Quantifying changes in bicycle volumes using crowdsourced data

ABSTRACT

Most cities in the United States lack comprehensive or connected bicycle infrastructure, therefore, inexpensive and easy-to-implement solutions for connecting existing bicycle infrastructure are increasingly being employed. Signage is one of the promising solutions. However, the necessary data for evaluating its effect on cycling ridership is lacking. To overcome this challenge, this study tests the potential of using readily-available crowdsourced data in concert with machine-learning methods to provide insight into signage intervention effectiveness. We do this by assessing a natural experiment to identify the potential effects of adding or replacing signage within existing bicycle infrastructure in 2019 in the city of Omaha, Nebraska. Specifically, we first visually compare cycling traffic changes in 2019 to those from the previous two years (2017-2018) using data extracted from the Strava fitness app. Then, we use a new three-step machine-learning approach to quantify the impact of signage while controlling for weather, demographics, and street characteristics. The steps are as follows: Step 1 (modeling and validation) build and train a model from the available 2017 crowdsourced data (i.e., Strava, Census, weather) that accurately predicts the cycling traffic data for any street within the study area in 2018; Step 2 (prediction) use the model from Step 1 to predict bicycle traffic in 2019 while assuming new signage was not added; Step 3 (impact evaluation) use the difference in prediction from actual traffic in 2019 as evidence of the likely impact of signage. While our work does not demonstrate causality, it does demonstrate an inexpensive method, using readily available data, to identify changing trends in bicycling over the same time that new infrastructure investments are being added.

Keywords: Cycling, Strava, Cycling Infrastructure, Bicycle Signage, Sharrows, Machine Learning.

INTRODUCTION

Communities are increasingly interested in promoting bicycling for their potential health and sustainability benefits. For example, multiple studies showed that those who ride bikes are less likely to have diabetes or hypertension (Beura et al., 2020; Dill, 2009; Fishman, 2016a; Fishman, 2016b; Hong et al., 2019; Pettit et al., 2016) and have a lower illness related to sedentary behaviors (Kriit et al., 2019) compared to people who shuttle in motor vehicles. Moreover, cycling helps to address traffic congestion and air pollution (Ogilvie et al., 2004). Despite the myriad benefits of bicycling, policy and funding in the United States prioritizes auto-infrastructure over all else, leaving bicycle infrastructure underfunded and difficult to implement (Piatkowski, Daniel, and Marshall, 2018). The effects of this "system of automobility" (Urry, 2004) have left most cities in the US lacking the high-quality, connected cycling infrastructure necessary to make bicycling convenient and safe (Moudon et al., 2005). To put things in perspective, cyclists are more likely to experience injuries or death from accidents compared to motor vehicle drivers (Jeroen Johan de Hartog et al., 2010).

A safe, connected, and easy-to-navigate bicycle infrastructure system is essential to fostering cycling in any city (Buehler and Dill, 2016; Dill, 2009; Goodman et al., 2014; Hull and O'Holleran, 2014; Larsen et al., 2013). However, given the systemic pressures blocking investments in bicycling in the US, many cities lack the political capacity and motivation (compared to auto infrastructure interest) to invest in bicycle infrastructure (Dill et al., 2017; Robartes et al., 2021). This problem is amplified since typical

bicycle infrastructure such as bike lanes or cycle tracks may require significant changes to existing road geometries (Buehler and Dill, 2016) or involve contentious changes such as removing on-street parking (Piatkowski, Daniel P., et al., 2019).

Because of these challenges, inexpensive and less-complex infrastructure solutions, such as adding bicycle signage or painting sharrows are favored (Dill et al., 2017). Such improvements to bicycle networks can utilize existing infrastructure in new ways by guiding cyclists and automobile drivers. However, these modest infrastructure investments lack evaluation of their impacts. For example, while bicycle signage is ubiquitous in the US, its impact is unclear (Bopp et al., 2018). Thus, there is an urgent need to understand the impacts of modest, low-cost cycling interventions to guide evidence-based planning decisions. As such, this paper addresses two gaps in the literature: (i) demonstrating an evaluation method using readily-available data and machine-learning to identify changes in bicycle volumes on specific corridors pre-and-post modest interventions while controlling for multiple relevant external factors (Heesch and Langdon, 2016; McArthur and Hong, 2019; Rodriguez-Valencia et al., 2019); and (ii) improving applied research for quantifying effects of modest investments on cycling rates over time (Bopp et al., 2018).

To fulfill our research aims, this work takes advantage of a natural experiment to quantify the effect of adding signage to existing infrastructure, or to connect disparate bicycle infrastructure. It is important to note that we are not claiming signage to be an ideal solution for infrastructure connectivity, but seek to demonstrate whether adding signs may be better than no infrastructure changes. The City of Omaha, Nebraska (USA) assigned 24 miles of bicycle routes in March 2019 with new signage. The added bicycle signage was distributed in the city center near major universities in the area. In our analysis, we use bicycle-volume data collected via the Strava smartphone app for Omaha from January 2017, to December 2019, and the cycling infrastructure location data provided by the regional planning authority, Metropolitan Area Planning Agency (MAPA). To add more dimensions to our analysis, we use weather, OpenStreetMap (OSM), and demographics data.

The organization of this paper is as follows. First, we review the literature related to our work. Then, we identify the relevant data sources used in our analysis for bike traffic in Omaha, Nebraska. Next, to confirm the representation of Strava data in the actual Omaha bike traffic, we provide a correlation analysis between Strava and several bike-counts data sources in Omaha. Afterward, we highlight the changes in bicycle infrastructure in Omaha within the last three years. Finally, we use a predictive machine learning approach trained by Strava data to quantify the effect of newly added signage on cycling traffic, while controlling for potentially confounding factors like weather and street characteristics. While this work is not able to demonstrate causality, our data-driven approach can be utilized by researchers for assessing correlation over time between newly added infrastructure and cycling volumes. Also, it provides a cost-effective approach for analyzing big cycling data to understand the factors associated with cycling volumes in a specific location.

LITERATURE REVIEW

This study is conducted in Omaha, Nebraska. The city and the region (in the Midwestern United States) lag behind much of the US in terms of bicycling mode share, the extent of major bicycle infrastructure, and investments in bicycling (The League of American bicyclists, 2019). As such, low-cost infrastructure additions such as sharrows and signage are the most common in the study area (Fig.1). Sharrows are solid white marked lines that are usually painted on roads or pavements. However, unlike bike lanes, sharrows do not give exclusive portions of the road for cycling purposes, and only indicate that

bicyclists are permitted on this lane. Of all forms of bicycle infrastructure, signage is considered the least sophisticated infrastructure, as it only informs cyclists and drivers that cyclists are permitted using roadside signs. While the effectiveness of these types of cycling infrastructure can vary widely, there is some consensus in the literature that the ability of bicycle infrastructure to increase cycling depends on the convenience, quality, connectivity (Dill, 2009; Fishman et al., 2015; Fishman, 2016a; McArthur and Hong, 2019; Meuleners et al., 2019; Pucher et al., 2009; Yang et al., 2019), and the proximity of cycling infrastructure to where people live (McArthur and Hong, 2019; Rodriguez-Valencia et al., 2019).

Figure 1: An example of on-street infrastructure types (Omaha, NE). Sharrows and signage are considered the lowest cost types of bike infrastructure.

Sharrows and signage have gained prominence for their low cost and ease of implementation (Weigand et al., 2013). Planners install sharrows to inform drivers to slow down and expect cyclists riding on the street, and to direct bicyclists to move towards the center of the travel lane away from parked cars (Buehler and Dill, 2016; DiGioia et al., 2017; Kovacs, 2017; Pucher and Buehler, 2016). However, the use of sharrows has been controversial, and there is conflicting evidence concerning the safety hazards of this type of infrastructure. On one hand, some researchers found that sharrows decrease cycling injuries on the roads compared to not using cycling infrastructure (Kovacs, 2017). On other hand, other researchers argue that more cyclist injuries occur on streets after sharrows intervention (Wall et al., 2016).

Signs are placed to inform cyclists of preferred routes designated by city planners but do not provide separate lanes or on-street lane markers like sharrows. Signed routes are commonplace, but there is limited research concerning the impacts of this type of infrastructure on cycling routes (Bopp et al., 2018; Buehler and Dill, 2016; Pucher et al., 2009). One study showed a positive correlation between signed routes and

better opinions that cycling infrastructure was of good quality (Sener et al., 2009). Another study showed that cyclists preferred residential roads with signed routes compared to not having signs at all (Abraham et al., 2002).

In general, evaluating the effect of bicycle infrastructure such as sharrows and signage on ridership presents several challenges for planners, who typically rely on low-quality data. Common examples include the US Census or the National Household Travel Survey, which are systematically biased against bicycling (Handy et al., 2014; Krizek et al., 2009). Moreover, those traditional data sources lack temporal and spatial details (Boss et al., 2018). Thus, they cannot capture the effect of bicycle interventions over a large area or an extended period (Heesch and Langdon, 2016). To supplement low-quality population data provided by national datasets, communities have also used bicycle-count data and live-point data. Live-point data are collected on intersections using cameras on traffic lights, counting stations, or sensors. Journey data provides information about the origin and destination of a trip, but it does not provide trip details. Instead, trip details can be collected in real-time from bikeshare programs. However, such data only provide information within the bikeshare area (Romanillos et al., 2016). To address these limitations, a "big data" approach has recently emerged to evaluate cycling infrastructure.

Big cycling data is richer than bicycle count data and includes data collected using live-point data, journey data, bikeshare programs, and GPS. For example, social fitness network companies collect data from their users. These GPS data are very detailed and spatially accurate and represent a good sample of the total population of cyclists (Helleland, 2017; Pritchard, 2018; Rogers and Papanikolopoulos, 2000; Romanillos et al., 2016). One example of a social fitness company is Strava. Strava app data contains a vast amount of spatial and temporal details to predict cycling trip patterns and provides a good approximation of the most-used routes and the peak months and times. To protect privacy, the Strava dataset reports trips using data based on street segments or intersections without disclosing each individual's origin and destination. While a portion of cyclists may use Strava to log their cycling trips, the app rarely tracks trips for users of other transportation modes such as driving a car (Fishman, 2016b; Helleland, 2017; Pettit et al., 2016; Romanillos et al., 2016). In our analysis, the occurrence of such data was rare and was removed manually before analysis.

The Strava big cycling data is collected voluntarily by Strava app users and hence is essentially self-reported survey data (i.e., biased based on the subset of the population who self-select to use cycling apps). A growing body of research, however, has addressed this potential bias (Conrow et al., 2018; Livingston et al., 2020; Roy et al., 2019; Watkins et al., 2016). Their findings indicate that Strava data is a robust proxy for actual bicycling, with studies identifying significant correlations between the Strava data and the ground-truth data obtained from counting stations (Boss et al., 2018; Hong et al., 2019; Jestico et al., 2016).

The literature also demonstrates that the usefulness of Strava data is not limited to analyzing and predicting the temporal and spatial behavior of cyclists, but is also good for evaluating the effect of adding new cycling infrastructure (Heesch and Langdon, 2016; Hong et al., 2019). Thus, this paper demonstrates a new multiple-phase machine learning method that learns from big data that includes Strava data to quantify the impact of added inexpensive cycling infrastructure in Omaha. Like other uses of Strava data, (Colorado Department of Transportation, 2018; Villamagna et al., 2019) findings from this paper can be used by urban planners to improve cycling infrastructure. Moreover, combining other data sources with Strava crowdsourced data strengthens the data interpretability by adding more spatial and temporal context to the crowdsourced data analysis.

Finally, it is worth mentioning that besides bicycling infrastructure, several additional factors affect bicycling. Those include weather (Zhao et al., 2018; Zhao et al., 2019), socio-demographics (e.g., population, gender, age, employment, and education) (Hochmair et al., 2019; Jestico et al., 2016; Musakwa and Selala, 2016; Romanillos and Gutiérrez, 2020), and street characteristics (Hochmair et al., 2019; McArthur and Hong, 2019; Orellana and Guerrero, 2019; Sultan et al., 2017). Accounting for those factors in our machine learning model approach adds a new perspective to the body of research on using Strava data.

DATA SOURCES

Quantifying cycling volumes, particularly temporal and spatial variations, requires multiple data sources. To this end, we acquire data from OSM, the City of Omaha, and MAPA. We also use weather data extracted from the National Oceanic and Atmospheric Association (NOAA). Finally, we use census data to understand the sociodemographic effects on cycling. Each of these data sources, over three years, is combined with Strava data to answer our research questions. Weather, census, and OSM data are free to download. Thus, this research provides the ability to utilize and combine many data sources at a relatively low cost, which could be beneficial for other researchers and communities to use. Next, we briefly describe the data sources we use in this project.

Strava Metro Data

Strava is a company that provides a phone fitness application for cyclists. The application, at the time of writing, is the most popular application of its kind designated for cyclists (Lee and Sener, 2021). It provides the ability to record a user's journey data, including statistics about their performance, speed, and distance traveled. Also, Strava challenges its app users in the same area by showing rankings of the highest performing people. The company may provide the data for researchers working on transportation related projects. The Strava data are categorized into hourly, monthly, and yearly data. For privacy reasons, Strava records the trips based on edges (street segments between intersections), not actual individual trips. Thus, it provides cycling count volumes across each edge. Furthermore, it hides a portion of the data if a specific edge has low cycling traffic (less than three cyclists) within the used timeframe (hours, months, or years). This rule explains why some trips may appear in the monthly data, but due to low traffic volume, might not appear in the hourly data. Finally, every street segment comes with an ID that can be joined with geographic datasets from OSM to provide the ability to create maps. OSM provides a variety of spatial information that can be utilized for a better understanding of street characteristics and land use, as explained next.

OpenStreetMap (OSM)

OSM is an open-source map built by a community of mappers that maintains spatial information about roads and land use in many places all over the world (Haklay and Weber, 2008; OpenStreetMap, 2008). In this paper, we use OSM data (Haklay and Weber, 2008) to create maps that display cycling data to extract road classes (e.g., major, minor, service, or suitable for walking and cycling) and characteristics (e.g., one-way street and maximum speed permitted). We incorporated this in the model by combining the OSM's geographic dataset with Strava for each street segment in the study area.

Omaha-Council Bluffs Metropolitan Area Planning Agency (MAPA) Data

MAPA provides existing cycling infrastructure maps and the treatment events between 2017 – 2019 (MAPA, 2020). Treatment events include infrastructure updates, such as adding or removing bicycling infrastructure and are tagged with their geographic locations and update dates. To use the data, we manually intersected and superimposed the location of each infrastructure item and event onto a single master map together with the Strava data.

Census Data

Census data provides demographic information (e.g., population count, cost of living, gender, employment rate, number of houses, and the median age in specific areas) at the zip code level. In this paper, we utilize such census data in building machine learning models that account for demographic effects on cycling. Zipcodes are postal codes that represent geographical areas and in this study are around 2 miles radius.

Weather Data

We extract weather data over the years (2017-2019) from the National Oceanic and Atmospheric Administration (NOAA). NOAA provides daily weather data that includes information about average temperature, wind speed, rain and snow precipitation, thunderstorms, and fog. We encoded weather data in our machine learning model as either continuous input (e.g., temperature) or binary (e.g., rain or no rain).

CORRELATION ANALYSIS

Concerns regarding bias in crowdsourced Strava data have proven to be largely unfounded. Multiple studies showed a strong correlation exists between Strava data and traditional bike-counts data (Boss et al., 2018; Hong et al., 2019; Jestico et al., 2016). The finding of those studies is particularly relevant to our research as Strava data is much easier to collect and access than creating and maintaining a city-wide bike-count program. However, to ensure robust findings, we first confirmed this finding for Omaha, as explained next.

Using the Shapiro-Wilk test (Shapiro and Wilk, 1965), we find that Strava data is not normally distributed in our sample region (the city of Omaha). Specifically, the ride counts per edge are skewed to the left. Similar findings were reported in the literature for other regions (Sun and Mobasheri, 2017). However, Strava data was also shown to be approximately normally distributed in some regions (Haworth, 2016). Therefore, for future research using Starva data, we recommend conducting a normality test of the Strava data whenever it is statistically required.

For the correlation study, we use Spearman's rank method, which deals with skewed data (Kutner et al., 2005), to study correlations between Strava data and available bike-count data from four different sites in Omaha. Table 1 shows a strong spearman's rank coefficient between weekly Strava and counters data in Omaha during the years 2017-2019. While Spearman's rank does not assume linearity, the relationship between Strava and bike counters is a linear positive relationship, as shown in Figure 2 (i.e., Strava counts increase with bike counter counts). In addition to Figure 2, the adjusted R^2 between Strava and bike counter data in Table 1 confirms the strong positive linear relationship between Strava and ground counter data. This strong correlation may justify the use of Strava data to capture the cycling trips in Omaha for places with no bike counts.

Table 1: Summary of counter – Strava correlation for several locations in Omaha

Counter Location	Spearman Coefficient	Adjusted R^2	Confidence Interval (95%)	Number of Data Points
Keystone Trail	0.93	0.85	[0.85- 1.0]	24
Big Papio Trail	0.87	0.80	[0.76 – 1.0]	16
Bob Kerry Bridge	0.87	0.73	[0.81 – 0.93]	93
Field Club	0.88	0.70	[0.79 – 0.96]	36

All Counters	0.89	0.66	[0.86 – 0.93]	169

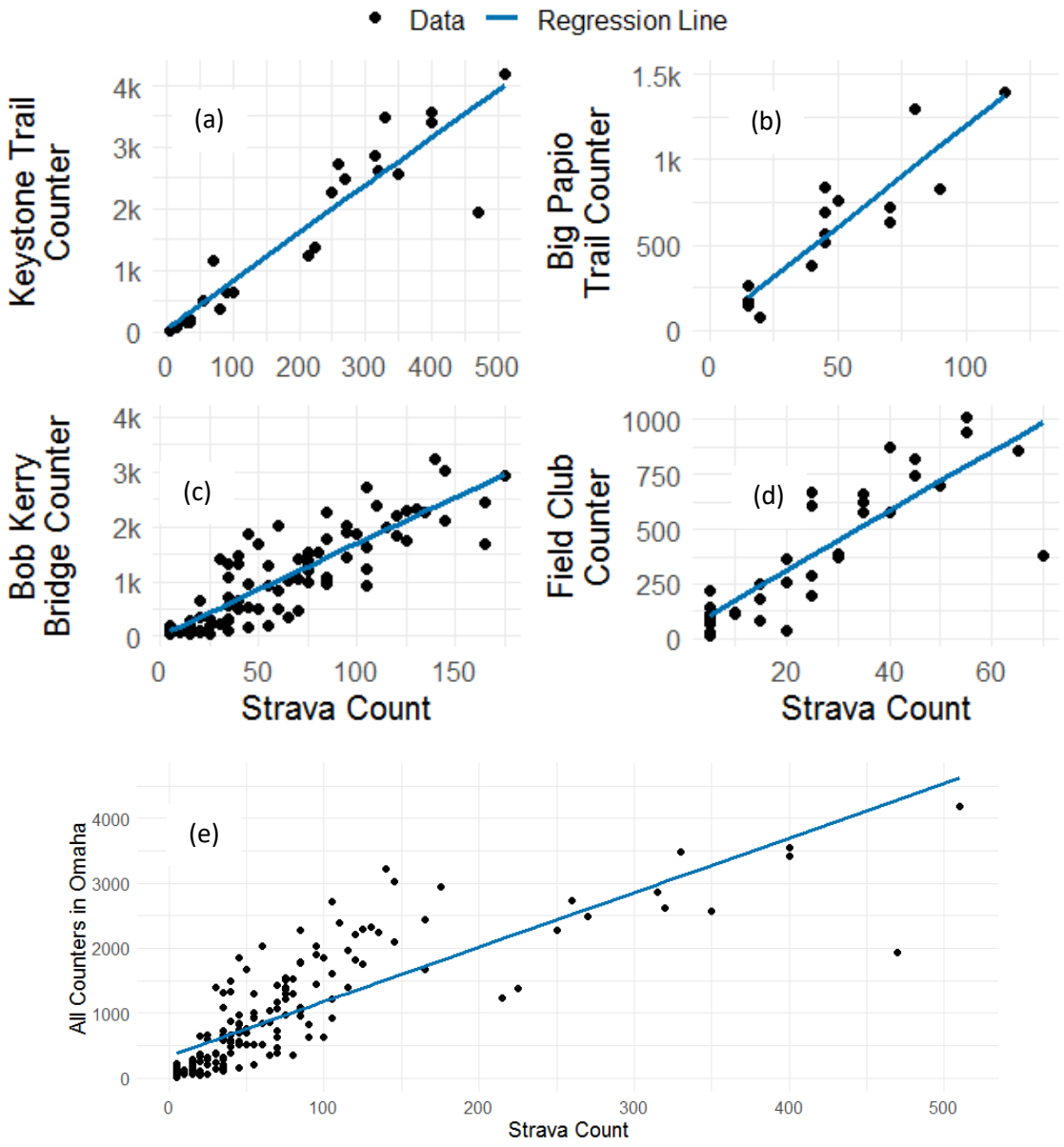

Figure 2: The relationship between Strava data and bike counter data on a weekly basis in four different locations in Omaha: a) Keystone Trail, b) Big Papio Trail, c) Bob Kerry Bridge, d) Field Club Trail, e) all counters combined.

MACHINE LEARNING ANALYSIS

This section presents the methods and major findings of our work. First, we provide a descriptive exploration of on-street infrastructure and bicycling volumes over the years 2017-2019 at selected sites in Omaha, which had major biking infrastructure changes within the study period. The second section presents the findings of our predictive analysis approach, using machine learning, to evaluate the impact of signage addition on cycling volumes in these sites.

Exploration of On-Street Infrastructure on Cycling: Lanes, Sharrows, and Signage

In total, on-street cycling infrastructure in Omaha includes 9 miles of bike lanes, 24 miles of bicycle routes with signage installed in March 2019, and 14 miles of sharrows. The bicycling infrastructure of interest in this study is concentrated in the city center near major universities. In this area, most of the bike lanes and sharrows were added before 2017. However, six miles of sharrows were removed and replaced by signage in March 2019. This provides an opportunity to study the effect of removing and replacing sharrows with signage. Signage was also added to streets that originally did not have any bicycle infrastructure and some streets with bike lanes. This provides another opportunity to take advantage of a natural experiment to study changes in bicycle traffic volumes over the same period and along the same corridors as changes were made to bicycle infrastructure. In Figure 3, we show the cycling infrastructure in this area in 2019 when signs were added.

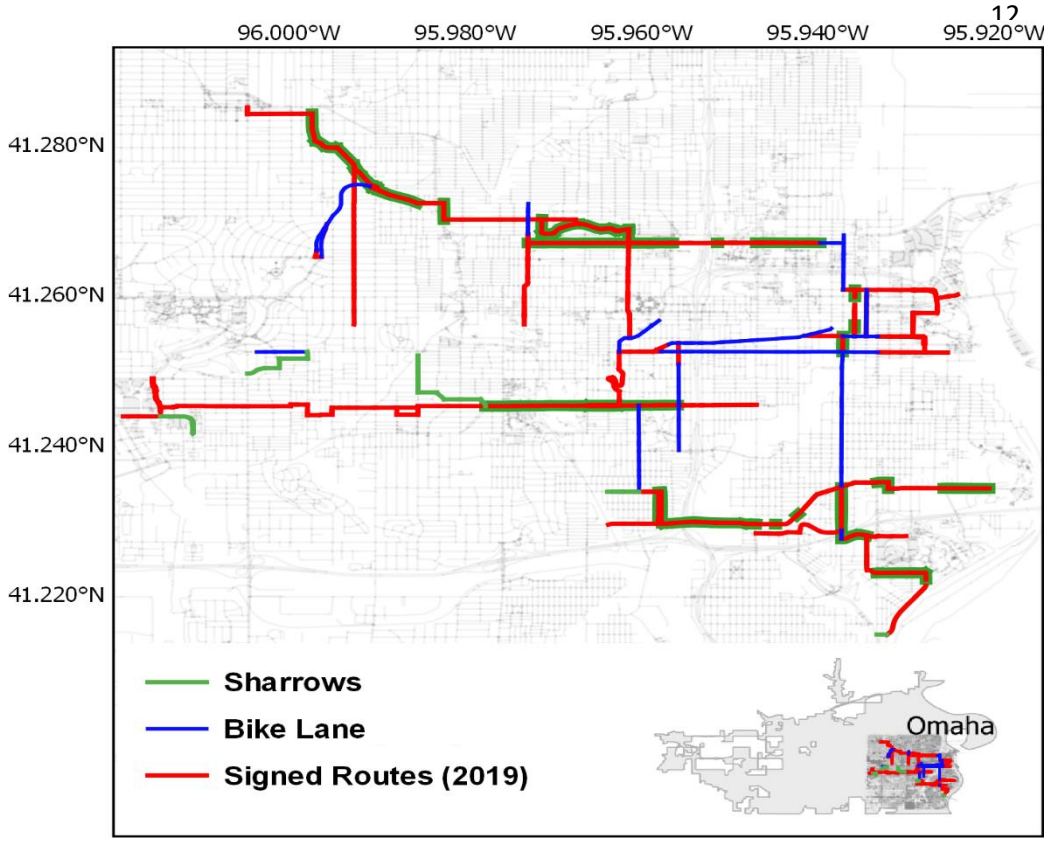

Figure 3: A map view showing the location of bicycle infrastructure in Omaha in 2019. The green line indicates the existence of sharrows. The blue line indicates the existence of bike lanes. The red line indicates the addition of signs in 2019. The overlap between the colors indicates the existence of multiple infrastructures, for example, red (signed routes) and green (sharrows) lines indicate that the street has both sharrows and signage.

Figure 4.a shows the normalized monthly total cycling traffic volume per street segment for streets that had sharrows but were replaced by signage in 2019. The figure covers the cycling volumes between 2017 and 2019. In this figure, for better visualization, the total cycling volumes are normalized to the maximum monthly total volume counts in 2017. This also helps to quantify the percentage change of cycling volumes in the following years. In this figure, the normalized total counts, rather than the absolute

total, is used to address the discrepancy in the number of streets for each type of infrastructure. Otherwise, using the total traffic will show higher numbers for bicycle traffic on the most dominant infrastructure compared to the rest of the infrastructure and may bias our analysis. Figure 4.a shows a significant increase in the normalized total cycling volumes in 2019. This effect may be associated with replacing sharrows with the new signage. In the next section, we test this hypothesis and quantify its impact by using machine learning models. For comparison purposes, Figure 4.b shows the normalized cycling monthly volume per street segment for streets that had sharrows that were not replaced with signage in 2019. The figure shows that for these streets, there was no increase in cycling volume. Interestingly, the figure shows a slight decrease in cycling volume in 2019. This decrease may be explained by the fact that these sharrows were located very near the newly added signed routes as shown in Figure 3. Thus, the newly added signed routes may have attracted some bikers who had been using these sharrow routes before the addition of the signage. Figure 4.c shows a tripling of cycling traffic when signs were added to streets with bike lanes, demonstrating the greater usability of routes that have signage combined with other types of infrastructure. Figure 4.d provides another piece of evidence of the effect of signages on cycling. By comparing cycling traffic volume before and after adding signs on streets with no infrastructure, we find that traffic increased after adding signage. While Figure 4 does not imply causality, there appears to be a correlation, since cycling traffic volume did not increase on all types of infrastructure in 2019, specifically, those without signage. For instance, cycling volumes on sharrows only decreased in 2019. Moreover, cycling traffic did not increase on bike lanes in 2018 (before adding signs in 2019). Finally, Figure 4 doesn't account for other factors, besides signage, that might affect cycling traffic such as weather and changes in street characteristics. Next, we explore the use of machine learning to filter those effects.

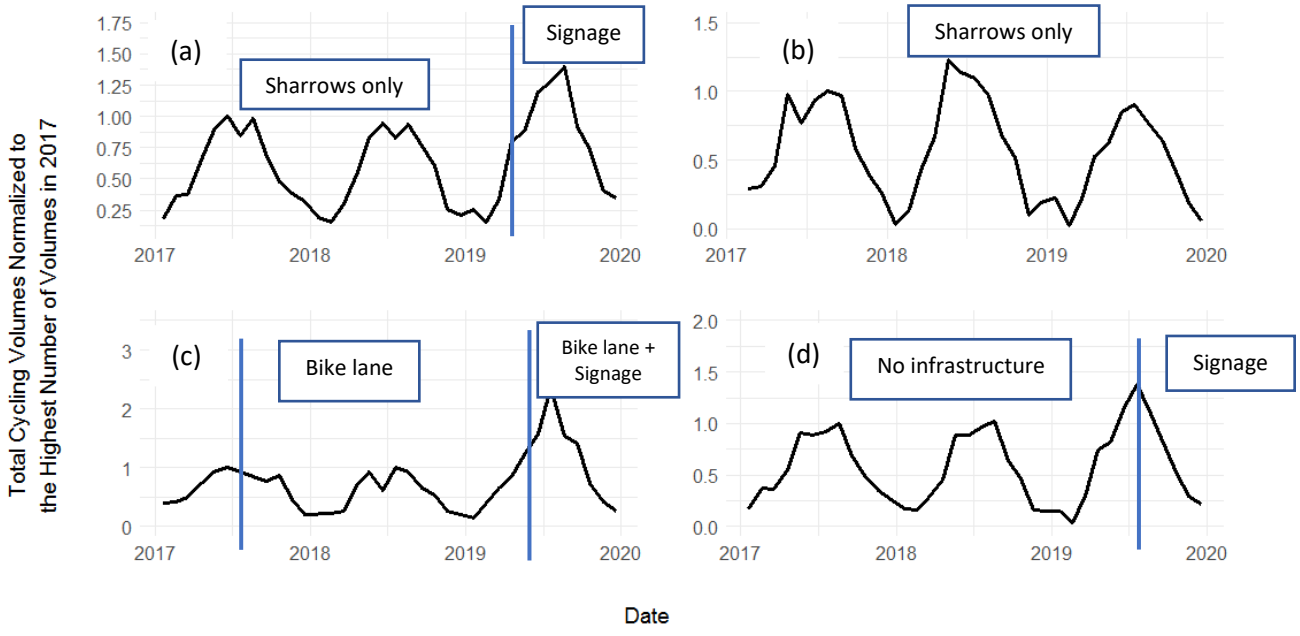

Figure 4: Total normalized monthly cycling volumes from 2017-2019 for streets with sharrows that a) were replaced by signed routes in March 2019 (12 miles), b) were not replaced (2 miles), c) streets with bike lanes that added signage in March 2019 (1 mile), d) streets with no infrastructure but that added signage in March 2019 (13 miles).

Identifying Changes in Bicycle Volumes Due to Added Signage using Machine Learning While Filtering other Factors

In this section, we provide a predictive analysis of the impact of changes in cycling volumes due to installing signage. We propose using machine learning algorithms because of their ability to account for many variables and nonlinearities that ordinary statistical models fail to handle. In this paper, we test several machine learning algorithms to predict cycling volumes on street segments across the study area. We also aim to identify correlations between changes in cycling volumes and the introduction of signage on streets in Omaha. Table 2 provides a brief description and the types of data used in the machine learning models discussed later in this section. It is worth mentioning that during modeling temporal data, like weather, the data was replicated for all street segments of the same period. Spatial data (i.e., OSM and census data) are all related to unique street segments and as such are replicated for each street segment every month. Moreover, OSM data is directly joined with Strava data. These datasets can be joined using a database management software like any SQL-based software or using a GIS software like QGIS and ArcGIS. Census data is replicated for each street segment within the appropriate zip code area for each month. Combining data sources in this way allows us to preserve the street segment (the spatial level of data provided by Strava) as the unit of analysis.

Table 2: Data Description

Feature	Type	Description	Data Source
Strava trips	Int	Monthly total number of cycling trips on streets segment	Strava
Edge ID	Int	Unique street segment ID	Strava
Month	int	Chronological months of the year	Strava
One-way street	VARCHAR	Is this a one-way road? “F” means that only driving in direction of the line string is allowed. “T”	OSM
Wind Speed	float	Average monthly wind speed	Weather
Rain precipitation	float	Total monthly rain precipitation	Weather
Snow precipitation	float	Total monthly Snow precipitation	Weather
Average Outside Temperature	float	Average monthly temperature	Weather
fog	int	Number of foggy days in each month	Weather
Heavy fog	int	Number of days that encountered heavy fog in each month	Weather
thunder	int	Number of days that encountered thunderstorms in each month	Weather
Ice pellets	int	Number of days that encountered ice pellets in each month	Weather
hail	int	Number of days that encountered hail in each month	Weather

rime	int	Number of days that encountered rime in each month	Weather
Drifting Snow	int	Number of days that encountered drifting snow in each month	Weather
Heavy rain	int	Number of days that encountered heavy rain in each month	Weather
Snowy Days	int	Number of Snowy days in each month	Weather
Snow	int	Number of days that snow was still precipitated in each month	Weather
fclass	VARCHAR	Street road category (major, minor, no car)	OSM
Season	VARCHAR	Spring, summer, fall, and winter	Weather
Population	int	Number of people living in each zip code area	Census
Population Density	int	Population Density of the zip code area	Census
Number of Houses	int	Number of houses in the zip code area	Census
Living Cost	int	Estimated living cost in each zip code area	Census
Unemployment Rate	float	Unemployment rate in the zip code area	Census
Commute Time	float	Estimated commute time in the zip code area	Census
Median Age	float	Estimated Median age in the zip code area	Census
Gender	int	Estimated number of females/males in the zip code area	Census

The three-step machine learning model is demonstrated in Figure 5. It starts by integrating cycling count data, street characteristics, weather, and demographic data into one database. Then, it separates the integrated data from 2017 and 2018 into a training dataset (2017 data: 5756 rows, 28 features), and a validation set (2018 data: 5499 rows, 28 features). The training dataset is used to train different machine learning models to predict monthly bike volumes per street segment, and the validation data is used to select the model that most accurately matches the 2018 data (Step 1). Next, the best model is used to predict the bike counts in 2019 (5644 rows, 28 features) assuming there was no new signage installed (Step 2). It is important to note that as bicycle signage was only added in 2019, the best model, trained and validated using only 2017 and 2018 data, does not account for this major intervention event. Thus, if the model predicts less traffic in 2019 than the actual traffic reported by Strava, a possible explanation is that the added bicycle signage, that was not accounted for in the model, may be associated with an increase in the bicycle volumes. In contrast, if the model predicts similar traffic volumes in 2019 to what the data shows, it could indicate that the added signage had minimal effect on cycling volumes (Step 3).

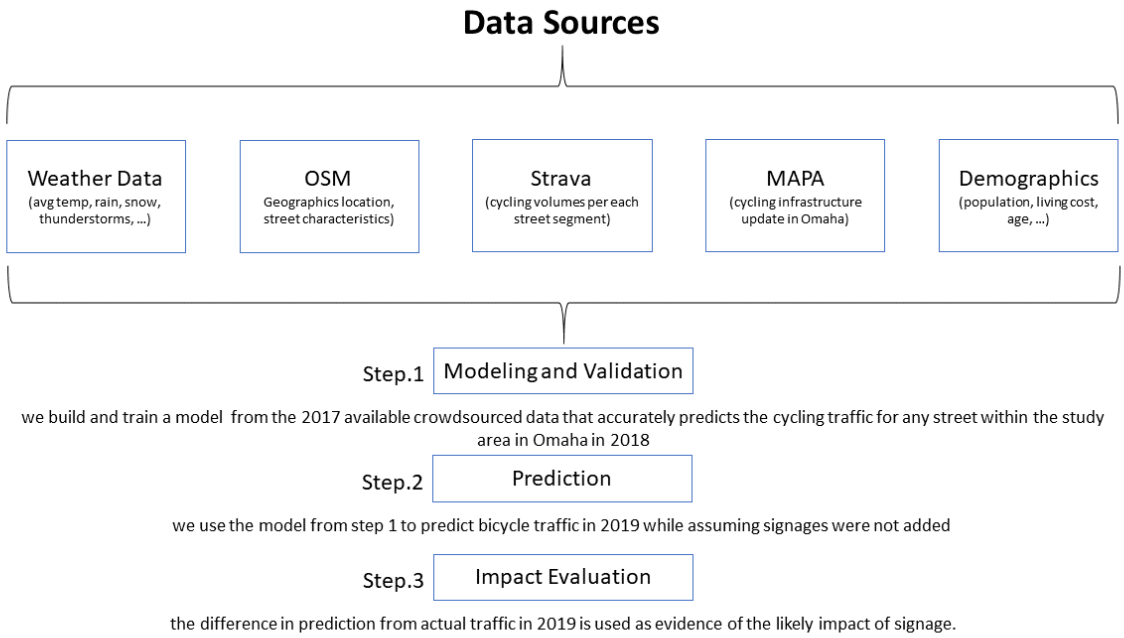

Figure 5: Flow chart demonstration of the 3-step machine learning analysis to predict Strava cycling volumes and signage impacts.

In step 1, multiple machine learning models (M. Stone, 1974) using the CARET Library in R (Kuhn, 2008) were trained from 2017 data. The machine learning methods include the Support Vector Machines (SVM) (Awad and Khanna, 2015), Random Forest (Liaw and Wiener, 2002), and XGBoost. The XGBoost comes in two versions: the Linear Booster and Tree Booster (Chen and Guestrin, 2016). It is noteworthy that these machine learning algorithms have been used in many application fields in the literature including cycling and transportation (Dadashova et al., 2020; Khasawneh et al., Apr 2020; Litzenberger et al., 2018; Mahmoud et al., 2021; Munira and Sener, 2020; Sun and Mobasheri, 2017). To compare their performance, we tuned each model parameter separately to minimize the mean absolute error (MAE) and to maximize R^2. Both MAE and R^2 are frequently used in assessing regression models (Mayer and Butler, 1993). After testing the trained models using the 2018 data, the XGBoost (linear) model showed the greatest accuracy, as shown in Figure 6, with its tuning parameters listed in Table 3. Finally, Figure 7.a confirms the high accuracy of the XGBoost model to predict the cycling volumes in 2018 while controlling for other factors such as weather, demographics, and street characteristics.

In step 2, the XGBoost model was used to predict the cycling volumes in 2019 as shown in the red plot in Figure 7.b. As mentioned before as the model was trained using 2017 data, before adding the signs in 2019, so the model assumes no signage was added. In step 3, the actual cycling volumes in 2019 (black plot in Figure 7.b) are compared to the prediction. Figure 7.b shows an increase of up to 56% in the actual cycling volume in 2019 (particularly during the summer peak in cycling) when compared to the model prediction. Thus, Figure 7 provides preliminary evidence that cycling volume increased from the baseline suggested by the machine learning model once signage was installed in the study area.

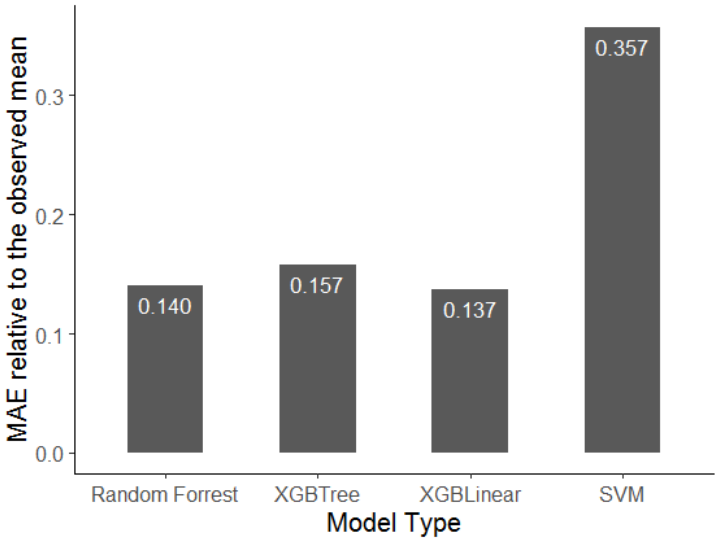

Figure 6: Error analysis of each model tested on the 2018 validation set showing that Xgboost (linear Booster) achieved the lowest error values.

Table 3: Tuning XGBoost Model Hyperparameters

Tuning Parameter	Function	Value
nrounds	Controlling the maximum number of iterations	600
lambda	to avoid overfitting	1
alpha	Feature selection	1
eta	Learning rate	0.3

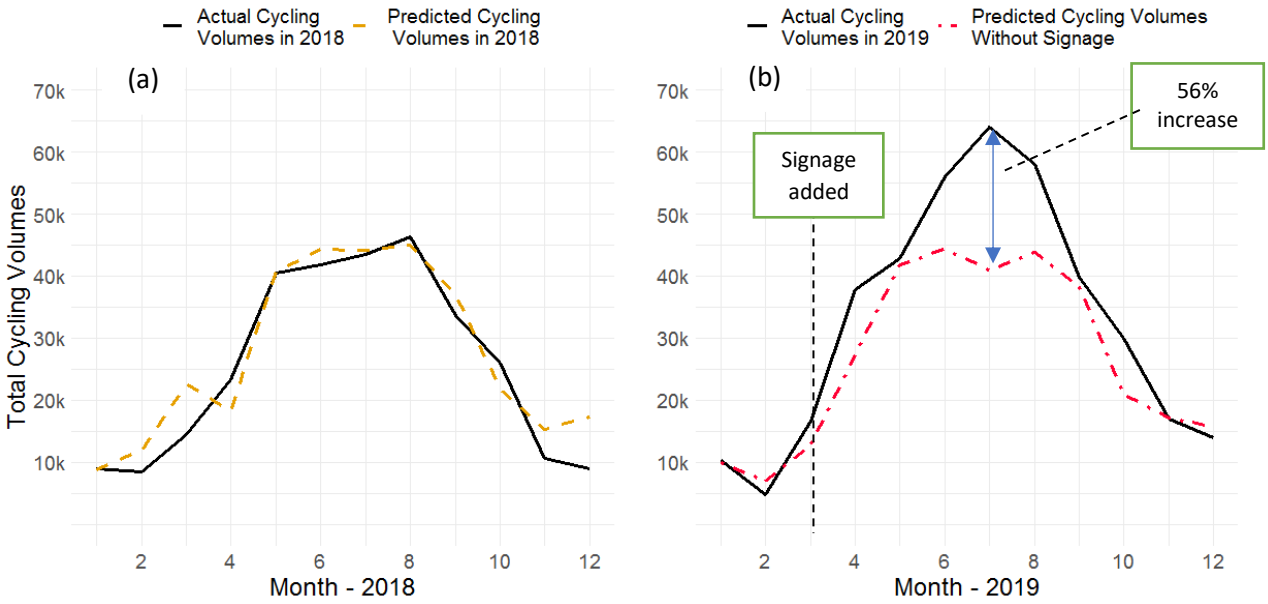

Figure 7: Comparison between 2018 and 2019 cycling volume predictions for all street types with added signage in 2019. a) accurate predictions of 2018 cycling volumes, b) increase in cycling volumes from prediction after installing signage.

DISCUSSION

This paper describes a proof-of-concept application of machine learning using crowdsourced data to estimate bicycle traffic volumes and identify correlations between said traffic volumes and investments in bicycle signage over the same period. Descriptive analysis in Figure 4 shows an increase in the total monthly bicycle count volumes on street segments in the study area that had new signage added in 2019 (Figure 4c illustrates that cycling volumes nearly tripled during peak months on street segments in which signage was installed). We also find exploratory evidence that signage may be useful in connecting existing infrastructure (Figure 4) because almost every street that had signs added in 2019 encountered an increase in cycling traffic. Results also suggest that while investments in signage may be related to bicycle traffic volumes, sharrows may not be (Figure 4.b).

To quantify changes in bicycle volumes that may be due to investments in signage while controlling for other parameters such as weather and demographics, we built a machine learning model that predicts cycling traffic at the street segment level. Results demonstrate that the model accurately predicted the 2018 cycling volumes (Figure 6). However, the model predicted fewer cycling volumes in 2019. This difference provides preliminary support for a correlational relationship between bicycling volume and signage that may require further study. While these findings are promising, they do not demonstrate causality; our observation period is brief and controls for other external factors are limited. For instance, our model does not account for promotional campaigns, biking events, or complex economic factors.

Our analysis also shows that the impact of weather and other spatial and demographic parameters (ranked by their importance in Figure 8) in cycling traffic prediction is significant and can't be ignored. However, we believe they deserve future study, since, in general, understanding feature importance is an unmatured, ongoing field of research in the machine learning community. Nevertheless, our work further confirms that cycling cannot be analyzed without studying temporal and spatial factors, particularly when trying to understand the effects of new infrastructure on bicycle volumes.

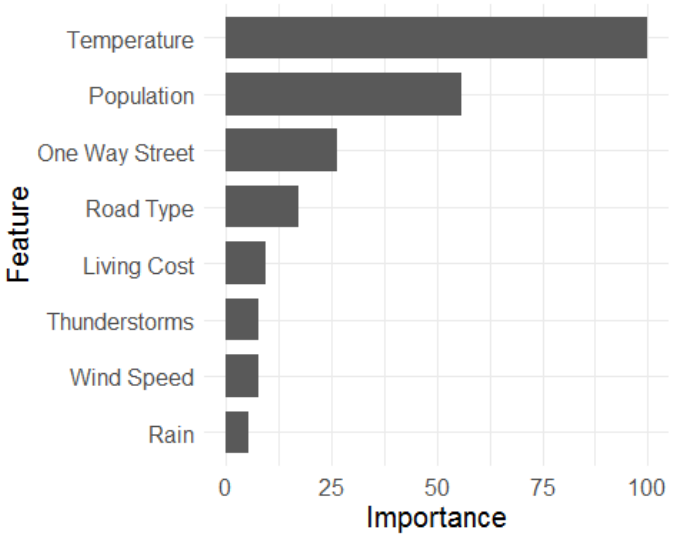

Figure 8: The most essential temporal and spatial factors affecting the prediction of cycling volume.

Finally, our analysis confirms that machine learning can handle complex datasets and can utilize nonlinearities to provide highly accurate predictions. The 3-phase machine learning approach we use in this study is highly flexible and can be used to predict how many kinds of changes (spatial or temporal) affect cycling volumes. For example, the same approach can be adapted to measure the effect of introducing other types of infrastructure such as bike lanes.

CONCLUSION

This research used a natural experiment to identify correlations between cycling volumes and new investments in signage at the street segment level. The study demonstrated the impact of bicycle signage on cycling volume using a machine learning model with three years of data from the fitness app, Strava. Specifically, in the third year of the study, bicycle signage was installed in much of the study area. By training the model to predict route bicycle volumes, we used the difference in predicted versus actual bicycling volume as a measure of the impact of signage on volume of route use. We found that cycling increased after installing the signs. Bicycle signage is not necessarily an exclusive piece of cycling infrastructure and is often used with other types of cycling infrastructure such as bike lanes. Whether alone or in concert with other infrastructure, signage may guide cyclists towards safer, higher-quality infrastructure, thereby providing a cost-effective means of assisting in the development of a connected bicycle infrastructure system.

This paper demonstrated a data-driven approach using readily available data and open-source software in assessing existing and newly added cycling infrastructure. The method used is not confined by a certain type of infrastructure, which means it can be used to monitor the effects of many spatial or temporal parameters on cycling volumes. But intervention evaluation is challenging. Demonstrating a causal relationship outside of a laboratory or clinical setting is nearly impossible, but communities must use whatever data is available to make evidence-based decisions about infrastructure investments. Despite these limitations, this work offers a crucial step forward in assessing data and methods for intervention evaluation in real-world settings. While it is impossible to make causal claims based on this work, we believe that future research may forward this goal.

Finally, while we have demonstrated significant correlations between bicycling rates and the installation of bicycle signage along specific corridors, additional research is still needed to understand and test these complex relationships to provide practical guidance to communities. For instance, we cannot generalize the predictors used in this study to all cycling locations, nor can we rule out additional controls not included in our analysis. Other parameters such as cycling events, promotional campaigns, and even broader social and economic factors should be studied to understand how they may affect bicycling at the city or corridor-specific level. Moreover, additional research requires a spatial autocorrelation study of the area. In future research, we recommend studying the long-term effects of this type of intervention and how it affects the neighboring areas in addition to studying the effects of each independent variable on cycling volume predictions. Continued advances in data availability, methods, and open-source software could allow communities significant access to the necessary data and tools to evaluate infrastructure investments and inform evidence-based policy and planning.

REFERENCES

2019 Bicycle Friendly State Report Cards. (2019) Available at: https://bikeleague.org/content/state-report-cards.

Abraham JE, McMillan S, Brownlee AT and Hunt JD (2002) Investigation of Cycling Sensitivities. In: *Transportation Research Board Annual Conference, Washington, DC*.

Awad M and Khanna R (2015) Support vector regression. *Efficient learning machines*: Springer, 67-80.

Beura SK, Kumar KV, Suman S and Bhuyan PK (2020) Service quality analysis of signalized intersections from the perspective of bicycling. *Journal of Transport & Health* 16: 100827.

Boss D, Nelson T, Winters M and Ferster CJ (2018) Using crowdsourced data to monitor change in spatial patterns of bicycle ridership. *Journal of Transport & Health* 9: 226-233.

Buehler R and Dill J (2016) Bikeway Networks: A Review of Effects on Cycling. *Transport Reviews* 36(1): 9-27.

Chen T and Guestrin C (2016) Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Colorado Department of Transportation (2018) Strava Metro Data Analysis Summary. Report. Colorado Department of Transportation, USA, June. Available at: https://www.codot.gov/programs/bikeped/documents/strava-analysis-summary_06-25-18.pdf

Conrow L, Wentz E, Nelson T and Pettit C (2018) Comparing spatial patterns of crowdsourced and conventional bicycling datasets. *Applied Geography* 92: 21-30.

Dadashova B, Griffin GP, Das S, Turner S and Sherman B (2020) Estimation of Average Annual Daily Bicycle Counts using Crowdsourced Strava Data. *Transportation Research Record* 2674(11): 390-402.

DiGioia J, Watkins KE, Xu Y, Rodgers M and Guensler R (2017) Safety impacts of bicycle infrastructure: A critical review. *Journal of Safety Research* 61: 105-119.

Dill J, Smith O and Howe D (2017) Promotion of active transportation among state departments of transportation in the US. *Journal of Transport & Health* 5: 163-171.

Ferenchak NN and Marshall W (2016) The Relative (in) Effectiveness of Bicycle Sharrows on Ridership and Safety Outcomes. In: *Transportation Research Board Annual Conference, Washington, DC*.

Fishman E, Böcker L and Helbich M (2015) Adult active transport in the Netherlands: An analysis of its contribution to physical activity requirements. *PloS One* 10(4): e0121871.

Fishman E (2016a) Bikeshare: A Review of Recent Literature. *Transport Reviews: Cycling as Transport* 36(1): 92-113.

Fishman E (2016b) Cycling as transport. *Transport Reviews: Cycling as Transport* 36(1): 1-8.

Goodman A, Sahlqvist S and Ogilvie D (2014) New Walking and Cycling Routes and Increased Physical Activity: One- and 2-Year Findings From the UK iConnect Study. *American Journal of Public Health* (1971) 104(9): e38-e46.

Haklay M and Weber P (2008) OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing* 7(4): 12-18.

Handy S, Van Wee B and Kroesen M (2014) Promoting cycling for transport: research needs and challenges. *Transport Reviews* 34(1): 4-24.

Haworth, J., 2016, June. Investigating the potential of activity tracking app data to estimate cycle flows in urban areas. In: *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* (Vol. 41, pp. 515-519). Copernicus Gesellschaft MBH.

Heesch KC and Langdon M (2016) The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour. *Health Promotion Journal of Australia* 27(3): 222-229.

Hochmair HH, Bardin E and Ahmouda A (2019) Estimating bicycle trip volume for Miami-Dade county from Strava tracking data. *Journal of Transport Geography* 75: 58-69.

Hong J, McArthur DP and Livingston M (2019) The evaluation of large cycling infrastructure investments in Glasgow using crowdsourced cycle data. *Transportation (Dordrecht)*: 1-14.

Hull A and O'Holleran C (2014) Bicycle infrastructure: can good design encourage cycling? *Urban, Planning and Transport Research* 2(1): 369-406.

Jennifer Dill (2009) Bicycling for Transportation and Health: The Role of Infrastructure. *Journal of Public Health Policy* 30(S1): S95-S110.

Jeroen Johan de Hartog, Hanna Boogaard, Hans Nijland and Gerard Hoek (2010) Do the Health Benefits of Cycling Outweigh the Risks? *Environmental Health Perspectives* 118(8): 1109-1116.

Jestico B, Nelson T and Winters M (2016) Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography* 52: 90-97.

Jestico, B., Nelson, T. and Winters, M., 2016. Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, 52, pp.90-97.

Helleland K (2017) *Big Cyclist Data*. Master Thesis. Aalborg University, Denmark.

Khasawneh N, Schulte C and Fraiwan M (Apr 2020) *A Framework for Crowd-Sourced Exercise Data Collection and Processing.* : IEEE.

Kriit HK, Williams JS, Lindholm L, Forsberg B and Nilsson Sommar J (2019) Health economic assessment of a scenario to promote bicycling as active transport in Stockholm, Sweden. *BMJ Open* 9(9): e030466.

Krizek KJ, Handy SL and Forsyth A (2009) Explaining changes in walking and bicycling behavior: challenges for transportation research. *Environment and Planning B: Planning and Design* 36(4): 725-740.

Kuhn M (2008) Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* 28(5).

Larsen J, Patterson Z and El-Geneidy A (2013) Build It. But Where? The Use of Geographic Information Systems in Identifying Locations for New Cycling Infrastructure. *International Journal of Sustainable Transportation* 7(4): 299-317.

Lee K and Sener IN (2020) Strava Metro data for bicycle monitoring: a literature review. *Transport Reviews*: 1-21.

Liaw A and Wiener M (2002) Classification and regression by randomForest. *R News* 2(3): 18-22.

Litzenberger S, Christensen T, Hofstätter O and Sabo A (2018) Prediction of Road Surface Quality during Cycling Using Smartphone Accelerometer Data. *Proceedings* 2(6): 217.

Livingston M, McArthur D, Hong J and English K (2020) Predicting cycling volumes using crowdsourced activity data. *Environment and Planning. B, Urban Analytics and City Science*: 239980832092582.

M. Stone (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)* 36(2): 111-147.

MAPA (2020) The Omaha-Council Bluffs Metropolitan Area Planning Agency (MAPA). Available at: https://mapacog.org/about/what-is-mapa/.

Mayer DG and Butler DG (1993) Statistical validation. *Ecological Modelling* 68(1-2): 21-32.

McArthur DP and Hong J (2019) Visualising where commuting cyclists travel using crowdsourced data. *Journal of Transport Geography* 74: 233-241.

Bopp M, Si D and Piatkowski D (2018) *Bicycling for Transportation*. US: Elsevier.

Mahmoud N, Abdel-Aty M, Cai Q and Zheng O (2021) Vulnerable road users' crash hotspot identification on multi-lane arterial roads using estimated exposure and considering context classification. *Accident Analysis and Prevention* 159: 106294.

Meuleners LB, Stevenson M, Fraser M, Oxley J, Rose G and Johnson M (2019) Safer cycling and the urban road environment: A case control study. *Accident Analysis and Prevention* 129: 342-349.

Moudon AV, Lee C, Cheadle AD, Collier CW, Johnson D, Schmid TL, et al. (2005) Cycling and the built environment, a US perspective. *Transportation Research. Part D, Transport and Environment* 10(3): 245-261.

Munira S and Sener IN (2020) A geographically weighted regression model to examine the spatial variation of the socioeconomic and land-use factors associated with Strava bike activity in Austin, Texas. *Journal of Transport Geography* 88.

Musakwa W and Selala KM (2016) Mapping cycling patterns and trends using Strava Metro data in the city of Johannesburg, South Africa. *Data in Brief* 9(C): 898-905.

Neter J, Wasserman W and Kutner MH (1985) *Applied Linear Statistical Models*. Homewood, Ill: Irwin.

Ogilvie D, Egan M, Hamilton V and Petticrew M (2004) Promoting walking and cycling as an alternative to using cars: systematic review. *Bmj* 329(7469): 763-766.

OpenStreetMap (2008) OpenStreetMap. Available at: https://www.openstreetmap.org/about.

Orellana D and Guerrero ML (2019) Exploring the influence of road network structure on the spatial behaviour of cyclists using crowdsourced data. *Environment and Planning. B, Urban Analytics and City Science* 46(7): 1314-1330.

Pettit CJ, Lieske SN and Leao SZ (2016) BIG BICYCLE DATA PROCESSING: FROM PERSONAL DATA TO URBAN APPLICATIONS. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* III-2: 173-179.

Piatkowski D and Marshall W (2018) We count what we care about: Advancing a framework for valuing investments in active modes. *Research in Transportation Business & Management* 29: 63-70.

Piatkowski DP, Marshall WE and Krizek KJ (2019) Carrots versus sticks: assessing intervention effectiveness and implementation challenges for active transport. *Journal of Planning Education and Research* 39(1): 50-64.

Pritchard R (2018) Revealed Preference Methods for Studying Bicycle Route Choice—A Systematic Review. *International Journal of Environmental Research and Public Health* 15(3): 470.

Pucher J and Buehler R (2016) Safer Cycling Through Improved Infrastructure. *AJPH*.

Pucher J, Dill J and Handy S (2009) Infrastructure, programs, and policies to increase bicycling: An international review. *Preventive Medicine* 50: S106-S125.

Robartes E, Chen E, Chen TD and Ohlms PB (2021) Assessment of local, state, and federal barriers to implementing bicycle infrastructure: A Virginia case study. *Case Studies on Transport Policy* 9(2): 488-496.

Rodriguez-Valencia A, Rosas-Satizábal D, Gordo D and Ochoa A (2019) Impact of household proximity to the cycling network on bicycle ridership: The case of Bogotá. *Journal of Transport Geography* 79: 102480.

Rogers S and Papanikolopoulos NP (2000). Counting bicycles using computer vision. In *ITSC2000. 2000 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 00TH8493)* (pp. 33-38). IEEE.

Romanillos G and Gutiérrez J (2020) Cyclists do better. Analyzing urban cycling operating speeds and accessibility. *International Journal of Sustainable Transportation* 14(6): 448-464.

Romanillos G, Zaltz Austwick M, Ettema D and De Kruijf J (2016) Big Data and Cycling. *Transport Reviews: Cycling as Transport* 36(1): 114-133.

Roy A, Nelson TA, Fotheringham AS and Winters M (2019) Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. *Urban Science* 3(2): 62.

Sener I, Bhat C and Eluru N (2009) An analysis of bicycle route choice preferences in Texas, US. *Transportation* 36(5): 511-539.

Shapiro SS and Wilk MB (1965) An Analysis of Variance Test for Normality (Complete Samples). *Biometrika* 52(3/4): 591.

Speck J (2018) Do Not Use Sharrows as Cycling Facilities. In *Walkable City Rules* (pp. 146-147). Island Press, Washington, DC.

Sultan J, Ben-Haim G, Haunert J and Dalyot S (2017) Extracting spatial patterns in bicycle routes from crowdsourced data. *Transactions in GIS* 21(6): 1321-1340.

Sun Y and Mobasheri A (2017) Utilizing Crowdsourced Data for Studies of Cycling and Air Pollution Exposure: A Case Study Using Strava Data. *International Journal of Environmental Research and Public Health* 14(3): 274.

Urry J (2004) The 'system'of automobility. *Theory, Culture & Society* 21(4-5): 25-39.

Villamagna A, Getts L and Young R (2019) *Active Transportation Accounting: Developing Metrics for Project Prioritization* (No. FHWA-NH-RD-26962R). Plymouth State University. Department of Environmental Science and Policy.

Wall SP, Lee DC, Frangos SG, Sethi M, Heyer JH, Ayoung-Chee P, et al. (2016) The Effect of Sharrows, Painted Bicycle Lanes and Physically Protected Paths on the Severity of Bicycle Injuries Caused by Motor Vehicles. *Safety* (Basel) 2(4): 26.

Watkins K, Ammanamanchi R, LaMondia J and Le Dantec CA (2016) Comparison of Smartphone-Based Cyclist GPS Data Sources. In: *Transportation Research Board Annual Conference, Washington, DC*.

Weigand L, McNeil N and Dill J (2013) Cost Analysis of Bicycle Facilities: Cases from Cities in the Portland, OR Region.

Weigand L, McNeil N and Dill J (2013) Cost analysis of bicycle facilities: Cases from cities in the Portland, or region. *FINAL DRAFT. Initiative for Bicycle & Pedestrian Innovation. Portland State University. Robert Wood Jonson Foundation*.

Yang Y, Wu X, Zhou P, Gou Z and Lu Y (2019) Towards a cycling-friendly city: An updated review of the associations between built environment and cycling behaviors (2007–2017). *Journal of Transport & Health* 14: 100613.

Zhao J, Guo C, Zhang R, Guo D and Palmer M (2019) Impacts of weather on cycling and walking on twin trails in Seattle. *Transportation Research Part D* 77: 573-588.

Zhao J, Wang J, Xing Z, Luan X and Jiang Y (2018) Weather and cycling: Mining big data to have an in-depth understanding of the association of weather variability with cycling on an off-road trail and an on-road bike lane. *Transportation Research Part A* 111: 119-135.