Transportation Research Part C 125 (2021) 102981

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Generalized model for mapping bicycle ridership with crowdsourced data

Trisalyn Nelson [a], Avipsa Roy [b], Colin Ferster [c], Jaimy Fischer [d], Vanessa Brum-Bastos [e], Karen Laberee [c,*], Hanchen Yu [b], Meghan Winters [d]

[a] Department of Geography, University of California, USA
[b] School of Geographical Sciences and Urban Planning, Arizona State University, USA
[c] Department of Geography, University of Victoria, Canada
[d] Faculty of Health Sciences, Simon Fraser University, Canada
[e] Department of Geography, University of St. Andrews, UK

ARTICLE INFO

Keywords:
Bias-correction
LASSO
Big data
Bicycling ridership
Exposure
Strava

ABSTRACT

Fitness apps, such as Strava, are a growing source of data for mapping bicycling ridership, due to large samples and high resolution. To overcome bias introduced by data generated from only fitness app users, researchers build statistical models that predict total bicycling by integrating Strava data with official counts and geographic data. However, studies conducted on single cities provide limited insight on best practices for modeling bicycling with Strava as generalizability is difficult to assess. Our goal is to develop a generalized approach to modeling bicycling ridership using Strava data. In doing so we enable detailed mapping that is more inclusive of all bicyclists and will support more equitable decision-making across cities. We used Strava data, official counts, and geographic data to model Average Annual Daily Bicycling (AADB) in five cities: Boulder, Ottawa, Phoenix, San Francisco, and Victoria. Using a machine learning approach, LASSO, we identify variables important for predicting ridership in all cities, and independently in each city. Using the LASSO-selected variables as predictors in Poisson regression, we built generalized and city-specific models and compared accuracy. Our results indicate generalized prediction of bicycling ridership on a road segment in concert with Strava data should include the following variables: number of Strava riders, percentage of Strava trips categorized as commuting, bicycling safety, and income. Inclusion of city-specific variables increased model performance, as the R^2 for generalized and city-specific models ranged from 0.08–0.80 and 0.68–0.92, respectively. However, model accuracy was influenced most by the official count data used for model training. For best results, official count data should capture diverse street conditions, including low ridership areas. Counts collected continuously over a long time period, rather than at peak periods, may also improve modeling. Modeling bicycling from Strava and geographic data enables mapping of bicycling ridership that is more inclusive of all bicyclists and better able to support decision-making.

* Corresponding author.
E-mail address: klaberee@uvic.ca (K. Laberee).

https://doi.org/10.1016/j.trc.2021.102981
Received 10 March 2020; Received in revised form 4 November 2020; Accepted 10 January 2021
Available online 27 February 2021

1. Introduction

In North America, growing awareness of health, environmental, and economic benefits of active transportation have led to an unprecedented commitment to pro-bicycling policies, including improved safety (Pucher et al., 2010) and investment in bicycling infrastructure (Dill, 2009; Garrard et al., 2008). However, lack of high-quality data on bicycling ridership is stalling research and evidence based pro-bicycling decision making (Broach et al., 2012).

While cities invest significantly in vehicle count programs, most cities have only a handful of locations where bicycle counts are monitored electronically and/or rely on periodic volunteer manual count programs to capture bicycling ridership levels (Hyde-Wright et al., 2014). Lack of bicycling volume data is a substantive barrier to safety studies, which require exposure data (Miranda-Moreno et al., 2011; Vanparijs et al., 2015). There is limited knowledge among practitioners on how bicycling volume flows throughout a city. More research is needed in order to understand where investments are most needed and identify locations that lack baseline data for monitoring change associated with new policy or infrastructure.

The proliferation of fitness apps to track individual bicycling has led to the creation of large and complex datasets that document bicycle ridership patterns. Strava is the most popular of these fitness apps, capturing over 822 million activities globally in 2019 (Strava.com). Cities worldwide are accessing Strava data to better understand ridership patterns. Although Strava is clearly 'big data', this source is only a sample of bicycling ridership, biased toward people who use the Strava app. Strava users are disproportionately young adults (25–35 years in age) and male (Roy et al., 2019). Women, children, older adults, and low-income bicyclists are under sampled by Strava data.

Researchers are developing models to overcome the sampling bias in Strava data (Chen et al., 2020; Garber et al., 2019; Jestico et al., 2016; Roy et al., 2019). By integrating Strava data with multiple data sources it is possible to generate maps of predicted total bicycling volume that are more representative of all ages and abilities of bicyclists. However, most studies (Chen et al., 2020; Jestico et al., 2016) are done on single cities and build models based on data availability. While individual models can optimize accuracy for a specific city, a generalized model offers an approach for bias correcting data from many cities and may provide a good starting point for understanding the general relationships between all ridership and Strava sampled ridership. As well, the generalized model can act as a baseline from which to compare models optimized for specific cities. Such a comparative approach provides insight into how factors impacting bicycling cultures vary by city. Development of a generalized approach to bias correcting crowdsourced bicycling data is timely as platforms such as Strava Metro have begun promoting open data policies resulting in data becoming more accessible for research and planning purposes. Increased data availability necessitates supporting bias correction more broadly, with generalized approaches and a better understanding of how to optimize model performance. Studies on single cities provide limited generalizable insights on best practices for modeling bicycling with Strava data.

Our goal is to develop a generalized approach to modeling bicycling ridership using Strava data and compare the performance of generalized and city specific models in order to develop recommendations for how to optimize accuracy of models across cities. For this purpose, we use Strava data, official bicycle counts, and geographic data to model Average Annual Daily Bicycling (AADB) in five cities: Boulder, Ottawa, Phoenix, San Francisco, and Victoria. These cities reflect a range of climates (i.e., temperate to extreme), sizes, topographies (i.e., flat to hilly) and bicycling cultures (i.e., few bicyclists to many bicyclists). Using a machine learning variable selection technique called Least Absolute Shrinkage and Selection Operation (LASSO) (Tibshirani 1996), we identify variables important for prediction in *all* cities and independently in *each* city. Using the LASSO-selected variables as predictors in Poisson regression we build generalized and city-specific models and compare accuracy. Our basic modeling approach is informed by the concept that the proportion of total ridership sampled by Strava ridership will vary with different infrastructure characteristics (Sun et al., 2017), and that variability can be captured by built environment (Hochmair et al., 2019) and socio-economic characteristics (Conrow et al., 2018). Modeling bicycling ridership by integrating Strava data with official counts and geographic data enables detailed mapping that is more inclusive of all bicyclists and better able to support decision making and social equity (Ilieva and McPhearson, 2018).

2. Material and methods

2.1. Study area

Our study includes five North American cities of varying sizes, demographic and socio-economic characteristics, as well as accessibility to varied bicycle infrastructure (Table 1). Boulder, CO has a high mode share of bicycling (9.9%) and more than 300 miles of bicycling facilities (City of Boulder, 2019). Ottawa, ON has the fewest Strava users and hot summer and cold winter conditions. Phoenix, AZ has a hot desert climate and is one of the fastest-growing metropolitan areas in the USA, with only 0.48 miles of bicycle facilities (Cynecki and Lee, 2014) and <1% mode share of bicycling ridership. San Francisco, CA has a temperate climate, hilly topography (Table 1), and the greatest number of Strava users. Finally, Victoria, BC has mild temperatures and the highest percentage (6.6%) of the population that bicycles to work in Canada (Statistics Canada, 2017).

2.2. Data

We used street segment counts of bicyclists from Strava Metro (Table 1), official counts of bicyclists (Table 1), and variables that quantified the people, neighborhoods, and characteristics of streets (Table 2).

We obtained official count data from local governments in each city (Fig. 1), and these demonstrate a range of approaches taken by local governments. In Boulder and San Francisco, we used continuous data from permanent counters that report number of bicyclists

Table 1
Summary of the climate & demographic characteristics of each city.

City	Region	Population [*]	Annual Temperature [**]		Mode [***] share of bicycling	Strava ridership	Year Strava data collected	Total no. of official counts	Temporal Resolution of official counts	Data collection mechanism
			Min	Max						
Boulder	Colorado, USA	326,078	3.2 °C	18.5 °C	8.9%	Medium	2017	15	1 min	Permanent counters
Ottawa	Ontario, Canada	934,243	1.9 °C	11.4 °C	2.6%	Smallest	2016	1066	8 hrs	Video counters
Phoenix	Arizona, USA	4,857,962	17.4 °C	30.4 °C	0.8%	Medium	2016	37	Daily	Video counters
San Francisco	California, USA	884,363	7 °C	21 °C	3.9%	Largest	2017	47	1 min	Permanent counters
Greater Victoria	British Columbia, Canada	235,689	5.6 °C	14.4 °C	6.6%	Small	2016	71	Peak periods	Manual counts

Sources:
[*] Population data were obtained from US census Bureau ACS 2015–2017 for US cities and Statistics Canada for Canadian cities.
[**] Temperature data for US cities were gathered from the National Climate Data Center hosted by NOAA and for Canadian cities from Environment Canada Climate Normals 1981–2010.
[***] Mode share of bicycling was gathered separately for each city in collaboration with the transportation authorities and Statistics Canada for Canadian cities.

Table 2
Sources of all datasets and variables used in the generalized and city-specific models across all 5 cities.

Category	Variables	Data Source	Data availability	Reference
Ridership	Official counts	DoT officials of each city	All cities	Griffin and Jiao (2015) and Jestico et al. (2016)
	Number of Strava Riders	Strava ridership GIS shapefile from DoT of each city		
	% Strava trips that are commute	Strava ridership GIS shapefile from DoT of each city		
Safety and Design	Bicycle crash density	Official crash data from DoT in each city. Crowdsourced crash reports from BikeMaps.org	All cities	Hauer (1995), Ferster et al. (2017), Sallis et al. (2012), Winters et al. (2010), Saelens et al. (2003), Moudon et al., (2005), Sanders et al. (2017), Hankey et al. (2012)
	Average traffic speed limit	OpenStreetMap.org		
	AADT	DoT officials of Boulder and Phoenix, MioVision Scout video counters in Ottawa.	Boulder, Ottawa, Phoenix. (For All Streets)	
	Bicycling comfort and safety levels**	Ottawa Bicycling Network shapefile	Ottawa	
	Number of traffic lanes	San Francisco DoT and Digital Road Atlas in British Columbia, and road network shapefile in Ottawa	Ottawa, San Francisco, Greater Victoria	
	Street type	British Columbia Digital Road Atlas and OpenStreetMap in, San Francisco	San Francisco, Greater Victoria	
	Bicycling infrastructure type	San Francisco DoT and bicycling network shapefiles from Canadian cities	San Francisco, Ottawa, Greater Victoria	
	City suggested route with no bicycling infrastructure	Bicycling network shapefile Ottawa	Ottawa	
	% Trucks crossing intersections	MioVision Scout video counters in Ottawa		
	Number of pedestrians at intersections	MioVision Scout video counters in Ottawa		
Land Use*	Distance to green spaces	Open Data Portal of the DoT in each city	All cities	Roy et al. (2019), Noland et al. (2011), Pikora et al.(2003), Griswold et al.(2011)
	Distance to residential areas	Open Data Portal of the DoT in each city		
	Distance to commercial areas	Open Data Portal of the DoT in each city		
	Distance to bike parking	San Francisco Open Data Portal	San Francisco	
	Distance to educational institutions	San Francisco Open Data Portal, British Columbia Open Data Catalogue	San Francisco, Greater Victoria	
	Distance to sea shore	San Francisco Open Data Portal	San Francisco	
	Distance to trailheads	Boulder County Open Data Portal	Boulder	
	Distance to water bodies	OpenStreetMap	Ottawa	
Demographics	% Population with at least college education	US Census Bureau 2015–17, Statistics Canada 2016 Census	All cities	Winters et al. (2010), Noland et al. (2011), Cervero et al. (2009), Sallis et al. (2013), Hankey et al. (2012)
	% Population with at least high school education	US Census Bureau 2015–17, Statistics Canada 2016 Census		
	% White Population	US Census Bureau 2015–17, Statistics Canada 2016 Census		
	% Female population	US Census Bureau 2015–17, Statistics Canada 2016 Census		
	Median age	US Census Bureau 2015–17, Statistics Canada 2016 Census		
	Population density	US Census Bureau 2015–17, Statistics Canada 2016 Census		
	% Veterans	US Census Bureau 2015–17, Statistics Canada 2016 Census	Boulder, San Francisco	
	% Male bicyclists	US Census Bureau 2015–17, Statistics Canada 2016 Census	San Francisco, Greater Victoria	
Socio Economic	Median household income	US Census Bureau 2015–17, Statistics Canada 2016 Census	All cities	Griswold et al. (2011), Strauss et al. (2013), Sallis et al. (2009)
	Crime density	Crime data portal in Boulder	Boulder	
	Unemployment rate	Statistics Canada 2016 Census	Ottawa	
Topography	Slope of the street	DEM provided by city officials		Broach et al. (2012) and Hood et al. (2011)

(*continued on next page*)

Table 2 (*continued*)

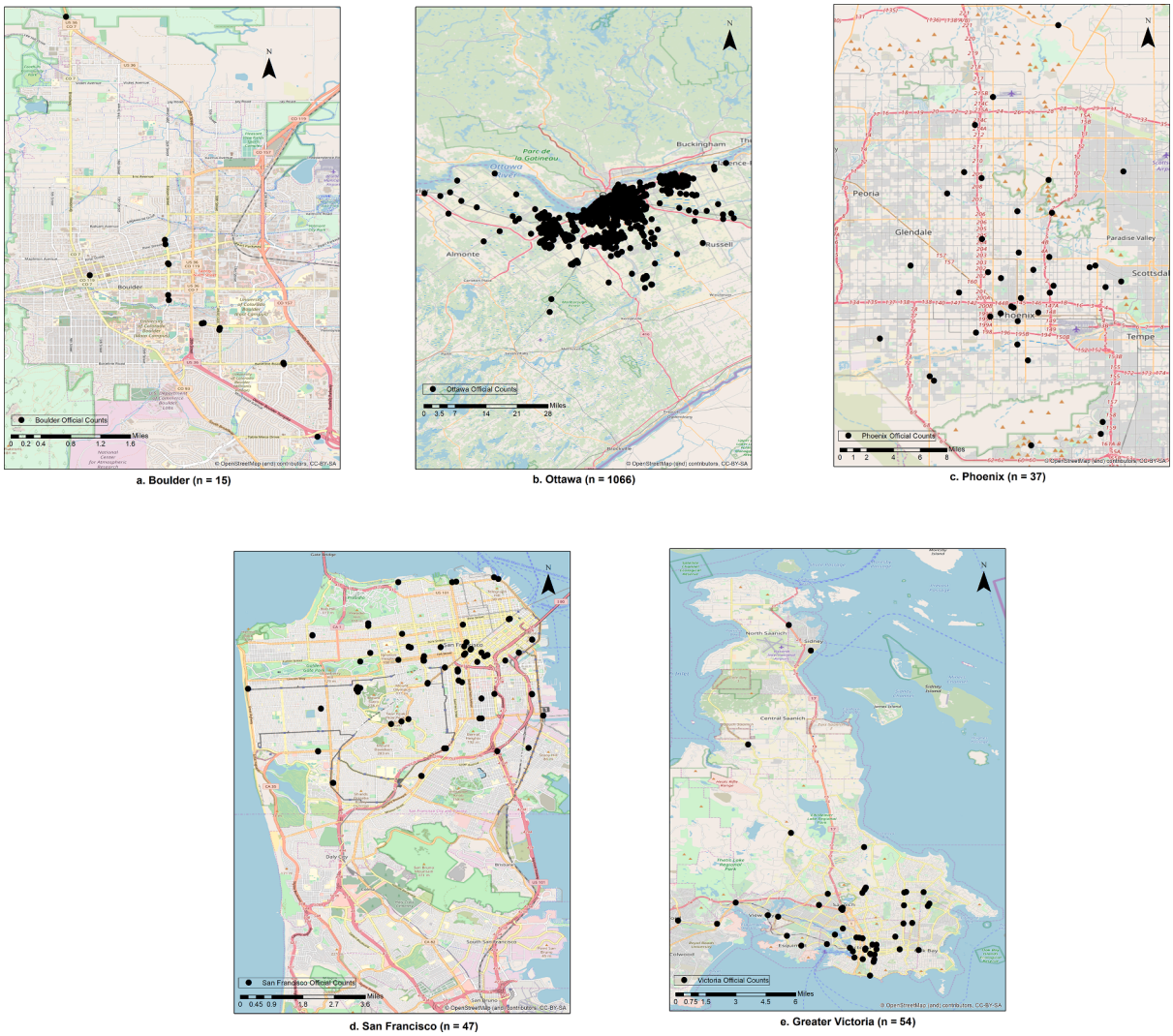
Category	Variables	Data Source	Data availability	Reference
Climate	Counts conducted in winter months	Temperature inferred by Canadian climate normals	Boulder, Ottawa, San Francisco, Greater Victoria Ottawa	Spencer et al. (2013)

***The description about how each variable was calculated and operationalized is listed in Supplementary Table S1.

* Distance variables were calculated in ArcGIS as Euclidean distances in US feet from a street segment to the nearest polygon of a particular land use type (Eg: residential are, commercial area, green space etc.).

** The comfort classification was derived using Can-BICS (Winters and Zanotto, 2019).

each minute. Though Boulder official counts had dense temporal coverage, they lacked spatial coverage as all 15 counters were closely spaced mostly in and around the core downtown area (Fig. 1a). In contrast, San Francisco had more plentiful official counts recorded by 47 counters that were more evenly spread throughout the city. In Ottawa, MioVision Scout video counters were used to collect data for 8-hour periods at over 1000 locations. Ottawa's counts were spread across the city but were only collected for one or two days per year at each location during peak commute hours (Fig. 1b). Phoenix video counters were used for one-week periods and included locations from outside the city (Fig. 1c), which increases the range of built environment conditions represented in the data. In Victoria, all count data (71 observations across 54 different locations) were collected manually by volunteer counters and included spatial



a. Boulder (n = 15)
b. Ottawa (n = 1066)
c. Phoenix (n = 37)
d. San Francisco (n = 47)
e. Greater Victoria (n = 54)

Fig. 1. Maps showing spatial distribution of all official count locations in all 5 cities.

locations beyond the core (Fig. 1e). Counts were limited to peak hours (7–9 am and 3–6 pm) for three weekdays in May and October.

Strava Metro is a data product generated from GPS data recorded using the Strava fitness app. Data are provided at street level and bicycle counts are reported for street geometries with one-minute temporal resolution. The raw counts are rolled up to the nearest hour in Strava Metro. The street network used by Strava is from OpenStreetMap (OpenStreetMap contributors, 2019). Strava data are reported for two spatial units: nodes (points), which represent intersections or terminals of street segments, and edges (lines), which represent lengths of roads or trails. To enable integration with manual counts, we used street segment counts for Boulder, Phoenix, San Francisco, and Victoria, and intersection counts for Ottawa.

For each street segment or node Strava variables include: *number of bicycle activities,* representing unique bicycle trips made by Strava users; *number of total bicyclists,* representing unique Strava users, and *type of trip (commute/ non-commute or recreational),* representing activities for transportation purposes determining whether a bicyclist is a commuter or not. The commute category is the result of a proprietary trip classification model that Strava has developed. In Strava's data model, "commuting" refers to all non-leisure trips. Strava reports the model to have 85% accuracy (Rodrigo Davis per comm., Dec 17, 2019). Strava data also includes a city-wide summary on gender and age of Strava users.

The Strava counts were matched at the street-segment level to the official counts using the method proposed by Roy et al. (2019). We performed an additional sanity test on the aggregated Strava data at all locations where we had official counts, to check whether the Strava counts exceeded the official bicyclist counts. In complex intersections it is possible that official counts only include bicyclists moving through specific lanes, while Strava counts include ridership in all directions. This occurs in very few locations (e.g., two locations in the Phoenix data). Any locations where Strava counts exceeded official counts were eliminated prior to our LASSO and Poisson regression analysis to ensure the model was representative.

To account for bias in the crowdsourced Strava data, we included geographical covariates that represent bicycling safety, road design, land-use, demographics, socio-economics, topography, and climate (Table 2). Bicycle crash density is a function of exposure (i.e. the number of bicycle trips) (Hauer, 1995), and crowdsourced bicycle crash density also relates to rates of digital participation (Ferster et al., 2017). In addition, features of the built environment such as increased intersection density and traffic calmed streets are related to higher bicycling trips (Winters et al., 2010). Higher rates of bicycling have been associated with parks and playgrounds (Noland et al., 2011) and demonstrate how land use can be used to predict ridership (Roy et al., 2019). Previous research on the demographics of who bicycles can also be used to predict ridership. For example, Sallis et al. (2013) found young, educated white males were more likely to ride bicycles. Slope has been found to be important in route choice for utilitarian bicycling (Broach et al., 2012) and climate has been shown to impact the decision to bicycle (Spencer et al., 2013).

We attributed official count locations with Strava data and all geographic covariates. Both Strava and official counts were represented at multiple temporal scales ranging from fine (daily) to coarse (annual) resolutions. For uniform interpretation of all ridership models, the response variable in the generalized and city-specific models was called Annual Average Daily Bicycling (AADB). However, in Ottawa "daily" referred to the 8 h of data collection, while in Victoria AADB represented peak daily volumes.

2.3. Analysis

As a preliminary step to identifying the bias-adjustment factors, we assessed how well Strava data alone explain the variation in overall bicycle ridership. In order to quantify the total number of bicyclists represented by a single Strava bicyclist, we used an ordinary least squares regression with Strava data as the only independent variable to predict AADB in each city separately (Table 3). Following this we then used variable selection (discussed in Section 2.3.1) to identify specific bias-adjustment factors in each city.

2.3.1. Identifying bias-adjustment factors for each city using LASSO

For each city we identified variables that are important predictors of ridership. Using a machine learning variable selection technique called Least Absolute Shrinkage and Selection Operation (LASSO) (Tibshirani 1996), we measured variable multicollinearity using variance inflation factor and selected variables that explain the maximum variance in the total bicycling ridership. LASSO reduces the number of variables by reducing the sum of the absolute value of their regression coefficients to be less than a fixed value, which forces coefficients of certain variables to be set to zero, effectively resulting in a simpler model that does not include those variables. A tuning parameter 'α' determined the optimal number of uncorrelated adjustment factors that explain maximum variance in the outcome.

Variables used in the LASSO were based on data availability and a city's geography and bicycling culture. Variables considered for each city are listed in Table 2. After applying a 10-fold cross-validation, LASSO returned the most important adjustment factors specific

Table 3
Annual Average Daily Bicycling (AADB) for each city represented by 1 Strava bicyclist. We use the regression coefficient of an Ordinary Least Squares (OLS) regression between Strava counts and official counts (AADB) to derive this value.

City	Year Strava data collected	Total no. of official count observations retained	$^{+}$AADB ~ 1 Strava rider (OLS)
Boulder	2017	15	**271**
Ottawa	2016	1058	**41**
Phoenix	2016	35	**53**
San Francisco	2017	47	**109**
Greater Victoria	2016	71	**55**

to each city. We ranked LASSO selected variables using a score function and chose all variables with a score above 0.2 to include in the models. We tested the sensitivity of the score function and found that a threshold of 0.2 resulted in reasonable regression fits without overfitting.

2.3.2. Modeling bicycling ridership and assessing accuracy

Using the unique set of variables identified through LASSO (referred to as "adjustment factors" herein) we fit a pair of Poisson regression models on a log scale for each of the five cities, resulting in a total of ten different models. The first model in each city, which we call the *generalized model*, used the variables that were consistently important in all of the cities as the predictors of bicycling ridership. The second model in each city, which we call the *city-specific model*, included the variables in the generalized model along with all additional variables identified by the LASSO as predictors of bicycling ridership in that particular city. In both models, we used the AADB count from the official counts as our response variable. Considering that Strava counts may have an endogeneity problem, we constructed a system of equations for each model and tested the endogeneity by Hausman test. The tests showed Strava counts are not an endogenous variable for all cities except Ottawa (Table 4). Hence, we estimated the models in all cities except Ottawa by OLS but estimated the models in Ottawa by instrumental variables estimation. The instrumental variables are number of Strava bicyclists and number of Strava bicycling activities in the previous year.

The high temporal resolution of Strava data make it possible to map the number of bicyclists at any temporal resolution. While we could map every minute throughout a year, such detailed resolution is generally not practical, and average values by day, month, or year tend to be more useful. However, daily and monthly temporal resolution maps can be difficult to make because it requires all official counts to be taken during that same time. In other words, to make a map of average ridership in September, all the official counts must be collected in September. In reality, official counts are usually collected a few times per year or throughout the year making annual average ridership most appropriate.

Owing to lack of sufficient ground truth, we used all the locations in each city for training as well as testing our model. We used a k-fold cross-validation technique to split the entire dataset into 10 folds and trained the model. Once an optimal fit was achieved, we calculated the in-sample accuracy determined by the coefficient of determination of an OLS regression between the true AADB and predicted AADB.

We calculated the prediction error as the absolute difference between actual AADB and predicted AADB separately for both generalized and city-specific models in each city. Additionally, we generated a cumulative error distribution and reported prediction error for 25%, 50%, 75%, and 99% of the segments in each city.

3. Results

At 1209 official count locations in five cities we compared the Strava sample of ridership and all ridership. Based on how the OLS model would change if Strava sampled ridership was increased by one bicyclist, a single Strava bicyclist represents between 41 and 271 total bicyclists (Table 3). Strava sampled the highest proportion of total bicycling in Ottawa, where, on average, one Strava sample bicyclist represented 41 officially counted bicyclists, and the lowest proportion of bicycling in Boulder, where, on average, one Strava sample bicyclist represented 271 officially counted bicyclists (Table 3).

3.1. Lasso variable selection

In Fig. 2 we show results of the LASSO and indicate the strength of each variable for prediction. The normalized color scale on the right of Fig. 2 indicates a variable's predictive strength with darker colors showing more important variables. Variables that did not have a score above 0.2 in any city were not shown. Of note, the number of Strava bicyclists was amongst the most important predictor variable in all five cities. Other variables that were consistently important were the percentage of Strava bicyclists commuting, the percentage of the general population who are female, bicycle crash density, and median household income.

3.2. Generalized model fit

The in-sample accuracy for the models in each city ranged from 0.08 to 0.80 (Fig. 3), when the generalized model was applied.

Table 4
Hausman tests of the generalized model and the *city-specific* model.

	Generalized model				
	Boulder	**Ottawa**	**Phoenix**	**San Francisco**	**Greater Victoria**
Hausman test	0.09	20.01	0.01	0.76	0.21
p-value	0.76	0	0.92	0.38	0.65
	City-specific model				
	Boulder	**Ottawa**	**Phoenix**	**San Francisco**	**Greater Victoria**
Hausman test	0.12	11.3	0.33	0.00	1.00
p-value	0.7253	0.0008	0.567	0.9593	0.3181

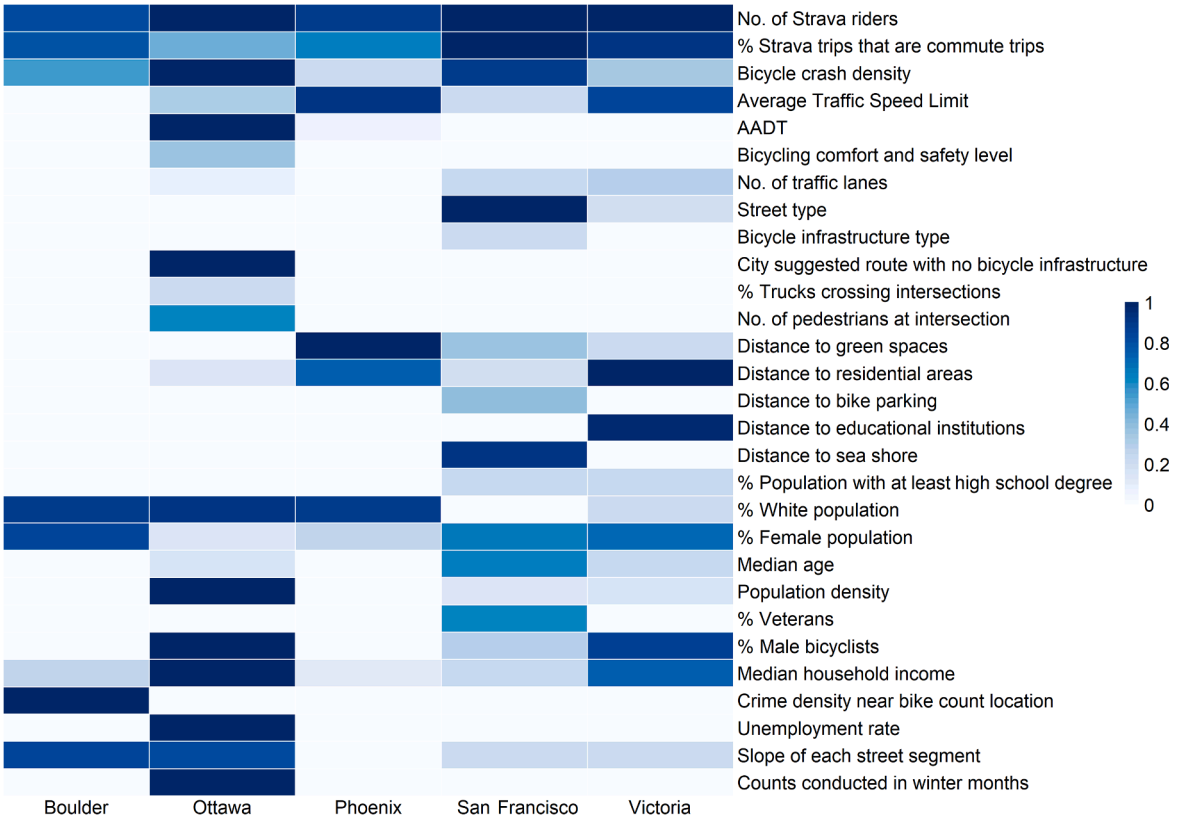


Fig. 2. Normalized LASSO ranks for all variables used in *city-specific* model.

Table 7 shows how well ridership was predicted, by calculating the difference between observed and modeled bicycling ridership levels. For example, with the generalized model, Boulder had 25% of streets predicted to within ±46 AADB, 50% predicted to within ±101 AADB, 75% of streets predicted to within ±120 AADB, and 99% of streets predicted to within ±428 AADB. When we compared accuracy of the generalized model across all cities, Victoria and Phoenix were the most accurate and San Francisco and Boulder the least accurate (Fig. 4).

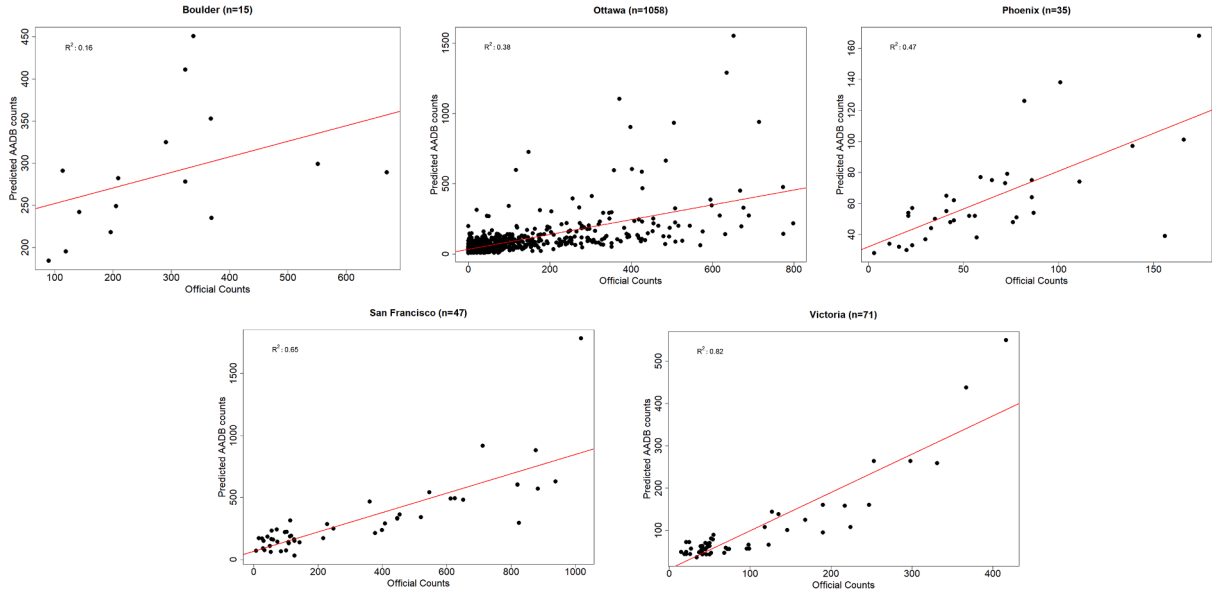


Fig. 3. In-sample fit of predicted AADB using *generalized* model.

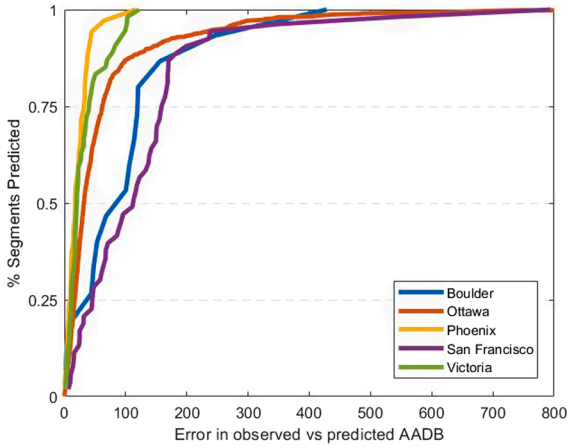


Fig. 4. Accuracy comparison of bias-corrected AADB using *generalized* model.

3.3. City-specific model fit

City-specific variables were also identified from the LASSO variable selection and reflected topography, weather, and socio-economic characteristics of streets. As expected, the AIC fits were better (smaller values indicate greater information content) in the models dedicated to each city compared to the generalized model for that city. We see that the AIC for Boulder improved from 1238.42 (Table 5) to 434.26 (Table 6) when we applied the city-specific model. Similarly, for Victoria, Phoenix, and San Francisco the AIC improves from 1214, 753, and 4864 (Table 5) to 748, 493, and 1739 (Table 6) respectively, when we chose the city-specific model over the generalized model. For Ottawa the AIC reduced from 100,546 (Table 5) to 58,624 (Table 6); the AIC was typically higher for Ottawa due to a large number of observations.

In terms of variable importance, for Boulder and San Francisco the slope of the street segment accounted for the overall variation in bicycle ridership from geographic factors for the city-specific models, whereas for Phoenix it was proximity to green spaces (Table 6). Other factors such as topography, job density class, and suggested route with no bicycle infrastructure were occasionally predictors. Distance to commercial areas, distance to water bodies, distance to trailheads, and percentage population with at least college education were never selected by LASSO for any of the cities and do not need to be included in models.

The city-specific models showed a large improvement in accuracy for 99% of the predicted segments for all five cities. The error margins were reduced the most for Ottawa bicyclists from ±1287 to ±558 (Table 7). However, for 75% of the predicted segments Boulder and San Francisco had the largest reduction in error margins (33 and 50 bicyclists, respectively) (Table 7).

3.4. Model accuracy comparisons for generalized and city-specific models

The in-sample accuracy for city-specific models ranged from 0.68 to 0.92, indicating a better fit compared to the generalized model (Fig. 5). The fit was lowest for Ottawa and highest for Victoria. Prediction accuracy was determined by calculating the difference between observed and modeled bicycling (Table 7). When we considered the prediction accuracy for 75% of the segments in each city (Table 7), we found that Boulder, Ottawa, Phoenix, San Francisco, and Victoria showed an overall increase in accuracy for the city-specific models compared to the generalized models as the errors were reduced by ±33, ±15, ±10, ±50, and ±11 bicyclists, respectively. Predicted values were closest to observed values for Victoria with error margins as low as ±10 bicyclists for 25% of the

Table 5
Poisson Regression coefficients of the *generalized* model. Note: All the coefficients are exponentiated. Values less than 1 indicate a decrease in AADB by a percentage (1-coefficient) * 100 and those above 1 indicate an increase by a percentage (coefficient-1) * 100.

	Dependent Variable: AADB				
	Boulder (1)	**Ottawa (2)**	**Phoenix (3)**	**San Francisco (4)**	**Greater Victoria (5)**
Median household income	1.00^{***}	0.89^{***}	1.03^{***}	2.61^{***}	9.69^{**}
Bicycle crash density	1.34^{***}	1.14^{***}	1.05^{***}	1.05^{***}	1.08^{***}
No. Strava riders	1.00^{***}	1.07^{***}	1.11^{***}	1.02^{***}	1.14^{***}
% Strava trips that are commutes	0.95^{***}	1.17^{***}	1.18^{***}	0.43^{***}	0.97^{***}
Observations	15	1058	35	53	54
Log Likelihood	−614.2	−50268.11	−371.56	−2427.2	−602.14
AIC	**1238.42**	**100546.2**	**753.11**	**4864.41**	**1214.29**
Note:					$^{**}p < 0.05$; $^{***}p < 0.01$

Table 6
Poisson Regression coefficients of *city-specific* model.

		Dependent Variable: AADB				
		Boulder	**Ottawa**	**Phoenix**	**San Francisco**	**Greater Victoria**
		(1)	**(2)**	**(3)**	**(4)**	**(5)**
Category	**Variables**					
Ridership	*No. Strava riders*	1.01***	1.05***	1.10***	1.02***	1.11***
	% Strava trips that are commutes	0.90***	1.16***	1.17***	0.72***	1.15***
Safety and Design	*Bicycle crash density*	1.11***		1.02***	1.84***	0.04
	Average traffic speed limit			0.90***	0.73***	0.89***
	AADT		0.98***			
	Bicycling comfort and safety level (High or Medium)		1.29***			
	Number of traffic Lanes (1)				2.10***	
	Number of traffic lanes (2)					0.67***
	Number of traffic lanes (3)					0.67***
	Number of traffic lanes (4)					0.55***
	Street type (Main)				1.67***	
	Bicycle infrastructure type (Protected / off-street bike lane)				1.23***	
	Bicycle infrastructure type (Painted bike lane with no physical protection)				1.56***	0.92***
	Bicycle infrastructure type (Signed bike route)				2.22***	1.11***
	City suggested route with no bicycle infrastructure		1.45***			
	% Trucks crossing intersections		0.63***			
	No. of pedestrians at intersection		1.00***			
Land Use	*Distance to green spaces*			0.72***	0.36***	0.65***
	Distance to residential areas			1.10***		
	Distance to bike parking				0.34***	
	Distance to educational institutions					0.83***
	Distance to sea shore				0.97***	
Demographics	*% Population with at least high school degree*				1.00***	0.99***
	% White population	0.99***				1.04***
	% Female population	0.73***			1.17***	0.57***
	Median age				0.71***	1.09***
	Population density		1.04***			
	% Veterans				1.36***	
	% Male bicyclists		1.58***		1.65***	0.94***
Socio Economic	*Median household income*		0.93***	0.99***	0.99***	0.97***
	Crime density near bike count location	1.21**				
	Unemployment rate		0.88***			
Topography	*Slope of each street segment*	0.65***	0.83***		0.90***	1.43***
Climate	*Counts conducted in winter months (True)*		0.48***			
	Observations	15	1058	35	53	54
	Log Likelihood	−209.13	−29297.07	−237.89	−850.89	−355.22
	AIC	**434.26**	**58624.15**	**493.78**	**1739.78**	**748.44**
	Note:	*p < 0.1; **p < 0.05; ***p < 0.01				

segments and ±91 for 99% of the segments. For Ottawa the predictions had the most error with the largest error margin of ±558 for 99% of the segments, which we explain in the discussion relates to the temporal extent of official data.

The cumulative error distribution plots in Fig. 4 and Fig. 6 highlight that the generalized model, as well as the city-specific models, were most accurate for Victoria and Phoenix. The least accurate predictions for both models were found for Boulder and San Francisco (Fig. 4; Fig. 6). As expected, there was a reduction in the error margins for all cities (Table 7) when we added city-specific variables.

4. Discussion

For North American cities aiming to map overall bicycle ridership, Strava is a promising data source that is continuous over space and time. Researchers have demonstrated that by integrating Strava data with official counts and GIS covariates it is possible to map bicycling ridership (Jestico et al., 2016; Roy et al., 2019). However, from studies using Strava data from individual cities it is difficult to determine best practices for extending modeling to other cities or for implementing models on regions or multiple cities simultaneously. A strength of this research is the ability to build hypotheses on why variation in prediction performance occurs. By observing how predictive accuracy was influenced by modeling inputs and approaches, we build suggestions for best approaches to modeling of bicycle volumes with Strava data for a range of available data.

Table 7
Model error comparison between *generalized* model and *city-specific* model.

City	*n* – official data	% Segments predicted	Error margins of predicted AADB		Summary of predicted AADB			
			Generalized	City-specific	Generalized		City-specific	
					Median	IQR*	Median	IQR*
Boulder	15	25%	±46	±46	301	84	240	131.5
		50%	±101	±63				
		75%	±120	±87				
		99%	±428	±272				
Ottawa	1058	25%	±17	±8	50	55	36	70
		50%	±32	±20				
		75%	±62	±47				
		99%	±1287	±558				
Phoenix	35	25%	±10	±8	54	28	57	50
		50%	±18	±15				
		75%	±32	±22				
		99%	±116	±74				
San Francisco	47	25%	±46	±20	207	182	183	394
		50%	±113	±42				
		75%	±160	±110				
		99%	±794	±352				
Greater Victoria	71	25%	±12	±10	61	45	56	73
		50%	±20	±21				
		75%	±40	±29				
		99%	±122	±91				

* IQR – Inter-Quartile Range.

Recommendation 1. Variables to always use

When building a generalized model, we recommend always including the following four variables: 1) number of Strava bicyclists; 2) the percentage of Strava bicyclists commuting; 3) bicycle crash density; and 4) median household income. Income data are typically

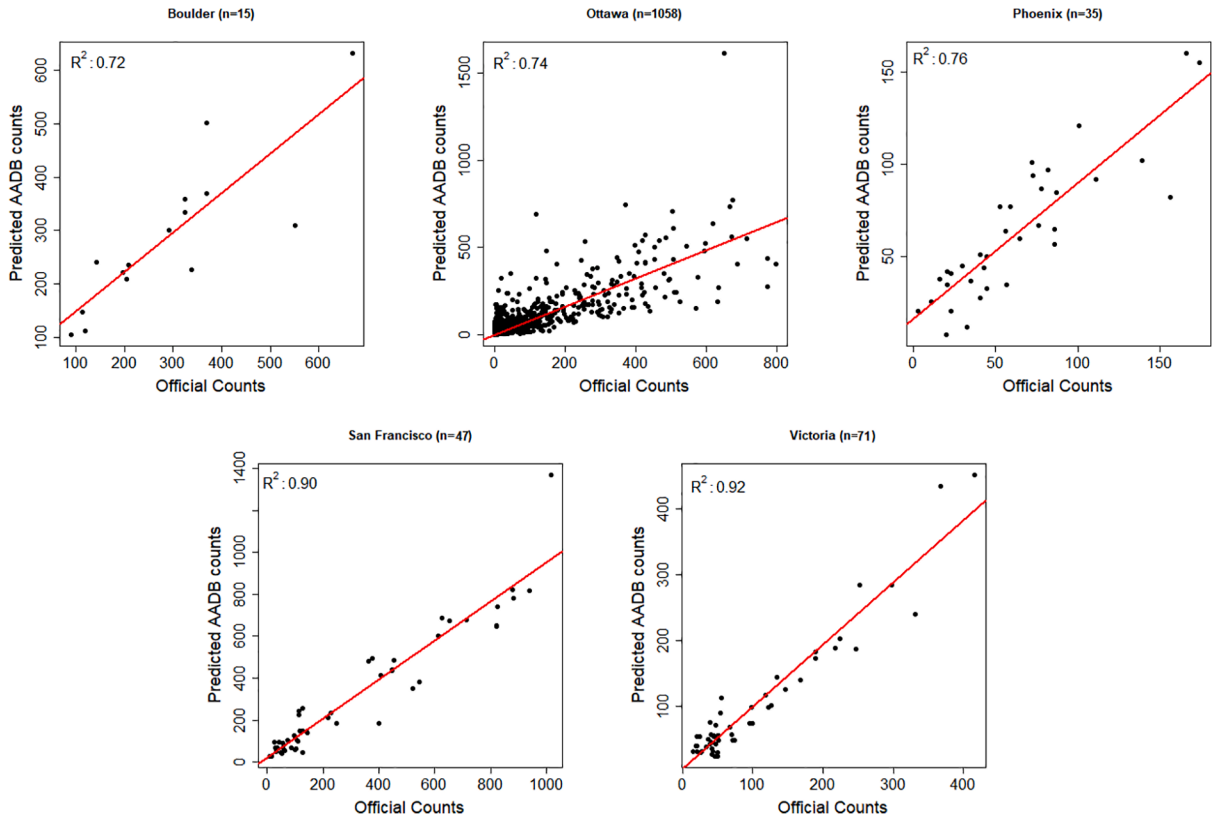


Fig. 5. In-sample fit of predicted AADB using *city-specific* model.

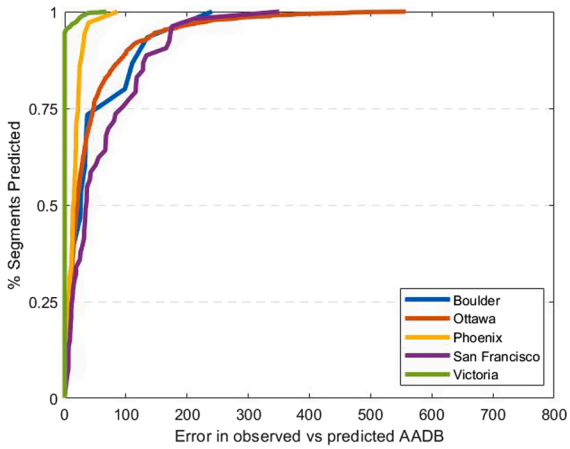


Fig. 6. Accuracy comparison of bias-corrected AADB using *city-specific* model.

available from the census, while bicycling safety data may be acquired from police, insurance companies, or Departments of Transportation. New sources of bicycling safety data, such as those being generated from the crowdsourcing project BikeMaps.org (Nelson et al., 2015), can also enhance data by filling in official reporting gaps.

Income and safety are key variables that have been identified as critical to bicycling ridership as they influence who has access to bicycling. While lower income individuals are often required to bicycle or walk in unsafe road conditions (Noland et al. 2013; Yu, 2014), in some cities, higher bicycling levels have been associated with higher incomes (Fuller and Winters, 2017). It is not uncommon in North America for access to safe bicycling, either low volume streets or bicycling infrastructure, to be associated with higher income neighborhoods (Braun et al., 2019). Finally, research has shown that concerns about safety are the number one barrier to increased bicycling (Porter et al., 2020) and how safe a road is will impact who will bicycle there and how often it is selected as part of a bicycling route.

Recommendation 2 – Variables to use when available

As one might expect, the city-specific models, which included more nuanced data, also had better accuracy. It is interesting that variables important in all cities represent the people on bicycles or road safety. No land use, topography, or climate characteristics were universally selected; however, these variables can become important for specific cities. For example, topography is a useful predictor in San Francisco where there are many hills and seems to be in agreement with previous studies (Hood et al., 2011; Menghini et al., 2009) that found slope of the terrain to be an important predictor of ridership. Variables that were predictive for at least two to four cities include: distance to residential areas (Roy et al., 2019), distance to commercial areas, and city designated route (with no infrastructure) supporting previous work that availability of bicycle infrastructure affects ridership in general (Dill and Carr, 2003).

Recommendation 3. It is all about official counts

The models' accuracy was heavily influenced by the number of official count locations and how representative the official counts are of the full range of conditions on the bicycling network. In fact, the nature of official count data seems to be more important than adding additional predictor variables into a city-specific model. As evidence, consider Victoria and Phoenix, which, regardless of the model inputs, had the best predictions of all bicyclists (Table 7). That the model performance was similarly high in Victoria and Phoenix is interesting given the differences in climates, number of Strava users, and mode share of bicycling. In both cities official counts were collected by regional governments, which mean that streets outside the downtown core were surveyed and therefore represent more diverse conditions. In San Francisco, which had much lower accuracy, official counts only covered the core downtown area.

Typically, official count programs favor high ridership streets, but moderate and low ridership streets are critical for model prediction. Recently, our team developed an approach to using Strava data to stratify locations for official counts to enhance representativeness for modeling (Brum-Bastos et al., 2019). But, as with any modeling it is not possible to predict outside the range of conditions that the model is trained on. If the model is only built for high bicycling streets or for high income areas, it is likely to perform poorly in other locations. Therefore, a holistic sampling strategy in terms of placing bicycle counters is of utmost importance.

It is also important to note that official data were collected for peak periods in Victoria and so the official counts were taken during specific conditions, when ridership is high. In contrast, daily data collected in Phoenix has more variability representing low and high ridership times. The difference shows the challenges of predicting lower ridership. In a sense, the high accuracy of Victoria is biased due to the focus on small windows of time with high ridership, even though it captures bicyclists who are commuters and may be less biased towards Strava riders.

The duration of official counts may be more important than the number of counters. Ottawa is a typical example. Although Ottawa

had over 1000 counts, it suffered from a weak in-sample fit (Figs. 3 and 5). In Ottawa, counts were mostly measured for short time periods and sampling was designed for motorized traffic (Table 1). Ottawa has a limited (n = 12) set of permanent bicycle counters that collect temporally continuous data, however, they represented a very narrow set of geographic conditions with high ridership. Another issue with short duration counts is that prediction at finer time periods will be hampered due to missing data. Strava data are recorded every minute. Here we predict AADB, but with continuous official counts, like those available from eco-counters, it would be possible to use Strava for prediction of ridership by season or even daily.

A final recommendation on official count programs is to use screenlines instead of intersection counts. This was a clear learning from the massive data formatting effort required to preprocess official counts from the range of formats across cities. Intersection counts are more difficult to integrate with Strava, as traffic flows includes multiple directions of travels and turns.

Recommendation 4. Avoid overselling results

It is important to remember that the ridership volume generated by statistical integration of official bicycle counts, Strava data, and GIS data are a modeled representation of average conditions. The best use of these data should be as categories of bicycling volume. If predicted ridership estimates are represented as categories of ridership (i.e., from low to very high ridership) the data can be used with greater confidence. Categorical ridership maps, that are continuous across a city, offer huge advantages over more traditional datasets that capture only a few locations. Many applications that benefit from city wide data do not need precise values everywhere. By representing data with maps showing low to high ridership categories we do not run risk of overselling precision and create data that are more accurate. Furthermore, we caution that users avoid interpreting small variances in number of total average annual daily ridership as meaningful, especially on street segments with lower ridership. The values in Table 7 that show the range of error associated with prediction are a good guide for determining appropriate categories.

5. Conclusions

In North American cities it is possible to use Strava data as an input to mapping all bicycle ridership. When predicting all bicycling ridership with Strava data we suggest including variables on income and bicycling safety. If available, additional covariate data, which represent geographic and socioeconomic conditions of a specific city, can be included. The biggest factor that will impact prediction accuracy is how the official counts are collected. It is critical that official counts are taken at the range of bicycling conditions across the bicycling network, including low ridership areas. Temporally continuous counts and screenline counts can also be beneficial in improving accuracy. Finally, while models predict values with the precision of a single bicyclist, modeled results should perhaps be communicated categorically: such as low, medium, and high ridership. Strava is a biased sample of ridership, but it is continuous in space and time and integration of GIS data in a statistical model enables Strava to be used to map spatial variation in all ages and abilities of bicycling. By mapping spatial variation in all bicycling throughout these five cities we have information that better informs inclusive and healthy transportation solutions.

CRediT authorship contribution statement

Trisalyn Nelson: Conceptualization, Funding acquisition, Writing - original draft, Supervision, Methodology. **Avipsa Roy:** Formal analysis, Data curation, Methodology, Writing - original draft. **Colin Ferster:** Formal analysis, Data curation, Methodology, Writing - review & editing. **Jaimy Fischer:** Formal analysis, Data curation, Methodology, Writing - review & editing. **Vanessa Brum-Bastos:** Formal analysis, Data curation, Methodology, Writing - review & editing. **Karen Laberee:** Project administration, Writing - review & editing. **Hanchen Yu:** Formal analysis. **Meghan Winters:** Conceptualization, Writing - original draft, Methodology.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgements

This work was supported by a grant (#1516-HQ-000064) from the Public Health Agency of Canada. MW is supported by a Scholar Award from the Michael Smith Foundation for Health Research.

The authors would like to thank Tetsuro Ide of the City of Ottawa for providing intersection count data; Joe Castiglione of the San Francisco County Transportation Authority for providing spatial data on the transportation infrastructure in San Francisco city; Tori Winters of the San Francisco Municipal Transportation Agency (SFMTA) for providing spatial data on the official counts for San Francisco; Jamie Parks (SFMTA) for providing official count data for San Francisco; Alex Phillips and Alex Hyde-Write from the City of Boulder and Boulder County for sharing official count data and supporting our work throughout; and to Jay Douillard and John Hicks from the Capital Regional District (Victoria, BC) for providing manual count data.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.trc.2021.102981.

References

Braun, L.M., Rodriguez, D.A., Gordon-Larsen, P., 2019. Social (in)equity in access to cycling infrastructure: Cross-sectional associations between bike lanes and area-level sociodemographics in 22 large US cities. J. Transp. Geogr. 80, 102544 https://doi.org/10.1016/j.jtrangeo.2019.102544.

Broach, J., Dill, J., Gliebe, J., 2012. Where do cyclists ride? A route choice model developed with revealed preference GPS data. Transport. Res. Part A: Policy Pract. 46 (10), 1730–1740. https://doi.org/10.1016/j.tra.2012.07.005.

Brum-Bastos, V., Ferster, C.J., Nelson, T., Winters, M., 2019. Where to put bike counters? Stratifying bicycling patterns in the city using crowdsourced data. Transport Findings https://doi.org/10.32866/10828.

Cervero, R., Sarmiento, O.L., Jacoby, E., Gomez, L.F., Neiman, A., 2009. Influences of built environments on walking and cycling: lessons from Bogotá. Int. J. Sustain. Transport. 3 (4), 203–226. https://doi.org/10.1080/15568310802178314.

Chen, C., Wang, H., Roll, J., Nordback, K., Wang, Y., 2020. Using bicycle app data to develop Safety Performance Functions (SPFs) for bicyclists at intersections: A generic framework. Transport. Res. Part A: Policy Pract. 132, 1034–1052. https://doi.org/10.1016/j.tra.2019.12.034.

City of Boulder, 2019. Bicycling in Boulder blog. https://bouldercolorado.gov/goboulder/bike (accessed 25 Feb 2020).

Conrow, L., Wentz, E., Nelson, T., Pettit, C., 2018. Comparing spatial patterns of crowdsourced and conventional bicycling datasets. Appl. Geogr. 92, 21–30. https://doi.org/10.1016/j.apgeog.2018.01.009.

Cynecki, M.J., Lee, J.C., 2014, November 1. Comprehensive Bicycle Master Plan, City of Phoenix report. https://www.phoenix.gov/streetssite/Documents/Bicycle Master Plan/2014bikePHX_Final_web.pdf (accessed 25 Feb 2020).

Dill, J., 2009. Bicycling for transportation and health: the role of infrastructure. J. Public Health Policy 30 (1), S95–S110. https://doi.org/10.1057/jphp.2008.56.

Dill, J., Carr, T., 2003. Bicycle commuting and facilities in major US cities: if you build them, commuters will use them. Transp. Res. Rec. 1828 (1), 116–123. https://doi.org/10.3141/1828-14.

Ferster, C.J., Nelson, T., Winters, M., Laberee, K., 2017. Geographic age and gender representation in volunteered cycling safety data: A case study of BikeMaps.org. Appl. Geogr. 88 (September), 144–150. https://doi.org/10.1016/j.apgeog.2017.09.007.

Fuller, D., Winters, M., 2017. Income inequalities in Bike Score and bicycling to work in Canada. J. Transport Health 7 (B), 264–268. https://doi.org/10.1016/j.jth.2017.09.005.

Garber, M.D., Watkins, K.E., Kramer, M.R., 2019. Comparing bicyclists who use smartphone apps to record rides with those who do not: Implications for representativeness and selection bias. J. Transport Health 15, 100661. https://doi.org/10.1016/j.jth.2019.10066.

Garrard, Jan, Rose, Geoffrey, Lo K, Sing, 2008. Promoting transportation for cycling for women: the role of bicycle infrastructure. Prevent. Med. 46 (1), 55–59. https://doi.org/10.1016/j.ypmed.2007.07.010.

Griffin, G.P., Jiao, J., 2015. Where does bicycling for health happen? Analyzing volunteered geographic information through place and plexus. J. Transport Health 2 (2), 238–247. https://doi.org/10.1016/j.jth.2014.12.001.

Griswold, J.B., Medury, A., Schneider, R.J., 2011. Pilot models for estimating bicycle intersection volumes. Transp. Res. Rec. 2247 (1), 1–7. https://doi.org/10.3141/2247-01.

Hankey, S., Lindsey, G., Wang, X., Borah, J., Hoff, K., Utecht, B., Xu, Z., 2012. Estimating use of non-motorized infrastructure: Models of bicycle and pedestrian traffic in Minneapolis, MN. Landscape Urban Planning 107 (3), 307–316. https://doi.org/10.1016/j.landurbplan.2012.06.005.

Hauer, E., 1995. On exposure and accident rate. Traffic Eng. Control 36 (3), 134–138.

Hochmair, H.H., Bardin, E., Ahmouda, A., 2019. Estimating bicycle trip volume for Miami-Dade County from Strava tracking data. J. Transp. Geogr. 75, 58–69. https://doi.org/10.1016/j.jtrangeo.2019.01.013.

Hood, J., Sall, E., Charlton, B., 2011. A GPS-based bicycle route choice model for San Francisco, California. Transport. Lett. 3 (1), 63–75. https://doi.org/10.3328/TL.2011.03.01.63-75.

Hyde-Wright, A., Graham, B., Krista Nordback, P.E., 2014. Counting bicyclists with pneumatic tube counters on shared roadways. Inst. Transport. Eng. ITE J. 84 (2), 32–37.

Ilieva, R.T., McPhearson, T., 2018. Social media data for urban sustainability. Nat. Sustain. 1 (10), 553–565. https://doi.org/10.1038/s41893-018-0153-6.

Jestico, B., Nelson, T., Winters, M., 2016. Mapping ridership using crowdsourced cycling data. J. Transp. Geogr. 52, 90–97. https://doi.org/10.1016/j.jtrangeo.2016.03.006.

Menghini, G., Carrasco, N., Schussler, N., Axhausen, K., 2009. Route choice of cyclists: discrete choice modeling based on GPS-data. Transport. Res. Part A: Policy Pract. 44 (9), 754–765. https://doi.org/10.1016/j.tra.2010.07.008.

Miranda-Moreno, L.F., Strauss, J., Morency, P., 2011. Disaggregate exposure measures and injury frequency models of cyclist safety at signalized intersections. Transp. Res. Rec. 2236 (1), 74–82. https://doi.org/10.3141/2236-09.

Moudon, A.V., Lee, C., Cheadle, A.D., Collier, C.W., Johnson, D., Schmid, T.L., Weather, R.D., 2005. Cycling and the built environment, a US perspective. Transport. Res. Part D: Transport Environ. 10 (3), 245–261. https://doi.org/10.1016/j.trd.2005.04.001.

Nelson, T.A., Denouden, T., Jestico, B., Laberee, K., Winters, M., 2015. BikeMaps.org: A global tool for collision and near miss mapping. Frontiers. Public Health 3, 53. https://doi.org/10.3389/fpubh.2015.00053.

Noland, R.B., Deka, D., Walia, R., 2011. A statewide analysis of bicycling in New Jersey. Int. J. Sustain. Transport. 5 (5), 251–269. https://doi.org/10.1080/15568318.2010.501482.

Noland, R.B., Klein, N.J., Tulach, N.K., 2013. Do lower income areas have more pedestrian casualties? Accid. Anal. Prev. 59, 337–345. https://doi.org/10.1016/j.aap.2013.06.009.

OpenStreetMap contributors, 2019. https://www.openstreetmap.org (accessed 29 May 2019).

Pikora, T., Giles-Corti, B., Bull, F., Jamrozik, K., Donovan, R., 2003. Developing a framework for assessment of the environmental determinants of walking and cycling. Soc. Sci. Med. 56 (8), 1693–1703. https://doi.org/10.1016/S0277-9536(02)00163-6.

Porter, A.K., Kontou, E., McDonald, N., Evenson, K.R., 2020. Perceived barriers to commuter and exercise bicycling in US adults: The 2017 National Household Travel Survey. J. Transport Health 16, 100820. https://doi.org/10.1016/j.jth.2020.100820.

Pucher, J., Dill, J., Handy, S., 2010. Infrastructure, programs, and policies to increase bicycling: an international review. Prev. Med. 50, S106–S125. https://doi.org/10.1016/j.ypmed.2009.07.028.

Roy, A., Nelson, T.A., Fotheringham, A.S., Winters, M., 2019. Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists. Urban Sci. 3 (2), 62. https://doi.org/10.3390/urbansci3020062.

Saelens, B.E., Sallis, J.F., Frank, L.D., 2003. Environmental correlates of walking and cycling: findings from the transportation, urban design, and planning literatures. Ann. Behav. Med. 25 (2), 80–91. https://doi.org/10.1207/S15324796ABM2502_03.

Sallis, J.F., Bowles, H.R., Bauman, A., Ainsworth, B.E., Bull, F.C., Craig, C.L., et al., 2009. Neighborhood environments and physical activity among adults in 11 countries. Am. J. Prev. Med. 36 (6), 484–490. https://doi.org/10.1016/j.amepre.2009.01.031.

Sallis, J.F., Floyd, M.F., Rodríguez, D.A., Saelens, B.E., 2012. Role of built environments in physical activity, obesity, and cardiovascular disease. Circulation 125 (5), 729–737, 10.1161/CIRCULATIONAHA.110.969022Circulation.

Sallis, J.F., Conway, T.L., Dillon, L.I., Frank, L.D., Adams, M.A., Cain, K.L., Saelens, B.E., 2013. Environmental and demographic correlates of bicycling. Prev. Med. 57 (5), 456–460. https://doi.org/10.1016/j.ypmed.2013.06.014.

Sanders, R.L., Frackelton, A., Gardner, S., Schneider, R., Hintze, M., 2017. Ballpark method for estimating pedestrian and bicyclist exposure in Seattle, Washington: Potential option for resource-constrained cities in an age of big data. Transp. Res. Rec. 2605 (1), 32–44. https://doi.org/10.3141/2605-03.

Spencer, P., Watts, R., Vivanco, L., Flynn, B., 2013. The effect of environmental factors on bicycle commuters in Vermont: influences of a northern climate. J. Transp. Geogr. 31, 11–17.

Statistics Canada, 2017. Census Profile, 2016 Census. Journey to Work data (accessed 31 January 2020).

Strauss, J., Miranda-Moreno, L.F., Morency, P., 2013. Cyclist activity and injury risk analysis at signalized intersections: A Bayesian modelling approach. Accid. Anal. Prev. 59, 9–17. https://doi.org/10.1016/j.aap.2013.04.037.

Strava.com, 2019. Strava releases 2019 Year in Sport Data Report. https://blog.strava.com/press/strava-releases-2019-year-in-sport-data-report/ (accessed 31 Dec 2019).

Sun, Y., Du, Y., Wang, Y., Zhuang, L., 2017. Examining associations of environmental characteristics with recreational cycling behavior by street-level Strava data. Int. J. Environ. Res. Public Health 14 (6), 644. https://doi.org/10.3390/ijerph14060644.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc.: Ser. B (Methodol.) 58 (1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Vanparijs, J., Panis, L.I., Meeusen, R., de Geus, B., 2015. Exposure measurement in bicycle safety analysis: A review of the literature. Accid. Anal. Prev. 84, 9–19. https://doi.org/10.1016/j.aap.2015.08.007.

Winters, M., Brauer, M., Setton, E.M., Teschke, K., 2010. Built environment influences on healthy transportation choices: Bicycling versus driving. J. Urban Health 87, 969–993.

Winters, M., Zanotto, M., 2019. The Canadian Bikeway Comfort and Safety (Can-BICS) Classification System: A Proposal for Developing Common Naming Conventions for Cycling Infrastructure. Vancouver, BC.

Yu, C.-Y., 2014. Environmental supports for walking/biking and traffic safety: Income and ethnicity disparities. Prev. Med. 67, 12–16. https://doi.org/10.1016/j.ypmed.2014.06.028.