Bicycle Traffic Volume Estimation using Geographically Weighted Data Fusion

Frank R. Proulx\textsuperscript{a,b,*}, Alexei Pozdnukhov\textsuperscript{b}

\textsuperscript{a}Alta Planning + Design, 233 A Street Suite 703, San Diego, California, 92101
\textsuperscript{b}Department of Civil & Environmental Engineering, University of California, Berkeley, McLaughlin Hall, Berkeley, California, 94720

Abstract

In this paper we present a method for estimating bicycle traffic volumes on all links of a network by fusing various demand datasets. While the proposed method is agnostic to the specific datasets used, a case study of San Francisco, CA is presented using two regional travel demand models, crowdsourced data from Strava Metro, and usage data from Bay Area Bikeshare. These datasets are hypothesized to represent different types of travel, each of which is a component of the overall traffic mix. By fusing them, we can achieve a more comprehensive view of existing travel conditions which accounts for various types of bicycle travel, and which takes observed traffic volumes directly into account.

First, we present a metadata schema for describing demand datasets, which is intended to reveal the similarities and dissimilarities between datasets, as well as the steps that must be taken to homogenize the datasets. Once a common format is achieved, demand estimates are then fit to observed counts using Geographically Weighted Data Fusion (GWDF), a technique based on Geographically Weighted Regression (GWR). In addition to the traditional inverse-distance weights used in GWR, additional weighting schemes are also tested. It is shown through cross-validation that the proposed GWDF method outperforms any individual dataset’s prediction of volumes, provides a precise representation of traffic volumes across the network, and opens the door for future analyses which require accurate exposure metrics such as the quantification of crash risk.

Keywords: Bicycles, Traffic Volume, Data Fusion, Geographically Weighted Regression

1. Introduction

Cities, states, and countries around the world are increasingly turning to encouraging bicycling as a low-cost, space and energy efficient form of transportation and as a means of promoting public health through increased activity. Bicycle traffic volumes are a useful measure for monitoring progress on achieving increased rates of bicycling, informing decision-making on infrastructure decisions intended to promote bicycling, and perhaps most importantly as an input to bicycle crash risk models [12]. While historically achieving accurate estimates of bicycle traffic volumes has been difficult due to a lack of data, we are currently seeing a rapid expansion in the availability of bicycle demand data [14, 13]. As a non-exhaustive list, novel datasets include automated counters, crowdsourced GPS traces, and bikeshare usage data, in addition to the traditional manual counts and travel surveys.

For current and past conditions, bicycle volumes can be either directly observed or can be modeled. However, direct observation is practically limited to a small subset of links/intersections of the network due to the cost of data acquisition. In order to determine network-wide conditions, modeled volumes must be used.

There are two predominant approaches to modeling non-motorized traffic volumes: random utility models of individuals’ transportation and activity decisions and facility-based “direct-demand” models of observed volumes [9]. In random utility models, traveler decisions to be modeled include the decision to engage in
out-of-home activities, where to travel for these activities, the travel mode to take, and the route to take to reach the destination. The last stage of this process, trip assignment, is a necessity for determining traffic volumes, and has been demonstrated in Zorn et al. [18] using a bicycle specific route choice model estimated by Hood et al. [5]. Random utility models have a strong theoretical basis in decision theory, but are subject to errors at all stages of the modeling process and do not accommodate recreational trips well. Direct-demand models are formulated by regressing observed volumes on characteristics of the surroundings of the count site, such as land use and transportation network features [8, 4, 16, 10]. While this approach is appealing in its’ simplicity and the fact that it directly accounts for observed volumes, it is lacking in a strong behavioral foundation. For example, it is not clear how much of the surrounding environment should be included in the model. Bicycle trips might cross a link of the network but not have any relationship to its’ immediate surroundings. In addition to these two modeling techniques, there is a growing interest in “crowdsourced” demand data, where individuals report their bicycle trips to a transportation planning agency or to a private company, who then anonymizes their data for planning purposes [3, 6].

In this paper, we draw on the strengths of the various modeling approaches and novel sources of demand data to improve the estimation of network-wide directional link-level bicycle volumes. We first present a metadata schema intended to help understand how various datasets relate to one another in Section 2. Then, in Section 3 we discuss the data fusion approach and the specific datasets to be used in our case study. The approach suggested here is termed Geographically Weighted Data Fusion (GWDF), based on Geographically Weighted Regression [1]. Finally, in Section 4 we present the case study results and discussion, and in Section 6 we discuss our conclusions and opportunities for future work.

2. Metadata Schema

While travel demand data is available on an increasingly large scale, inferring demand patterns is an active area of inquiry in part due to the limitations of these datasets. No known dataset has full resolution into the spatial and temporal dimensions of the entire population’s travel patterns. However, they all describe these various dimensions to some degree. For example, permanent automated counters describe all peoples’ activity for all time but for only a single spatial location. Crowdsourced data, on the other hand, describes a subset of trips made by a subset of people with rich spatial and temporal detail. To further explore this idea, we propose the following metadata schema as a means of classifying demand datasets:

- Population Scope: Does the dataset describe the full population of people or a subset?
- Trip aggregation: Are individual trip records available, or are trips aggregated?
- Temporal scale: What is the temporal extent of the data, i.e. is it a full historical time-series, a partial time-series, or an average or “typical” estimate?
- Temporal resolution: How fine of slices of time are used to describe trip patterns? This can range from individual timestamps to annual totals.
- Spatial scale: How is the spatial nature of travel handled? The four categories identified here, to be discussed in greater detail below, are site-based, traces, origin-destination points, and origin-destination zones.
- Demographics: Any categorical descriptors of the trip or trip-maker(s), such as trip purpose, travel mode, age, gender, or helmet usage.

With these six dimensions, we can classify datasets in order to identify heterogeneities. While most of these descriptors should be straightforward, the spatial scale bears further discussion. “Site-based” datasets include information about trips passing a specific set of locations in space, typically along network links or through intersections. Traces are records of observed routes, namely those collected with GPS transponders or GPS-enabled smartphones. Origin-destination data only includes information on where trips begin and end, either as specific points or by Travel-Analysis Zone, but no details of the route taken.
3. Data and Methods

We will now discuss the two steps of the proposed data fusion method, and following that present the data used in the case study.

3.1. Fusion Model

In the first step, the demand datasets are homogenized to a similar representation in terms of their metadata dimensions. Once a common representation is achieved, the datasets are fused together by fitting to ground-truth counts.

3.1.1. Homogenization

The first step in fusing the 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 datasets that are intended to represent the full population 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. However, processes can be defined for conversion between any pair of representations. We will not exhaustively discuss these, but will highlight some to illustrate.

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. O-D pair, time of day, and mode). Construction of disaggregate trips from an aggregate representation requires simulation based on the details available. For example, if our travel demand model suggests that 20 bicycle trips have taken place for a given O-D pair between 3 and 6 P.M., we might assume a distribution of departure times peaking around 5 P.M. and draw departure times for the synthesized trips from this distribution.

The possible representations in temporal scale are a full time-series, a partial time-series, or an average/“typical” time-series. Conversion from a full time-series to a partial time-series simply entails truncating to the relevant interval, while simulating a full time-series based on a partial time-series (e.g. a short-duration manual count) requires making assumptions about how the partial time-series extends across time.

The spatial scale for all data sources is converted to a site-level representation, or in other words all are expressed in terms of the traffic volumes that they imply.

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.

3.1.2. Fitting

Once a common, site-level volume format is achieved for all datasets, they are fit to observed counts using a weighted Poisson-identity maximum likelihood model. The units of analysis here are directional links of the network. Overall bicycle traffic volumes are expressed as a linear combination of the volume estimates implied by each of the various datasets. The assumption underlying this formulation is that each of the datasets presents different travel patterns. For example, travel demand models are primarily focused on utilitarian travel such as commute trips, while user-focused crowdsourced data might provide a better representation of recreational patterns. Because travel demand models are intended to represent the full population of trips, we would hope that the corresponding weight would be on the order of 1, suggesting that each cyclist represented by the model corresponds to one “real-world” cyclist. For sub-population estimates such as crowdsourced data, the apparent travel patterns might be representative of trips made by a greater number of cyclists, in which case we would expect a parameter value greater than 1. On the other hand, if multiple full-population datasets are included, we would anticipate that the estimated parameters
would be less than 1 to compensate - that is, each of their “real-world cyclists” would only count for a portion of the overall observed total.

We restrict the parameters in our model to be non-negative to encourage interpretation as a fusion of demand estimates. A negative weight on a dataset would mean subtracting the cyclists that it represents from the overall traffic flow, where in actuality each dataset should either add cyclists to the modeled traffic mix for the patterns it is representing, or contribute no additional information \((\beta = 0)\).

The coefficients associated with the demand datasets are free to vary in space, and in this case we consider that their value could vary by link of the network. Variation in the parameters is induced by weighting observations according to an exogenously determined weighting matrix, \(W\), which defines weights for all observed links \(q\) for parameter estimates made on links \(i\). The weights here are flexibly defined, but we will primarily focus on weights that decay with distance, a technique known as Geographically Weighted Data Fusion (GWDF). In general, the GWDF model is specified (following after Fotheringham et al. [1] and Nakaya et al. [11]) as a locally evaluated conditional Poisson distribution with a linear link function:

\[
\begin{align*}
\lambda_i &= \beta_i v_i \\
p(y_i|\hat{\beta}_i, v_i) &= \frac{\lambda_i^{y_i}}{y_i!} e^{-\lambda_i}
\end{align*}
\]

where

- \(\beta_i = \) vector of parameters estimated at location \(i\)
- \(v_q = \) vector of directional link-volume estimates from set of datasets for link \(q\)
- \(y_q = \) observed volume on link \(q\)
- \(w_{iq} = \) weighting factor for regression point \(i\) and observation \(q\)

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. The resulting model can be interpreted as for every directional link \(i\), the estimated link volume is

\[
v_i^* = \beta_i v_i
\]

where again \(v_i\) denotes the vector of link-volume estimates implied by each of the datasets. \(\beta_i\) can thus be interpreted as a vector of coefficients by which each cyclist in the various datasets is weighted in terms of the overall traffic flow on link \(i\). The non-negativity constraint is used to ensure that bicyclists predicted by a given source aren’t being subtracted from the total, and the minimum link-flow that will be predicted is 0.

Based on this formulation, and following after Nakaya et al., we consider localized maximum likelihood estimation, where the “local” here is in geographic space, not attribute space, as elaborated upon by Nakaya et al [11]. The log-likelihood function at location \(i\) to be considered is

\[
L_i = \sum_q N \ln \lambda_q - \lambda_q w_{iq}
\]

\[
= \sum_q (y_q \ln (\hat{\beta}_i v_q) - \hat{\beta}_i v_q) w_{iq}
\]

We constrain our estimation to have positive values for \(\hat{\beta}_i\) at all locations, so the problem to be solved is:

\[
\begin{align*}
& \text{maximize} & & L_i \\
& \text{s.t.} & & \hat{\beta}_i \geq 0
\end{align*}
\]

In order to accommodate the non-negativity constraint, we reparameterize our equation by taking the element-wise log-transform of the \(\beta\) parameters as \(\theta_k = \ln (\beta_k)\). This transformation prevents the estimated parameters from being negative.
values from becoming negative-valued, and yields a final program to be solved of:

$$\text{maximize} \quad \sum_{q}^{N} (y_q \ln (\hat{\beta}_i \mathbf{v}_q) - \hat{\beta}_i \mathbf{v}_q) w_{iq}$$  \hspace{1cm} (6)

We will now suppress the $i$ index, which is implicit in the fitting of this objective separately at all locations. This gives us first-order conditions of:

$$\frac{\partial L}{\partial \theta_k} = \beta_k \sum_{q}^{N} \frac{y_q - 1}{\beta \mathbf{v}_q} v_{kq} w_q = 0 \quad \forall \ k$$  \hspace{1cm} (7)

And a Hessian calculated as:

$$\frac{\partial^2 L}{\partial \theta_k \theta_j} = \beta_k \beta_j \sum_{q}^{N} \frac{y_q v_{jq} v_{kq} w_q}{(\beta \mathbf{v}_q)^2}$$  \hspace{1cm} (8)

We maximize this quantity using the BFGS algorithm as implemented in SciPy Jones et al. [7].

Spatially varying coefficients are used to suggest that the rates by which the predicted volumes from each dataset $k$ are weighted could vary by location in a systematic way. This spatial non-stationarity could arise for a number of reasons, including:

- 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).

- 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. [17]).

There are various options available for parameterizing the weighting term $w_{iq}$, and we have only explored a small subset of the possibilities here. In general, we desire that the weights to be used be bounded on the range $[0, 1]$, and provide some means of relating location $i$ to location $q$. Some broad classes of relations include:

- $w_{iq} = 1$ – The traditional “global” regression model (i.e. coefficients do not vary in space - all observations contribute equally at all locations).

- $w_{iq} = \{(d_{iq})$ – Weights are some function of the distance from $i$ to $q$. This is similar to the traditional Geographically Weighted Regression.

- $w_{iq} = S_{iq}$ – Weights are based on some measure of “link similarity” between link $i$ and $q$, such as a binary score for whether they have the same type of bicycle facility present.

In the GWDF specification, a function relating distance between observations to the weighting value is typically referred to as the “kernel function.” Here we experiment with two specific kernels, namely:

- **Gaussian**
  \[ w_{iq} = \exp\left( -\frac{1}{2} \left( \frac{d_{iq}}{h} \right)^2 \right) \]

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

where in both cases the bandwidth $h$ is a model hyperparameter that affects the attenuation rate of the distance decay function.

In the link-similarity specification, we conceive that multiple factors $x_{iq}$ relating link $i$ to $q$ could be combined in a variety of ways, so long as the resulting weight is normalized on $[0, 1]$. Again, two possibilities were considered here:
• Product: If all features $P$ of $x_{iq}$ are constrained to the range $[0,1)$ or are binary, $w_{i,q} = \prod_{p \in P} x_{i,q,p}$

• Logistic: $w_{i,q} = \frac{\exp(x_i^\prime \theta)}{\exp(x_i^\prime \theta) + 1}$

where $\theta$ is a vector of model hyperparameters.

Regardless of the specific weighting scheme used, the model hyperparameters are set using a randomized search and cross-validating with Leave One Label Out (LOLO). LOLO entails iterating through the dataset, leaving out all observations sharing a common label on each pass for testing, and training on the remainder of the data, where in this case the labels are road network intersections. LOLO was used here to avoid simultaneously training and testing on observed volumes taken from turning movement counts at the same intersection, as these observations are necessarily highly correlated.

3.2. 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 weekday 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 [15]. 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.

3.2.1. Demand Datasets

The datasets used for this study are summarized in terms of their metadata characteristics in Table 1. The ground-truth bicycle counts are from a combination of 69 manual PM peak-period turning-movement counts and 25 automated counters. As directional links are the unit of analysis, the turning-movement counts were aggregated into link-flow totals for each of the approaches to the intersections. After removal of apparently miscoded counts, this resulted in a dataset of 536 observed directional links. For these links, we consider the average hourly volume during the PM peak as our outcome variable of interest.

“SF-CHAMP” and “MTC” refer to two full-population travel demand models. SF-CHAMP is managed by the San Francisco County Transportation Authority and considers a fine-grained spatial scale in the downtown San Francisco region, as the city of San Francisco is the focus of the model. Estimates from this model were available as bicycle origin-destination trip-tables for 5 aggregate time bins on a “typical weekday”. The MTC model is managed by the Metropolitan Transportation Commission, and has a regional planning focus.

“BABS” is historical demand data from the Bay Area Bikeshare system. BABS was launched in 2013 with San Francisco stations clustered along the Embarcadero waterfront, Market Street, and the South-of-Market neighborhood. All trips made on the system are recorded as timestamped observations, indicating the origin and destination stations.

Strava is a smartphone application that allows users to record their bicycle rides, runs, and other fitness activities via their phone’s GPS. These trips are then processed for the user to analyze their own activity patterns, such as frequently used routes and performance trends. User data is also anonymized by matching trips to the network and aggregating these observations by each minute of the year, and is then available for purchase for transportation planning purposes under the brand name Strava Metro. This crowdsourced data provides a rich account of the spatial and temporal patterns of Strava users, but there is some suspicion amongst researchers about the applicability of these patterns to the general population [2].

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. [5]. The resulting PM peak volume estimates from each dataset are shown in Figure 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 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. [6].

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

<table>
<thead>
<tr>
<th></th>
<th>Observed Volume</th>
<th>SFCTA</th>
<th>BABS</th>
<th>Strava Metro</th>
<th>MTC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed Volume</td>
<td>1.000</td>
<td>0.348</td>
<td>0.148</td>
<td>0.508</td>
<td>0.011</td>
</tr>
<tr>
<td>SFCTA</td>
<td>0.348</td>
<td>1.000</td>
<td>0.081</td>
<td>0.045</td>
<td>0.130</td>
</tr>
<tr>
<td>BABS</td>
<td>0.148</td>
<td>0.081</td>
<td>1.000</td>
<td>0.009</td>
<td>0.000</td>
</tr>
<tr>
<td>Strava Metro</td>
<td>0.508</td>
<td>0.045</td>
<td>0.009</td>
<td>1.000</td>
<td>0.000</td>
</tr>
<tr>
<td>MTC</td>
<td>0.011</td>
<td>0.130</td>
<td>0.000</td>
<td>0.000</td>
<td>1.000</td>
</tr>
</tbody>
</table>

In addition to geospatially mapping the volumes predicted by each dataset, we can consider the coefficient of determination matrix of volume estimates on the observed links, as presented in Table 2. 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
3.2.2. 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, \( w_{i,q}^{\text{sim}} \). Generally speaking, the features in \( x_{i,q} \) 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”,

Figure 1: September 2014 PM Peak volume estimates for each dataset.

peak period under consideration here.
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

For both the global and the local models, we compare model predictive accuracy according to Root Mean Square Deviation under Leave One Label Out Cross-Validation for all possible combinations of data sources. These results are shown in Table 3. The results reported for the local model with the Gaussian and Bisquare kernel correspond to the optimal bandwidth value. The best model, on the basis of Cross-Validation, is the local 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. Interestingly, including the MTC travel demand model worsens fit. It is also notable that predictive accuracy only improves substantially by using the local model when the feature set includes the crowdsourced data.

Table 3: Comparison of model predictive accuracy for global and local models using Leave One Label Out Cross-Validation for various combinations of data sources.

<table>
<thead>
<tr>
<th>Variable Subset</th>
<th>RMSD&lt;sub&gt;OLS&lt;/sub&gt;</th>
<th>RMSD&lt;sub&gt;Gaussian&lt;/sub&gt;</th>
<th>RMSD&lt;sub&gt;Bisquare&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>mtc</td>
<td>74.8</td>
<td>74.1</td>
<td>74.1</td>
</tr>
<tr>
<td>babs</td>
<td>70.8</td>
<td>66.7</td>
<td>66.9</td>
</tr>
<tr>
<td>sfcta</td>
<td>56.7</td>
<td>56.6</td>
<td>56.7</td>
</tr>
<tr>
<td>strava</td>
<td>56.7</td>
<td>34.7</td>
<td>35.4</td>
</tr>
<tr>
<td>babs, mtc</td>
<td>67.7</td>
<td>65.6</td>
<td>64.9</td>
</tr>
<tr>
<td>sfcta, mtc</td>
<td>56.7</td>
<td>56.6</td>
<td>56.7</td>
</tr>
<tr>
<td>sfcta, babs</td>
<td>54.6</td>
<td>54.9</td>
<td>55.3</td>
</tr>
<tr>
<td>strava, mtc</td>
<td>55.0</td>
<td>34.2</td>
<td>35.2</td>
</tr>
<tr>
<td>strava, babs</td>
<td>50.5</td>
<td>30.9</td>
<td>32.5</td>
</tr>
<tr>
<td>strava, sfcta</td>
<td>42.6</td>
<td>25.5</td>
<td>26.5</td>
</tr>
<tr>
<td>sfcta, babs, mtc</td>
<td>54.6</td>
<td>54.9</td>
<td>55.5</td>
</tr>
<tr>
<td>strava, babs, mtc</td>
<td>49.2</td>
<td>31.1</td>
<td>32.7</td>
</tr>
<tr>
<td>strava, sfcta, mtc</td>
<td>42.7</td>
<td>25.6</td>
<td>26.5</td>
</tr>
<tr>
<td>strava, sfcta, babs</td>
<td>40.0</td>
<td>24.3</td>
<td>24.8</td>
</tr>
<tr>
<td>strava, sfcta, babs, mtc</td>
<td>40.0</td>
<td>24.4</td>
<td>24.9</td>
</tr>
</tbody>
</table>

In addition to experimenting with varying the set of datasets used in prediction, we explore a variety of additional weighting variables for link similarity for the optimal combination of datasets, summarized in Table 4.

Figure 2 presents the pairwise distribution of directional traffic volume estimates on observed links between each pair of datasets. The “results” values depicted here are for the optimal model identified above, and are presented using held-out values to avoid overfitting. The comparison of the results against the observed volumes, highlighted in red, shows a very linear relationship as would be expected. We can compare this with the scatterplots of all of the other datasets against the observed volumes, all of which demonstrate relatively low correlation. This further confirms that argument that while relying on any of these datasets independently to predict traffic volumes is inadequate, fusing them can achieve a more accurate depiction of reality.

The volume estimates from the optimal local model are presented in Figure 3. As expected, the underlying patterns of heavy usage along the waterfront, Market street, through Golden Gate Park, and throughout the South of Market neighborhood 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.
Table 4: Comparison of model predictive accuracy for various link similarity measures in a local model, using SF-CHAMP, Strava Metro, and BABS datasets.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Weighting Link Variables</th>
<th>Optimal Parameters</th>
<th>RMSD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gaussian Identity BABS Hull</td>
<td>$h$: 2500</td>
<td>25.4</td>
<td></td>
</tr>
<tr>
<td>Gaussian Identity Bike Facility</td>
<td>$h$: 3000</td>
<td>28.0</td>
<td></td>
</tr>
<tr>
<td>Gaussian Identity Bearing</td>
<td>$h$: 2750</td>
<td>24.2</td>
<td></td>
</tr>
<tr>
<td>Gaussian Identity Highway Type</td>
<td>$h$: 3000</td>
<td>29.3</td>
<td></td>
</tr>
<tr>
<td>Gaussian Logit Bearing</td>
<td>$h$: 2328 $\alpha$: -2.43 $\theta_{\text{angle}}$: 1.67</td>
<td>24.0</td>
<td></td>
</tr>
<tr>
<td>— Logit Bike Facility</td>
<td>$h$: 2533 $\alpha$: -2.75 $\theta_{\text{angle}}$: 1.16 $\theta_{\text{facility}}$: 2.01</td>
<td>23.7</td>
<td></td>
</tr>
<tr>
<td>— Logit Highway Type Bike Facility</td>
<td>$\theta_{\text{hwy}}$: 7.32 $\theta_{\text{facility}}$: 4.27</td>
<td>39.7</td>
<td></td>
</tr>
</tbody>
</table>

5. Discussion

In this paper, 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 Data Fusion, 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.

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 local 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. 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.

Third, if an excessive number of zero-valued volumes were observed in the data conditioned on $v_i$, a more flexible functional form than the Poisson distribution could be used for the error terms. Popular alternative options include the negative binomial form and zero-inflated Poisson.

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.

6. Conclusions

This paper has presented a method for estimating link-level bicycle flows across a network by fusing together demand data from a variety of novel sources. This fusion method attempts to account for spatial biases in the datasets through the use of Geographically Weighted Data Fusion, wherein model parameters are free to vary in space, and observations are weighted at a given location based on proximity and similarity to the estimation location. By utilizing this weighting approach, we have demonstrated improved predictive accuracy over both all of the separate datasets on their own, and over using a simple linear combination of the demand dataset estimates with globally determined weights.
The research presented in this paper opens the door to ample future opportunities. The Geographically Weighted Data Fusion method here could readily be applied to other modes of transportation and datasets, as available. For example, in estimating motorized traffic one might envision pairing travel demand model estimates with third party datasets, either of overall traffic volume estimates, or of specific population subsets such as fleet data from taxi or freight companies.

7. Acknowledgements

This work was funded in part by grants from the Dwight D. Eisenhower Transportation Fellowship program and the University of California Center on Economic Competitiveness in Transportation (UC-CONNECT). Neither funding agency directed the research. The authors thanks Brian Riordan at Strava,
Inc. for providing access to the Strava Metro dataset, Siyu Chen for helping with data input, and Timothy Brathwaite for very helpful conversations and input. All views expressed and any errors made, however, are ours alone.

8. References


