimal geographically weighted model are presented in Figure 4.2. As expected, the underlying patterns of heavy usage along the waterfront, Market street, through Golden Gate Park, and throughout SoMa appear, although these distinct patterns were not well represented in the maps of individual dataset predictions. The volumes presented here are link totals (i.e. bidirectional), whereas the model predicted values are for each direction on the link separately.

5 Discussion

In this chapter, a novel method has been presented for estimating directional bicycle volumes on all links of a network for a single time interval based on a variety of demand data sources. The proposed method, utilizing Geographically Weighted Regression, attempts to overcome spatial differences in reporting rates amongst datasets by localizing parameters to a given link and weighting observations based on proximity. This method has been shown to improve estimation accuracy over using a global data weighting scheme.

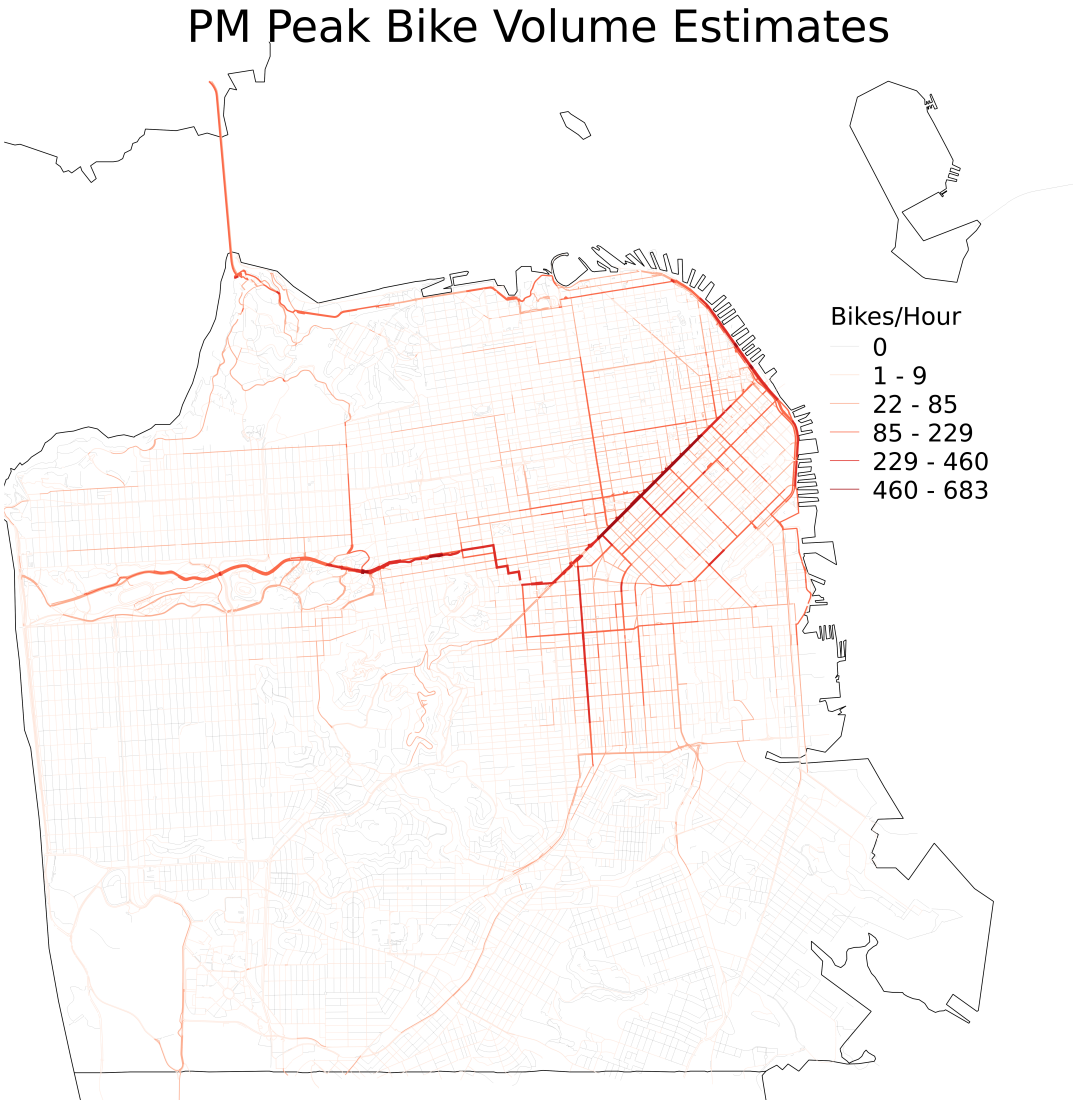


Figure 4.2: PM Peak bicycle volume estimates from Geographically Weighted Regression with a Gaussian kernel, 2800 foot bandwidth.

This work represents a major step forward in estimating bicyclist exposure. There have been multiple recent papers utilizing crowdsourced bicycle trip data to inform volume estimates. However, few of these have made use of ground-truth volumes to validate estimates, and none have overcome potential spatial biases in the data. By fusing together crowdsourced data, which appears to primarily represent recreational

travel in the case study presented, with utility-based travel demand model estimates and bikeshare data, we have presented a more complete view of directional bicycle volumes on the network. Furthermore, the methodology presented here is flexible to additional datasets that may become available in the near future, and provides a means for evaluating how these datasets fit into the overall representation of travel patterns.

There are ample opportunities to extend this work. First, within the context of the weighted model, we attempted multiple weighting schemes beyond a distance decay function, such as based on bicycle facility similarity between links. While none of these appeared to improve model fit in this case study, this could still be a useful way forward. Additional weighting schemes might include an estimate of the proportion of the trips that pass over the "fit location" and the "observation location" based on a trace dataset, or by similarities in surrounding land uses.

Second, we have only considered here spatial variation in the weighting scheme. However, given a greater availability of continuous count data, weights could be conceived to also vary temporally when considering panel data. In the case of crowdsourced data, for example, this might be interpreted as trip reporting rates being higher at certain times of the day. Similarly, it would be useful to consider separate models for weekdays and weekends, as the bicycle traffic activity patterns are likely substantially different on weekends.

Finally, it would be interesting to see how this work replicates in different cities or across a larger region. The improvements to predictive accuracy noted here by including crowdsourced data are due in part to an apparent deficiency in the inputs to the regional travel demand models. While this is a compelling argument for fusing these data sources together, it is expected that further improvements could be had in a region with greater variation in facilities, namely a higher preponderance of "primarily recreational" roads. Bicycle volumes on these roads due to recreational travel are virtually unpredictable within a utility-based model, but are an important part of evaluating risk factors.

Chapter 5

Temporal Extrapolation based on Signal Decomposition of Continuous Bicycle Volume Data

1 Introduction

In the previous chapter, the focus was on estimating bicycle traffic volumes during a single time interval given abundant directly observed counts. Here, the attention will turn to inferring the temporal distribution of traffic on a link, which the PM peak volume estimates developed in the previous chapter can then be used to scale upon yielding hour-by-hour traffic volume estimates across the entire week for all links on the network, as well as for standardizing the PM peak hour counts to an Annual Average Daily Bicyclists (AADB) representation. The variation in temporal patterns across the network, depending on both location and direction, means that "PM peak" estimates on different directional links are not directly comparable to one another in terms of total traffic volumes, and hence exposure.

The majority of the current state of knowledge on temporal extrapolation of non-motorized traffic volumes to generate estimates of Average Annual Daily Bicyclists is summarized in Chapter 4 of the *Traffic Monitoring Guide* (TMG) [Federal Highway Administration, 2013]. The TMG recommends collecting continuous counts on a randomly chosen subset of links to develop "extrapolation factors," or factors summarizing the proportion of traffic falling within a given subset of time. The temporal patterns of traffic volumes vary in space, and the TMG recommends classifying these patterns into "factor groups," or groups of sites with similar temporal patterns, and characterizing each group by the average temporal pattern.

Factor groups are typically established based on the hour-of-day patterns on weekdays, hour-of-day patterns on weekends, and day-of-week patterns. For example, based on these three sub-patterns Miranda-Moreno et al. [2013] identify four factor groups using induction loop data from multiple cities. These patterns are developed based on bidirectional traffic and are semantically labeled as "recreational", "mixed recreational", "mixed utilitarian," and "utilitarian." This method supposes that particular hour-of-day and day-of-week patterns are associated with each other within a given location, which might not necessarily be the case. Further, and perhaps most importantly, it does not readily provide a means of identification of the appropriate extrapolation factors for unobserved sites. This is a fundamental omission in the state of practice that has not been addressed in any systematic way - previous works have matched short-duration count sites to factor groups on the basis of shared facility types and/or common surrounding land uses, but there has been no validation or formal development of these methods [Roll, 2013, Hankey et al., 2014, Nordback et al., 2013].

As an alternative, in this chapter a decomposition method for link characterization is presented. The proposed method utilizes Latent Dirichlet Allocation (LDA) for signal decomposition. LDA, originally presented by Blei [2003], is a hierarchical Bayesian model developed for identifying latent topics in text documents. The LDA model is applied here under the assumption that there exist latent types of bicycle trips with distinct spatio-temporal patterns associated with them. That is, a given type of trip has a particular probability associated with being observed at a given time of week, and similarly has a varying probability of being observed on different directional links of the network. To be more concrete, an easy example is AM commuting activity which we might suppose has the highest probability of being observed between 7 and 9 AM on weekdays, and which is most likely to be observed on links oriented towards the central business district and located in close proximity.

Applying LDA as a signal decomposition tool here, we are able to identify a set of latent bicycle activity patterns. We then develop estimates for the proportion of overall traffic on observed links attributable to each trip type. These weights can then be used to estimate the overall distribution of traffic on observed links, and on unobserved links if the weights are estimated. Here we present a simple spatial/directional autoregressive model, which infers the mixture weights for the latent trip types on unobserved links based on the observed signal components on spatially and directionally adjacent links.

2 Data Sources

For this chapter, the primary data source is one week of automated counts collected between June 5 and June 12, 2016 on 75 directional links. Counts were collected using 74 bicycle-specific induction loop detectors. Similar to those used to detect and count motor vehicles, induction loop bicycle counters detect metal objects passing through their detection zone and classify the vehicle according to the currents that it induces in the loop.

The counters, including the type of facility on which they lie and their observed Average Daily Bicycle (ADB) volumes during the study period are shown in Table 5.1.

3 Methodology

We will now describe the methodology employed for inferring the temporal distribution of traffic volumes across space. The "temporal patterns" referred to here are specifically the proportion of traffic passing the link within each hour of the week relative to all traffic crossing the link in a week. We will denote this $f_{y,t}$, i.e. the factor for link y during time-slice t. As these are defined relative to the full week, by necessity

$$\sum_{t=0}^{167} f_{y,t} = 1 \; \forall \; y \tag{5.1}$$

This quantity is observed on a subset of links, but we desire an estimate on all links, so that "peak hour" volume estimates can be appropriately scaled to get volume estimates for an average day.

There are three main steps to the proposed process:

- Signal Decomposition: Automated count data from the observed links is decomposed using Latent Dirichlet Allocation. For each observed directional link, this results in a set of weights associated with each of the identified latent bicycle trip types, as well as a definition of what each of those trip types' temporal patterns looks like.

- Spatial Interpolation: In order to infer the trip type proportions on unobserved links, we use a fully spatio-directionally autoregressive fractional logit model.

[1]L: Bike Lane, BL: Buffered Bike Lane, P: Path, CT: Cycletrack, —: No Facility
[2]Average Daily Bicyclists

- Signal Reconstruction: Finally, a process is defined for reconstructing the overall temporal profile of trips on a given link based on the trip-type proportions associated with that link.

Signal Decomposition

In the first step of the proposed method, we seek to decompose the observed temporal bicyclist count patterns on each link into a lower dimensional representation. The Latent Dirichlet Allocation (LDA) model is employed. LDA is a Bayesian hierarchical model originally developed to identify latent topics in text corpora. In the original LDA model, it is supposed that we have a corpus of text documents. Each document is conceived as having a mixture of latent topics underlying it. For each word in each document, a topic is drawn from the document's topic mixture, and conditional on this topic a particular word is drawn. Both multinomial distribution parameters are given Dirichlet priors.

While LDA was originally conceived in the context of exchangeable text documents, it has also proven useful as a method of decomposing discrete time series data with multiple parallel signals into underlying components shared across the signals [Coffey and Pozdnoukhov, 2013]. In this application, rather than documents we consider time intervals (e.g. hours), which are comprised of a set of discrete observations at each of a set of locations. In this paper, these observations are bicyclists being counted on specific directional links of the network.

LDA was originally conceived as a generative model, meaning that it explicitly models the process by which the observations were generated (i.e. for each document, words are drawn from the underlying topic mixture). The analogous generative process for this application is:

1. For each time period t, draw $\Phi \sim \text{Dirichlet}(\alpha)$, the distribution of topics within the time period t.

2. For each observation of a bicyclist made in time period t:

   a) Draw a trip purpose $\phi_i \sim \text{Multinomial}(\Phi)$.

   b) Draw a link w_i from $\text{Multinomial}(w_i|\phi_i, \beta)$, which corresponds to the link where the observation occurred.

While this generative story can help elucidate the parallel to topic discovery in text, a probabilistic interpretation provides more intuition on the topic at hand. In this interpretation, for a given time period we have a certain probability of any observation arising from a set of trip purposes Z, denoted $p(z|t)$. For example,

during the 8:00 AM hour on a weekday, we might expect a high probability of a trip arising due to AM commuting. We would also expect that particular directional links of the network are more likely to have observations occur due to this latent AM commuting trip type (in particular, links oriented towards downtown would seem to be a reasonable guess).

Signal Reconstruction

This model as presented is useful for ascribing some meaning to the count observations taken on a set of fully (temporally) observed links. However, it is also useful as a dimensionality reduction tool. Now each observed link can be described in terms of the proportion of the observed bicyclists who are presumed to be a part of each of the latent trip types. The derivation of this probability will follow.

The output of the Latent Dirichlet Allocation gives us empirical estimates of:

- $p(l_i = y|z_i = k)$ – The conditional probability of link y being drawn for an observation from topic k.
- $p(z_i = k|t)$ – The probability of a given topic k for time period (document) t.
- $p(t)$ – The proportionate size of each time slice t, calculated as $\frac{N_t}{\sum_t N_t}$, where N_t is the total number of observations of cyclists in time period t.

From this, we could marginalize out the latent topics Z to get:

$$p(l_i|t) = \sum_k p(z_i = k|t)p(l_i = y|z_i = k)$$

We should still marginalize out the $p(t)$, yielding:

$$\begin{aligned} p(l_i) &= \sum_t \sum_k p(z_i|t)p(l_i|z_i)p(t) \\ &= \sum_k p(l_i|z_i) \sum_t p(z_i|t)p(t) \end{aligned}$$

The overall goal here is to get weights to associate with each topic k for each link y,$p(z_i = k|l_i = y)$, which by Bayes Theorem can be written as:

$$p(z_i|l_i) = \frac{p(l_i|z_i)p(z_i)}{p(l_i)} \tag{5.2}$$

While we don't exactly have $p(z_i)$, we can get it from $p(z_i) = \sum_t p(z_i|t)p(t)$.

So, overall, this gives us that:

$$\widehat{p(z_i|l_i)} = \frac{p(l_i|z_i)\sum_t p(z_i|t)p(t)}{\sum_k p(l_i|z_i)\sum_t p(z_i|t)p(t)} \tag{5.3}$$

Say that we want to construct the temporal profile for a link y given these weights. Let's say we have N^* total observation (i.e. in total across links and time periods). The temporal profile could be written as:

$$f_{y,t} = \frac{N_{y,t}}{\sum_t N_{y,t}} \tag{5.4}$$

where

$$N_{y,t} = p(l_i = y|t)p(t)N^*$$

bri so

$$\begin{aligned} f_{y,t} &= \frac{p(l_i = y|t)p(t)}{\sum_t p(l_i = y|t)p(t)} \\ &= \frac{\sum_k p(z_i|t)p(l_i|z_i)p(t)}{\sum_t \sum_k p(z_i|t)p(l_i|z_i)p(t)} \\ &= \frac{p(t)\sum_k p(z_i|t)p(l_i|z_i)}{\sum_t p(t)\sum_k p(z_i|t)p(l_i|z_i)} \\ &= \frac{p(t)\sum_k p(z_i|t)p(l_i|z_i)}{p(l_i)} \end{aligned} \tag{5.5}$$

The problem in general is that we do not know what $p(l_i|z_i)$ is. But, we can re-express this in terms of $p(z_i|l_i)$ as

$$p(l_i|z_i) = \frac{p(z_i|l_i)p(l_i)}{\sum_t p(t)p(z_i|t)}$$

Hence, we can write the temporal distribution $f_{y,t}$ as:

$$f_{y,t} = \frac{p(t)}{p(l_i)}\sum_k p(z_i|t)\frac{p(z_i|l_i)p(l_i)}{\sum_t p(t)p(z_i|t)} \tag{5.6}$$

$$= p(t)\sum_k \frac{p(z_i|t)\widehat{p(z_i|l_i)}}{\sum_t p(t)p(z_i|t)} \tag{5.7}$$

$$\tag{5.8}$$

This yields a way to reconstruct the temporal profile given topic probabilities for each link, but for all unobserved links we will not have $p(z_i|\widehat{l_i = y})$. Hence, we must infer.

Signal Interpolation

At this point, we have decomposed the bicycle counts on a set of observed links into a set of latent trip types, which are defined based on the spatio-temporal patterns underlying the data, and determined how to reconstruct the signals given a decomposed representation. The "decomposed representation" referred to here is the proportion of overall traffic on the link attributable to each latent trip type. Re-notating the conditional probabilities as $y_{ki} = P(z_i = k|l_i)$we can refer to the outcome variable matrix as $\boldsymbol{Y}$ which has dimensionality $L \times K$, where L is the number of links and K is the number of trip types. Our goal here is to estimate y_{ki}, which must obey two conditions:

- $\sum_k y_{ki} = 1 \ \forall \ i$
- $y_{ki} \geq 0 \ \forall \ k, i$

In other words, the vector $\vec{y_i}$ must lie on the standardized simplex. The prediction of a simplex outcome variable has been studied previously under the name "fractional regression." One approach to this problem is formulating the outcome vector $\boldsymbol{Y}$ as following a Dirichlet distribution, with the predictor variables linked to the concentration parameters of the Dirichlet [Woodland, 1979]. However, this proves problematic if any of the components of the outcome vector take a value of 0 as the log-likelihood tends towards $-\infty$, which is the case here with some links having nearly no weekend traffic component. As an alternative, Mullahy [2015] propose the use of the multinomial logistic distribution, specified with the fractional share as the outcome variable as opposed to the typical multinomial variable where the prediction is the probability of each discrete outcome occurring.

We will consider a similar formulation here, where we first define:

$$E[y_{ik}|\mathbf{Y}, \rho] = \xi_{ik} \in (0, 1), k = 1, ..., K \tag{5.9}$$

$$\sum_{m=1}^{K} \xi_{im} = 1 \forall i \in I \tag{5.10}$$

$$Pr(y_k = 0|\mathbf{Y}, \rho) \geq 0, k = 1, ..., K \tag{5.11}$$

$$Pr(y_k = 1|\mathbf{Y}, \rho) \geq 0, k = 1, ..., K \tag{5.12}$$

$$\tag{5.13}$$

As in Mullahy [2015], we'll first specify that

$$E[y_{ik}|\mathbf{Y}, \rho] = \xi_{ik} = \frac{e^{g_k(\mathbf{Y};\rho_k)}}{\sum_k e^{g_k(\mathbf{Y};\rho_k)}} \tag{5.14}$$

where $g_k(\bullet)$ relates the observation of y_{ik} on link i to observations $\boldsymbol{Y}$ on all other links J through the use of a spatial weighting matrix W. As in most spatial modeling contexts, the weighting matrix must be exogenously specified, which will be discussed shortly. We also have options to consider for the form of $g_k(\bullet)$. The most general option here would be to specify

$$g_k(\boldsymbol{Y}, \boldsymbol{\rho}) = \alpha_k + \sum_{m=1}^{K} \sum_{j=1}^{J} w_{ij} y_{jm} \rho_{mk} \tag{5.15}$$

where

$\alpha_k =$ Topic-specific constant for k. Must be fixed to 0 for one topic, e.g. $\alpha_K = 0$
$w_{ij} =$ spatial weight relating link j to link i
$y_{jm} =$ observation of topic m on link j
$\rho_{mk} =$ spatial autocorrelation between topic m and topic k

This is overly general, with potentially K^2 free parameters within $\boldsymbol{\rho}$. To make this tractable, we will start with some simplifying assumptions on $\boldsymbol{\rho}$. First, we might specify that $\rho_{mk} = \delta_{mk}\rho_k$; that is, assume that there are no cross-topic terms in the spatial autocorrelation. A further, more restrictive assumption would be that $\rho_{mk} = \delta_{mk}\rho$, implying that spatial auto-correlation is constant across topics.

Taking the first of these assumptions, our specification for expected topic shares would thus be:

$$\xi_{ik} = \frac{e^{\alpha_k + \sum_j w_{ij} y_{jk} \rho_k}}{\sum_m e^{\alpha_m + \sum_j w_{ij} y_{jm} \rho_m}} \tag{5.16}$$

Similar to in the Geographically Weighted Regression models presented in Chapter 4, a specification for the weighting matrix W_{ij} is required. For a first cut, we'll consider a joint Gaussian distance decay and angular deviation parameterization, where the element relating observation j to location i is:

$$w_{ij} = \begin{cases} e^{-(\frac{d_{ij}}{h})^2 \angle_{ij}^{\beta_{\text{angle}}}}, & \text{for } i \neq j \\ 0, & \text{for } i = j \end{cases}$$

where

$\angle_{ij} =$ is the normalized angular deviation between the oreintation of link i and link j
$\beta_{angle} =$ is a free parameter for the weight given to the angular deviation
$h =$ is the bandwidth for the Gaussian kernel
$d_i j =$ is the distance between link i and link j

Leave One Out cross-validation is utilized here to find the hyperparameters for the weighting matrix. Kullback-Leibler Divergence is considered for the loss-function, where fit is evaluated between the modeled and observed latent trip type shares on the held-out links.

4 Results

Applying LDA to the automated bicycle count dataset results in a set of latent travel patterns that can be attributed to underlying similarities in their spatio-temporal characteristics. The degree to which each of these latent temporal patterns reveals itself on a particular directional link of the network is presumably a function of a complex assortment of factors, most importantly land use patterns "upstream" and "downstream" of the count location. In this section, we present the observed patterns to reveal their temporal and spatial characteristics. We also hypothesize about what these observed patterns might correspond to in terms of bicyclist behavior, although fully understanding the underlying causes is not critical to the primary purpose of this chapter which is to infer the normalized temporal patterns on unobserved links of the network.

Decomposition on links with automated counts

The results of the LDA decomposition of automated counts can be visualized in various ways. First, we can consider the latent topic sizes within each time period, shown in Figure 5.1. Utilizing four topics, we uncover what appear to be an AM commuting pattern, a PM commuting pattern, a weekday pattern, and a midweek midday pattern. The "midday" pattern appears to also include some late night activity, which would fit with a hypothesized social cause of this activity.

In addition to visualizing the temporal distributions of the identified latent trip types, visualizing the spatial patterns of topic mixture weights can be informative. In Figure **??**, the relative share of counts on each observed link that has been probabilistically attributed to each of the latent topics is visualized. This geographic visualization lends further credence to the interpretation of the latent patterns, as we can see where in space and direction each trip type is most heavily represented.

For example, Figures **??**(a) and **??**(b) show the "AM Commute" and "PM Commute" patterns, the temporal profiles of which are shown in Figure 5.1. As can be seen, the travel pattern which we attribute to morning commuting (based on the times of day during which it is most prevalent) generally comprises the greatest share of overall travel on the links that are oriented towards downtown and located close to

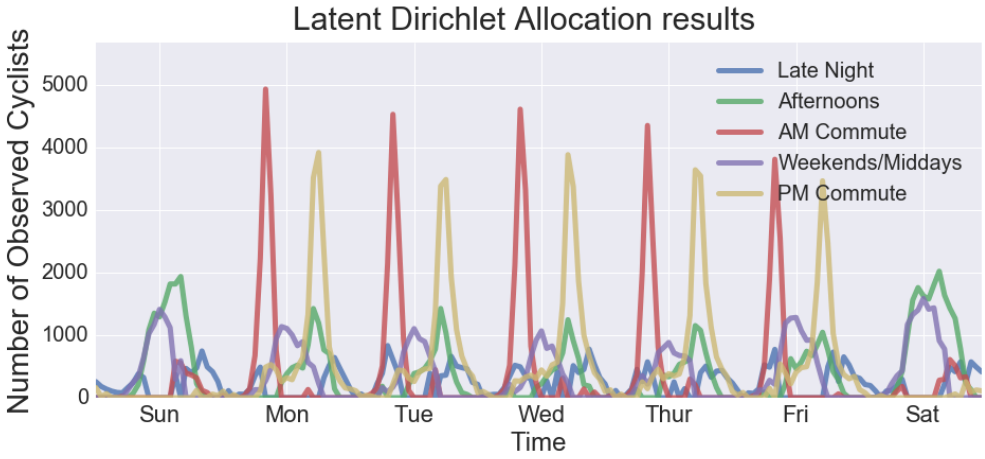


Figure 5.1: Latent topic sizes for each hour of the week for K=5 topics.

downtown. Similarly, we see the complementary evening commute patterns the most strongly on the same links in the opposing direction, out of downtown and particularly along the East - West direction. These patterns are very readily interpretable as arising due to weekday commuting behavior.

The other three topics identified here, termed "Midday", "Evenings", and "Late Night" are slightly less clear, but we can still hypothesize based on the spatiotemporal patterns. "Midday" occurs most prominently on weekends, with a bias towards the morning, as well as on weekdays during the midday. Spatially, this pattern has its most weight along the northern waterfront, as well in the western half of the city. This suggests some combination of recreational riding and lunchtime social activity. The "Evenings" pattern naturally has its strongest showing during the evening, particularly on weekends. This spatiotemporal pattern appears especially prominent on links near to and oriented away from the Golden Gate Bridge, suggesting that this may primarily be return trips from trips across the bridge, making it the complement to the "Midday" pattern similar to the duality of the AM/PM commute patterns. Finally, the "Late Night" activity pattern occurs most heavily late at night/in the early morning. This pattern does not comprise a particularly strong share of overall traffic on any particular link, which is not surprising given the overall relatively low amount of activity during these hours.

For all of the links with automated counters, we can consider how well the reconstruction of the signal performs (based on known topic shares) by plotting the reconstructions and the observed patterns as is shown in Figure 5.2. These plots depict the fraction of overall weekly bicycle traffic traversing the links within each hour. The volumes decomposed onto each latent topic are produced here by scal-

ing the overall topic-specific patterns (i.e. those in Figure 5.1) by the empirically determined conditional topic weights for each link $p(z_i|l_i)$. Importantly, the reconstructed overall signals, shown by the dotted grey lines, reasonably well reproduce the observed signals (solid grey lines).

Inference on spatial distribution of trip types

In addition to visualizing the latent patterns that appear to underly the time series signals from automated bicycle counters around the city, we also seek to infer the temporal distributions of traffic on unobserved links. As discussed in the methodology, this spatial inference is performed here using a purely autcorrelational model, where all variation in topic patterns is explained by observed variation in the same patterns on nearby links. The autocorrelation weighting matrix hyperparameters are first set using Leave One One cross-validation and a grid search across potential values, and the lowest average KL-Divergence is found for a bandwidth value of approximately $h = 10000ft.$ with a bearing parameter value of 1.75. The fact that the bearing parameter value is greater than one fits with expectations - this implies that links being closer to parallel than anti-parallel are asymetrically more similar to each other. In other words, links that "more or less" point in the same direction are more explanatory than those that are not, for equal deviations from perpendicular. With these hyperparameter values, the spatial autocorrelation coefficient has an estimated value of $\rho = 0.64$

5 Discussion

This chapter has presented a novel method for accounting for the spatiotemporal patterns underlying bicycle count data by using a signal decomposition method. These results have both theoretical and practical implications.

Theoretically, the identified latent trip patterns appear to correspond to particular types of travel. This method therefore can be used to provide some insight into the categories of travel underlying over bicycle traffic patterns, and by visualizing the latent topic proportions we can see what these dynamics look like across the city.

Practically, we have provided a new means of defining factor groups for purposes of extrapolating short-duration traffic volume estimates to daily averages, as well as for inferring traffic volumes across the week. While these factor group assignments have historically been performed in an ad hoc manner, by decomposing the reference count site observations into sub-patterns that are assumed to have some spatial and

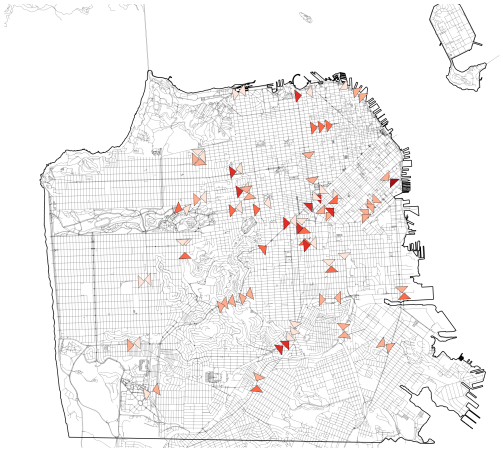
directional contiguity associated with them, we have shown a means of inferring patterns based on surrounding observed patterns, rather than rules of thumb.

There are opportunities for improvement at various stages of the proposed model. First, the signal decomposition model could be improved by accounting for dependencies in space and time. For example, rather than drawing a topic mixture independently for each time period with a Dirichlet prior, topic weights within a given time periods could be used to inform priors for the subsequent time period. This accounting for temporal dependence might help to smooth the decomposed signals shown in Figure 5.1, and further cement the interpretation of the latent topics as temporally types of travel. Similarly, correlation between links could be accounted for in the formulation of the conditional link probabilities.

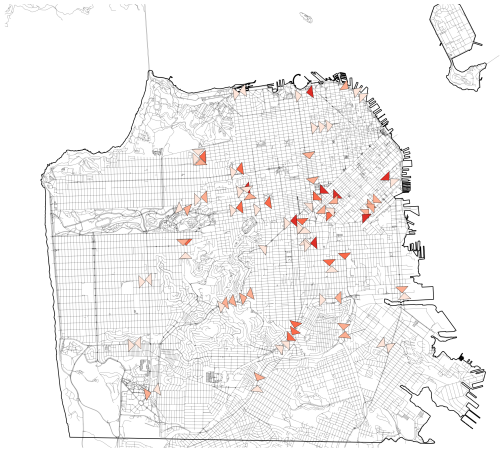
In addition, the spatial inference component as currently specified relies purely on spatial and directional autocorrelation to infer topic mixture components on each directional link. This could be improved upon by taking into account additional regressors, such as similarly identified temporal patterns within the other available bicycle demand datasets or land use variable as in a direct-demand model.

Table 5.1: Summary of counting stations, including observed traffic volumes.

Name	Facility[1]	ADB[2]	Name	Facility	ADB
NB 11th St. at Folsom	L	821	NB Illinois at Cesar Chavez	L	143
SB 11th St. at Folsom	L	590	SB Illinois at Cesar Chavez	L	148
EB 14th St. at Valencia	L	630	EB JFK at Conservatory	CT	1062
EB 17th St. at Mission	L	444	WB JFK at Conservatory	CT	1225
WB 17th St. at Mission	L	594	EB Lake at Arguello	L	85
NB 2nd St. at Townsend	—	224	WB Lake at Arguello	L	103
SB 2nd St. at Townsend	—	288	EB Marina at Baker	P	1206
SB 6th Ave. at Kirkham	L	58	WB Marina at Baker	P	2482
NB 7th Ave. at Kirkham	L	137	EB Market at 10th St.	CT	2565
SB 7th Ave. at Kirkham	L	121	WB Market at 10th St.	CT	2189
NB 7th St. at Townsend	L	543	EB Market at Castro	L	181
SB 8th St. and Folsom	BL	548	WB Market at Van Ness	CT	1842
NB Alemany at Ocean	L	73	WB Market at Van Ness	CT	1947
SB Alemany at Ocean	L	69	EB North Point at Polk	L	288
SB Arguello at Lake	L	251	WB North Point at Polk	L	417
NB Baker at Golden Gate	L	46	EB Oak at Scott	CT	1024
SB Baker at Golden Gate	L	38	EB Oakdale at Newhall	L	38
NB Bayshore at Cortland	BL	82	WB Oakdale at Newhall	L	45
SB Bayshore at Cortland	BL	81	EB Ortega at 20th Ave.	L	6
EB Cabrillo at Arguello	L	85	WB Ortega at 20th Ave.	L	9
WB Cabrillo at Arguello	L	69	EB Panhandle at Masonic	P	1544
EB Cesar Chavez at Harrison	L	641	WB Panhandle at Masonic	P	1964
WB Cesar Chavez at Harrison	L	1163	SB Polk at Grove	CT	671
EB Clipper at Grand View	L	11	SB Polk at Sutter	S	1000
EB Duboce Bikeway	P	540	EB Portola at O'Shaughnessy	BL	82
WB Duboce Bikeway	P	1343	WB Portola at O'Shaughnessy	L	63
NB Embarcadero at Bay	L	777	EB Post at Presidio	L	30
SB Embarcadero at Bay	L	1178	WB Post at Presidio	L	33
WB Fell St. at Scott St.	CT	1613	SB Potrero Avenue at 21st St.	L	95
EB Folsom at 8th	BL	1009	EB Robert C. Levy Tunnel	—	72
EB Golden Gate at Baker	L	116	NB San Jose at St. Mary's	BL	272
WB Golden Gate at Baker	L	75	SB San Jose at St. Mary's	BL	243
NB Harrison at 22nd St.	L	227	EB Sloat at 24th Ave.	BL	53
SB Harrison at 22nd St.	L	314	WB Sloat at 24th Ave.	BL	38
EB Holloway at Stratford	BL	33	NB Valencia at 14th St.	L	1213
WB Holloway at Stratford	BL	33	SB Valencia at 14th St.	L	887
WB Howard at 8th St.	BL	992	NB Valencia at 16th St.	L	1661

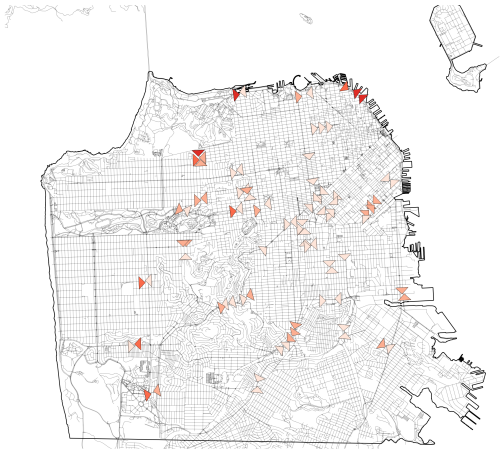


(a) AM Commute



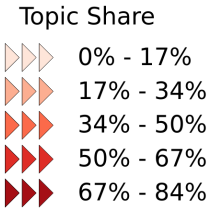
(b) PM Commute

(c) Midday



(d) Evenings

(e) Late Night



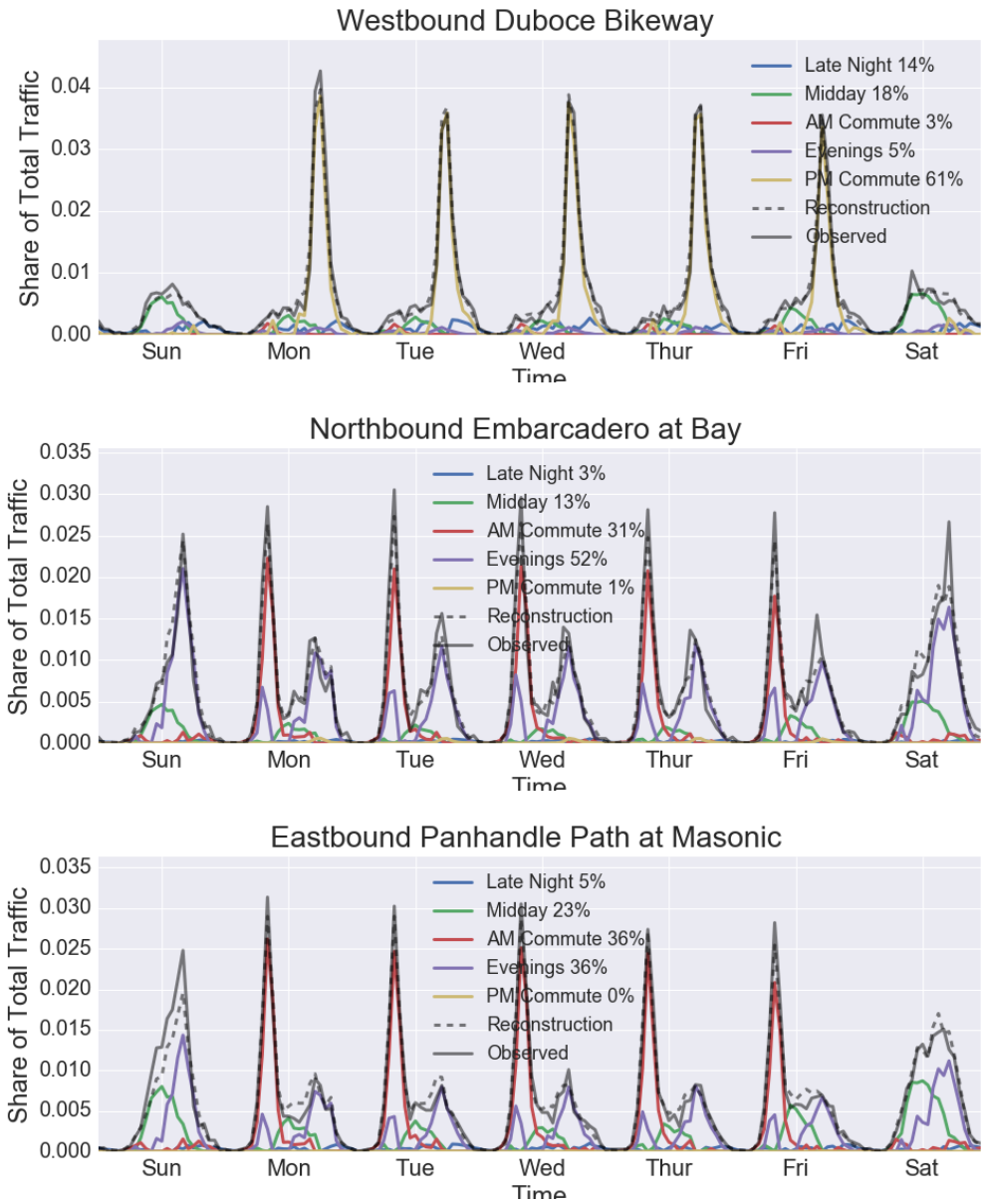


Figure 5.2: Reconstructed signals from 3 example counters.

Chapter 6

Conclusions

1 Contributions

This dissertation has pursued practical solutions to a long-standing problem in the field of traffic safety; namely, the lack of bicyclist exposure estimates (i.e. traffic volumes) across an entire street network. The increasingly widespread availability of high quality, spatially and temporally resolved data on bicycle demand has enabled the solutions proposed herein. In particular, here we have considered fusing traffic volume estimates from travel demand models, crowdsourced data, and bikeshare system data. Once reasonable peak-hour volume estimates are available, predicting the underlying temporal distribution of traffic at all locations is necessary for achieving annual daily traffic estimates.

The specific contributions of this dissertation are as follows:

1. Identified a "test case" location in San Francisco, CA with a relative abundance of bicycle demand data. This study site could easily be used for future similar studies.

2. Defined a metadata schema for use in classifying transportation demand data. The proposed schema is useful in defining how distinct datasets relate to one another.

3. Developed a data fusion method, based on Geographically Weighted Regression, that can be used to estimate peak period traffic volumes across an entire network. Improved predictive accuracy is shown for this model over a simpler "global" model, and over using any individual dataset independently.

4. Reconsidered the typical formulation of "factor groups" as is promulgated in the Traffic Monitoring Guide by applying Latent Dirichlet Allocation as a dimensionality reduction tool to permanent bicycle counter data. This novel formulation provides greater insight into the underlying mechanisms of temporal traffic patterns, and allows for more fine-tuned prediction of temporal patterns on a directional basis.

2 Future Work

This dissertation is just one step in using the ever-increasing abundance of data on bicycling to inform policy. Naturally there are ample opportunities for future work, which can be categorized as refinements upon the methods presented herein and logical next steps.

For the data fusion method presented in Chapter 4, the formulation of the weighting matrix could be revisited. Here, two distance-decay weighting schemes were experimented with, as well as some simple "similarity index" weights based on observations coming from locations with the same bicycle facility or road classification. More sophisticated similarity indices would be a worthwhile next step to see if additional refinements can be made in predictive accuracy. As more datasets become available, such as additional crowdsourcing streams, they can easily be incorporated into the proposed framework to improve estimates.

The temporal decomposition piece discussed in Chapter 5 also has room for improvement. At the signal decomposition stage, we have here assumed independence both between time periods and between network links. Naturally, we might expect that the probability of observed bicycle counts arising from a particular latent travel pattern might be correlated across time, and that links in close adjacency to one another would have similar correlation structures in terms of the their usage by trips under a common trip type. The refinement in the temporal dimension could be made by assuming a random walk process across time periods. The structure across directional links, however, is less intuitively clear, as this relationship should ideally be governed both by spatially proximity and directional similarity. One possibility might be to utilize GPS traces to inform this relationship. At the signal interpolation stage, at this point only autocorrelation has been used for prediction. Inclusion of additional regressors, such as land uses surrounding the count site or temporal patterns intimated by the additional demand datasets could potentially improve prediction of the topic shares.

Aside from these refinements to the model implementations shown here in San Francisco, application to regions with more variation in road types and functions

could be interesting. Specifically, the inclusion of more suburban and rural environs could exemplify the power of the data fusion approach taken here, as part of the underlying strength of this approach is presumed to be the lack of knowledge of recreational bicycle demand in regional travel demand models, and this type of riding is disproportionately prevalent on rural roads. Similarly, the temporal patterns present in this wider scope of places might benefit more from the temporal decomposition used here than at locations entirely within a fairly dense city, and provide more insight into the underlying predictors of these variations.

The natural next step of this work is to apply the exposure estimates developed herein to evaluating risks incurred by cyclists, especially due to crashes and air pollution. Because the exposure estimates developed here are both spatially and temporally resolved, novel analytical methods are enabled, including studying both the spatial factors affecting risk (such as roadway geometries and specifically bicycle facilities) and temporal factors such as variations in crash risk between night and day and variations in exposure to pollution across space and time.

While very promising results have been shown here in San Francisco, this study location was selected in part due to its abundance of available datasets. An additional future line of study would be to develop guidance for communities looking to apply these methods on how to allocate scarce data collections budgets, between conducting additional ground-truth counts, improving travel demand models, collecting or purchasing crowd-sourced data, and installing GPS transponders on bikeshare system bicycles. The relative value of these sources in terms of their information contributions towards exposure estimation could be evaluated.

Bibliography

M. Ben-Akiva, M.J. Bergman, A.J. Daly, and R. Ramaswamy. Modeling inter-urban route choice behavior. In *Proceedings of the 9th International Symposium on Transportation and Traffic Theory, VNU Press, Utrecht*, pages 299–330, 1984.

Chandra R. Bhat and Rachel Gossen. A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transportation Research Part B*, 38: 767–787, 2004.

Chandra R. Bhat and Frank S. Koppelman. *Handbook of Transportation Science*, chapter Activity-Based Modeling of Travel Demand, pages 39–65. Springer US, 2nd edition, 2003.

Andrew Y. Jordan Michael I. Blei, David M. Ng. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

Piet H.L. Bovy. On modelling route choice sets in transportation networks: A synthesis. *Transport Reviews*, 29(1):43–68, 2009.

Joseph Broach, Jennifer Dill, and John Gliebe. Where do cyclists ride? a route choice model developed with revealed preference gps data. *Transportation Research Part A: Policy and Practice*, 46, December 2012.

Edmund G. Brown Jr., Brian P. Kelly, Malcolm Dougherty, and Kome Ajise. Caltrans strategic management plan 2015-2020. Technical report, California Department of Transportation, 2015.

Cathal Coffey and Alexei Pozdnoukhov. Temporal decomposition and semantic enrichment of mobility flows. In *LBSN'13 at 21st ACM SIGSPATIAL GIS*, 2013.

Corey, Canapary & Galanis. Mode Share Survey 2011: Summary Report. Technical report, San Francisco Municipal Transportation Agency (SFMTA), 2011.

Mohamed El Esawey, Clark Lim, Tarek Sayed, and Ahmed Ibrahim Mosa. Development of daily adjustment factors for bicycle traffic. *Journal of Transportation Engineering*, 139(8):859–871, 2013.

Rune Elvik. Can a safety-in-numbers effect and a hazard-in-numbers effect coexist in the same data? *Accident Analysis and Prevention*, 60, 2013. doi: j.aap.2013.08.010.

Federal Highway Administration. *Traffic Monitoring Guide*. U.S. Department of Transportation, 2013.

A. Stewart Fotheringham, Chris Brunsdon, and Martin Charlton. *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*. Wiley, 2002.

Christopher Gallop, Cindy Tse, and Jinhua Zhao. A seasonal autoregressive model of vancouver bicycle traffic using weather variables. In *Transportation Research Board 91st Annual Meeting*, number 12-2119, 2012.

G. P. Griffin and J. Jiao. Crowdsourcing bicycle volumes: Exploring the role of volunteered geographic information and established monitoring methods. *URISA Journal*, 27, 2015.

Julia B. Griswold, Aditya Medury, and Robert J. Schneider. Pilot models for estimating bicycle intersection volumes. In *Transportation Research Board 90th Annual Meeting*, 2011.

Steve Hankey, Greg Lindsey, and Julian Marshall. Day-of-year scaling factors and design considerationsn for nonmotorized traffic monitoring programs. In *Transportation Research Board 93rd Annual Meeting Compendium of Papers*, 2014.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2009.

Ezra Hauer. Traffic conflicts and exposure. *Accident Analysis and Prevention*, 14: 359–364, 1982.

Matt Haynes and Seth Andrzejewski. Gis based bicycle & pedestrian demand forecasting techniques. Webinar, 2010. URL `http://media.tmiponline.org/webinars/2010/fehr_and_peers.pdf`.

Jeffrey Hood, Elizabeth Sall, and Billy Charlton. A gps-based bicycle route choice model for san francisco, california. *Transportation Letters: The International Journal of Transportation Research*, 2011. doi: 10.3328/TL.2011.03.01.63-75.

Ben Jestico, Trisalyn Nelson, and Meghan Winters. Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, 52:90–97, 2016.

Thomas Jonsson. *Predictive models for accidents on urban links-A focus on vulnerable road users*. PhD thesis, Lund University, 2005.

Lei Kang and Jon D. Fricker. Bicyclist commuters' choice of on-street versus off-street route segments. *Transportation*, 40:887–902, 2013.

J. Richard Kuzmyak, Jerry Walters, Mark Bradley, and Kara M. Kockelman. Estimating bicycling and walking for planning and project development: A guidebook. Technical Report NCHRP Report 770, National Cooperative Highway Research Program, Transportation Research Board of the National Academies, 2014.

Greg Lindsey, Yuling Han, Jeffrey Wilson, and Jihui Yang. Neighborhood correlates of urban trail use. *Journal of Physical Activity & Health*, 3:S139, 2006.

Dominique Lord and Fred Mannering. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5):291–305, 2010.

Michael G. McNally. *Handbook of Transport Modeling*, chapter The Four Step Model. Pergamon, 2008.

Luis F. Miranda-Moreno, Thomas Nosal, Robert J. Schneider, and Frank Proulx. Classification of bicycle traffic patterns in five north american cities. *Transportation Research Record: Journal of the Transportation Research Board*, 2013. doi: 10.3141/2339-08.

John Mullahy. Multivariate fractional regression estimation of econometric share models. *Journal of Econometric Methods*, 4(1):71–100, 2015.

Debbie A. Niemeier. Longitudinal analysis of bicycle count variability: Results and modeling implications. *Journal of Transportation Engineering*, 122:200–206, 1996.

Krista Nordback, Wesley E. Marshall, Bruce N. Janson, and Elizabeth Stolz. Estimating annual average daily bicyclists: Error and accuracy. *Transportation Research Record: Journal of the Transportation Research Board*, 2339, 2013.

Thomas Nosal and Luis F. Miranda-Moreno. The effect of weather on the use of north american bicycle facilities: A multi-city analysis using automatic counts. *Transportation Research Part A: Policy and Practice*, 66:213–225, 2014. doi: doi:10.1016/j.tra.2014.04.012.

Josh Frank Roll. Bicycle traffic count factoring: An examination of national, state and locally derived daily extrapolation factors. Master's thesis, 2013.

Paul Ryus, Erin Ferguson, Kelly M. Laustsen, Robert J. Schneider, Frank R. Proulx, Tony Hull, and Luis Miranda-Moreno. Guidebook on pedestrian and bicycle volume data collection. Technical Report NCHRP Report 797, National Cooperative Highway Research Program, Transportation Research Board of the National Academies, 2015.

Peter T. Savolainen, Fred L. Mannering, Dominique Lord, and Mohammed A. Quddus. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident Analysis & Prevention*, 43(5): 1666 – 1676, 2011.

Ipek N. Sener and Chandra R. Bhat. Modeling the spatial and temporal dimensions of recreational activity participation with a focus on physical activities. *Transportation*, 39:627–656, 2012.

Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

Jillian Strauss and Luis F. Miranda-Moreno. Spatial modeling of bicycle activity at signalized intersections. *The Journal of Transport and Land Use*, 6:47–58, 2013. doi: 10.5198/jtlu.v6i2.296.

Jillian Strauss, Luis F. Miranda-Moreno, and Patrick Morency. Cyclist activity and injury risk analysis at signalized intersections: A bayesian modelling approach. *Accident Analysis and Prevention*, 59:9–17, 2013. doi: 10.1016/j.aap.2013.04.037.

Strava, Inc., 2014. URL `http://metro.strava.com/`. "Accessed 6 June, 2015".

Sandar Tin Tin, Alistair Woodward, Elizabeth Robinson, and Shanthi Ameratunga. Temporal, seasonal and weather effects on cycle volume: an ecological study. #mar# 2012. doi: http://dx.doi.org/10.1186/1476-069X-11-12.

Shane Turner, Shane Binder, and Aaron Roozenburg. Cycle safety: Reducing the crash risk. Technical Report 389, NZ Transport Agency, October 2009.

U.S. Census Bureau. American Community Survey, Commuting Characteristics By Sex, 2009-2013 American Community Survey 5-Year Estimates, Table S0801. URL <http://factfinder2.census.gov>. generated by Frank Proulx, using American Fact Finder.

Kari Watkins, Rohit Ammanamanchi, Jeff LaMondia, and Christopher A. Le Dantec. Comparison of smartphone-based cyclist gps data sources. In *TRB 95th Annual Meeting Compendium of Papers*, number 16-5309, 2016.

Meghan Winters, Kay Teschke, Michael Grant, Eleanor M. Setton, and Michael Brauer. How far out of the way will we travel? built environment influence on route selection for bicycle and car travel. *Transportation Research Record: Journal of the Transportation Research Board*, 2190:1 – 10, 2010.

A. Woodland. Stochastic specification and the estimation of share equations. *Journal of Econometrics*, 10(3):361–383, 1979.