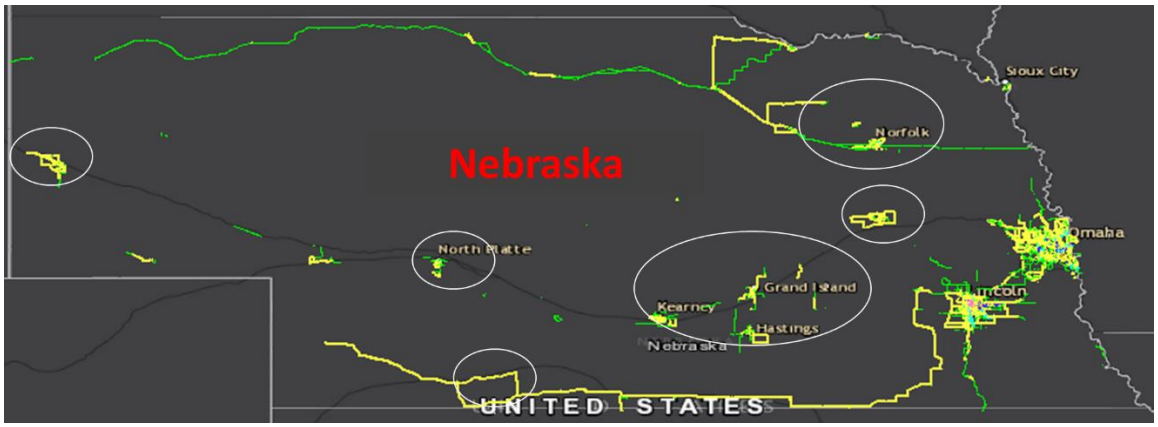
A Big Data Approach for Improving Nebraska Cycling Routes

Fadi Alsaleem, Ali Al-Ramini, Mohammad Ali Takallou, and Daniel Piatkowski.

Architectural Engineering and Construction Department of University of

Nebraska-Lincoln (Durham School)

FINAL REPORT



Sponsored By

Nebraska Department of Transportation and U.S. Department of Transportation Federal Highway Administration

December 2020

NEBRASKA
Good Life. Great Journey.
DEPARTMENT OF TRANSPORTATION

THE DURHAM SCHOOL
OF ARCHITECTURAL ENGINEERING AND CONSTRUCTION

1. Report No. M095	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle A Big Data Approach for Improving Nebraska Cycling Routes		**5. Report Date** December 2020 **6. Performing Organization Code**
7. Author(s) Fadi Alsaleem, Ali Al-Ramini, Mohammad Ali Takallou, and Daniel Piatkowski.		**8. Performing Organization Report No.** If applicable, enter any/all unique numbers assigned to the performing organization.
9. Performing Organization Name and Address University of Nebraska-Lincoln		**10. Work Unit No.** **11. Contract** SPR-P1(20) M095
12. Sponsoring Agency Name and Address Nebraska Department of Transportation Research Section 1400 Hwy 2 Lincoln, NE 68502		**13. Type of Report and Period Covered** Final Report July 2019 – December 2020 **14. Sponsoring Agency Code**

15. Supplementary Notes

If applicable, enter information not included elsewhere, such as translation of (or by), report supersedes, old edition number, alternate title (e.g. project name), or hypertext links to documents or related information.

16. Abstract

More people are becoming interested in creating healthy lifestyle habits for themselves. It has been proved that cycling has several health benefits. Therefore, governments are planning towards more cycling-friendly infrastructures and environments. To understand the cycling activities in Nebraska, we analyze Strava data for the last three years, providing valuable insight into cycling activities for urban planning purposes.

We analyze commute and recreational cycling patterns varying between weekdays and weekends to identify peak hours. Moreover, we study the effect of weather on cycling activity patterns for the last three years. Also, we use the impact of the outside temperature as a parameter to model the monthly cycling activities in Nebraska using polynomial regression. Furthermore, we analyze the spatial and demographic factors that influence cycling activities in Nebraska. Finally, we use machine learning regression to understand the effect of adding signed routes on cycling activities.

We found a strong association between Strava and counter data counting for all cyclists in a specific location. Also, we found a correlation between weather conditions and cycling, where the average outside temperature has the most significant effect. However, other factors influence cycling activities like rainfall, wind speed, thunderstorms, and fog in different extents. Weekends and weekdays showed different cycling patterns during peak hours, explained by the fact that most people cannot ride their bike in working hours. The spatial analysis showed that cycling is profoundly affected by the existence of cycling infrastructure. Also, it showed that trails are the most used for recreational activities.

Additionally, including bike lanes and signed routes to the street infrastructure are the most used for commute purposes. Moreover, the added bike lanes in Lincoln between 2017 and 2019 showed a significant rise in cycling activities. In Omaha, we showed a significant increase in cycling activities because of the installation of signed bicycle routes.

Key Words	18. Distribution Statement
Cycling, Strava, Cycling Infrastructure, Signed Routes, Bicycle Signage, Sharrows, Machine Learning.	No restrictions. This document is available through the National Technical Information Service. 5285 Port Royal Road Springfield, VA 22161

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	61	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. The contents do not necessarily reflect the official views or policies neither of the Nebraska Department of Transportations nor the University of Nebraska-Lincoln. This report does not constitute a standard, specification, or regulation. Trade or manufacturers' names, which may appear in this report, are cited only because they are considered essential to the objectives of the report.

The United States (U.S.) government and the State of Nebraska do not endorse products or manufacturers. This material is based upon work supported by the Federal Highway Administration under SPR-P1(M095). Any opinions, findings and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the Federal Highway Administration."

Acknowledgments

This project is funded by the Nebraska Department of Transportation (NDOT). Special thanks go to Ryan Huff, Craig Wacker, and David Hansen from NDOT, Julie Harris from Bike Walk Nebraska, and MAPA team for their valuable feedback and comments. Finally, special thanks for Lieska Halsey and Mark Fischer for organizing and attending our updated meetings.

Table of Contents

List of Figures

List of Tables

Introduction

1.1 Motivation

Cycling is becoming more popular, as people shift to adopt healthier methods of transportation. Also, the increase in cycling mobility has numerous environmental, social, and health benefits because it is a low cost and energy-efficient mode of transportation. While providing the opportunity for a healthy lifestyle, cycling poses a solution for congestion traffics, air pollution, and less injury risk to road users (1). However, it requires providing a safe infrastructure environment for cyclists. For instance, cyclists are more likely to experience injuries or even death in accidents compared to motor vehicle drivers (2).

Unfortunately, the infrastructure in the U.S. traditionally caters to automobile traffic creating impediments for bikers and impacting their safety. To accommodate cycling, a significant challenge is the lack of data to assess the attributes of existing assets accurately and to inform additional investments to integrate bicycles into our transportation system. Traditionally, access to high-quality data has limited our understanding of cycling behavior and route choice in the face of these myriad factors. Toward that end, this project uses citywide bicycle travel data (i.e., Strava Metro Data) to provide a comprehensive description of cycling activities in a mid-size American state, as a proof-of-concept approach to planning for cycling.

1.2 Literature Review

1.2.1 Health and cycling

A healthy lifestyle is the most significant benefit of cycling. Physical activities are more likely to increase by introducing more attractive and friendly infrastructure. Thus, developing more cycling-friendly environments encourages society to implement a healthier lifestyle. Several studies confirm that cycling and physical activities have various health benefits. Furthermore, people who adopt cycling activities are less likely to have diabetes or hypertension compared to automobile transporters (3, 4, 5, 6, 7, 8). Also, investing in urban cycling infrastructure is cost-effective from the health sector perspective. In addition to the benefits of physical activities, decreased air pollution, and climate change rates would lower illness related to sedentary behaviors and pollution (9).

1.2.2 Safety and Cycling Infrastructure Planning

Building a safe bicycle infrastructure is an essential aspect that must be studied through planning for cycling environments because of its crucial role in encouraging people towards cycling. A route might be used for recreational, commute purposes, or both. Bike trails typically aim to

accommodate recreational cycling purposes. At the same time, On-street infrastructure is often targeted to serve commute purposes. Cycling infrastructures provide a sense of safety for cyclists. Therefore, the number of cyclists increases if transportation planners install the appropriate infrastructure. Moreover, the increase in cycling activities depends on the safety and the convenience of the cycling infrastructure (3, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18).

To understand the importance of safety precautions in cycling, many researchers studied the effect of cycling infrastructure on intersections in the context of cyclist's safety. Several data sources are used, including GPS data, count data, and crash frequency data to predict crashes for different infrastructure models. Additionally, researchers went further to develop safety performance functions to predict the safety of bicycle traffics in a specific region. Also, they used network screening methods to understand the effect of adding a cycling infrastructure on road safety and crash frequency. Furthermore, they studied the severity of bicycle crashes and the factors that affect it. Moreover, many exploited the ability to build machine learning models to analyze and predict crashes (16, 19, 20, 21, 22, 23, 24, 25, 26, 27). However, despite the growth in cycling safety research, more research is needed in terms of traffic data, street characteristics, and crash data (28).

Studies showed that implementing bike lanes infrastructure increased cycling traffics. Also, they found that outer areas could have a delayed increase in cycling activities because of its remoteness from the grid (5). Moreover, the sociodemographic analysis showed that cyclists who live near bike lane infrastructures tend to use their bikes more than others (29). Other cycling infrastructures are shared lane markers (Sharrows) and signed bicycle routes (30, 31). Planners install sharrows to inform drivers that cyclists have the priority on the street by increasing the spacing and lowering the vehicle speed (30). The use of sharrows has been controversial, and there are different opinions concerning the safety hazards of this type of infrastructure. For example, some researchers found that sharrows decreased cycling injuries on the roads compared to not using cycling infrastructure (32). On the other hand, other researches argued that shared street between cyclists and motorized vehicles is causing more injuries (33, 34).

The signed bicycle route is a roadway sign informing the cyclists of the preferred course, which does not provide separate lane or lane markers. Signed routes are commonly used in the US. However, there are limited studies concerning the effect of the signed routes on cycling. One study showed a positive correlation between signed routes and people's opinions towards good quality cycling infrastructure. Another study showed that cyclists preferred residential roads with signed routes compared to not having signs at all (15, 30, 31, 35, 36).

1.2.3 Big Data and Cycling

Various governments and organizations are utilizing big data to evaluate their cycling infrastructure. Big cycling data are usually collected using live point data, journey data, Bike-Share Programs (BSP), and GPS. For instance, live point data are collected on intersections using cameras on traffic lights, counting stations, or even sensors. While the journey data provide information about the origin and the destination of the trip, it does not give the details of the trip. This set of data could be collected from BSP or by other sources like online questionnaires. BSP

data are complete and in real-time. However, these data sets only give information within the area of its location. Strava is considered a GPS program made available by the social fitness network company Strava. Strava utilizes the Open Street Map (OSM) to deliver its data. These GPS data are very detailed and historical but represent a small sample of the total population of cyclists (37, 38, 39, 40, 41).

1.2.4 Strava Metro Data

Strava app data contains a vast amount of spatial and temporal details to predict cycling activity patterns. It provides a good approximation of the most-used routes and the peak months and hours. To protect privacy, the Strava data set is combined into population datasets. While a small portion of cyclists may use Strava to log their trips, the app might track trips for users using other transportation methods (6, 8, 39, 40). Researchers investigated the potential bias in the crowdsourced data like Strava metro data because it only represents a small sample of cyclists and proposed ways to correct the unfairness of the data (42, 43, 44).

On the other hand, several studies showed that there is a strong correlation between the Strava data, and the ground-truth data obtained from counting stations (5, 45, 46). The usefulness of the data was not limited to analyzing and predicting the temporal and spatial behavior of cyclists. Researchers used the data to evaluate the effect of adding a new cycling infrastructure (5, 47). In the United States, many state planners such as Colorado, Oregon, and New Hampshire started to use crowdsourced data, including Strava Metro data in urban planning and improvement of cycling infrastructure (48, 49).

1.2.5 Other Factors Affecting Cycling

Cycling is affected by several factors such as weather, time of the day, infrastructure, congestion, environment, income, public transportation, health, population density, the slope of the street, and cultural view towards cycling (45, 46, 50, 51, 52, 53, 54, 55, 56, 57, 58). A primal factor that affects cycling is the weather. Cycling is the most severely affect mode of transportation when it comes to weather variations (18, 59).

1.3 Project Objectives and Overall Results

This project aims to provide a comprehensive analysis of cycling routes in Nebraska. As we presented in the Literature Review, many factors shape cycling traffics categorized into temporal and spatial factors. Throughout the project, we provide correlation studies between Strava and bike counter data to understand the representability of Strava cycling data in context with ground-truth counter data. The results showed a strong relationship between Strava and bike counters and the correlation study help us to ensure that Strava cycling patterns follow the patterns of the cyclist population.

Subsequently, we investigate the effect of temporal factors on cycling patterns in Nebraska. Temporal factors include weather, day of the week, and the difference between weekdays and

weekends in terms of commute and recreational cycling, hour to hour analysis, and peak hours. The peak hours, and hour to hour analysis showed that the peak of commuting cycling hours is highly concentrated before and after working hours on weekdays, but on weekends, recreational cycling increases in the middle of the day.

Moreover, we model the relationship between weather conditions and monthly cycling trips in Nebraska. The analysis in this regard showed that cycling is profoundly affected by weather conditions and the outside temperature. As a result, cycling tends to increase in warmer weather and significantly decreases in cold weather.

Consequently, we consider specific locations in Omaha and Lincoln for spatial analysis. We choose the sites based on the infrastructure dynamics between 2017 and 2019. We study the effect of adding or removing bike lanes, bike signage, and sharrows. Moreover, we provide an analysis of cycling traffics across trails in Nebraska. The primary outcomes of this work are providing a generalized summary of cycling routes in Nebraska and proposing solutions to improve cycling infrastructure conditions and paving the path for future work. According to this study, cyclists are responding positively to the existing and added cycling infrastructure, and their activities tend to increase in designated cycling infrastructure like bike lanes, signed routes, and trails.

Discussion

2.1 Data Collection

Understanding cycling behaviors requires the utilization of many data sources. Crowdsourced data is hardly enough to perform an adequate cycling analysis. However, using other data combined with Strava data provides a powerful combination of datasets. To build a more robust database, we acquired data from several different sources. Other data sources include OpenStreetMap (OSM), Nebraska Department of Transportation (NDOT), Census data, and Metropolitan Area Planning Agency (MAPA). Also, we added more information provided by the City of Lincoln, City of Omaha, and the cycling community in Nebraska. Figure 1 shows a brief description of data flow management and the utilization of different sources and software tools.

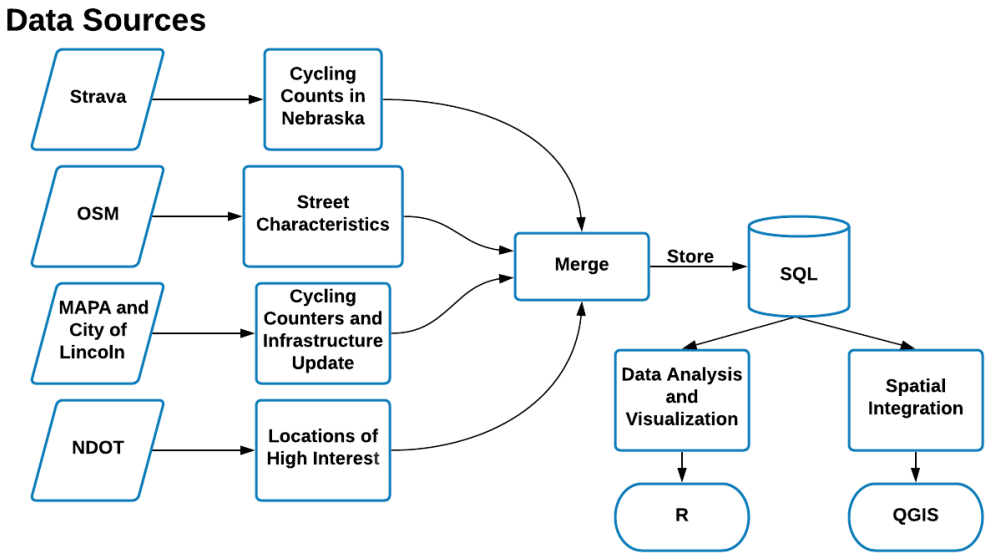


Figure 1: Flow chart of the data collection and preparation process.

2.1.1 Strava Metro Data

Strava is a company that provides a fitness application for cyclists. The application is the most known among fitness applications designated for cyclists. As most used cycling applications, the Strava app offers its customers the ability to record their journey data, including stats about their performance and the distance and speed traveled. Also, it challenges its customers by showing a ladder of the highest performance people using the app in the same area.

Strava app provides raw app data for research groups and government planners. In addition to state departments of transportation, many researchers used the Strava Metro data for cycling urban planning. Strava app categorizes its data into three classes, edges, nodes, and origin-destination pairs (OD). Edges are street segments between nodes (intersections), Strava provides the amount

of cycling traffic across each edge. Similarly, it provides the nodes data, which is cycling traffics across intersections. The OD is provided as spatial polygons, which represent the most used areas for starting or finishing a cycling trip. Moreover, all the classes mentioned earlier are categorized into hourly, monthly, and yearly data.

For privacy protection reasons, the Strava data set is combined into population datasets, which means it provides the data by edge, node, or origin-destination ID. Furthermore, it hides a portion of the information if a specific edge had low cycling traffic (less than three cyclists) within the used timeframe. This rule creates a problem for researchers studying the data. For example, some trips appear in the yearly data, but due to low traffics, these trips might not appear in the hourly or monthly data.

Every edge, node, or OD comes with an ID that can be joined with shapefiles to provide the ability to create maps using OSM with numerous spatial information utilized for a better understanding of cycling patterns and behaviors.

2.1.2 Open Street Maps (OSM)

OSM is an open-source built by a community of mappers that maintains spatial data about roads, land use, and places all over the world (60, 61). In this project, we use OSM to create maps demonstrating cycling data. Moreover, we extract information about the roads mapped. For example, the road class is classified as major, minor, service, or suitable for walking and cycling. Also, OSM provides the ability to extract street characteristics and land use data.

2.1.3 Nebraska Department of Transportation (NDOT)

Nebraska Department of Transportation funded this project to have a better understanding of the cycling patterns in Nebraska. Moreover, the NDOT guided us in this project by providing us with valuable information regarding the areas with high interest.

2.1.4 Metropolitan Area Planning Agency (MAPA) and The City of Omaha

The Omaha-Council Bluffs Metropolitan Area Planning Agency (MAPA) is a regional council of Governments (62). MAPA and the City of Omaha provided data about cycling infrastructure treatment. The usefulness of MAPA data comes when studying the effect of adding or removing a cycling infrastructure on the Strava cycling trips. In section 6, we provide an in-depth explanation of this data and its usefulness. Also, MAPA provided us with bike counter data for four different locations in Omaha. This data helps us to study the association between Strava data and bike counter data.

2.1.5 City of Lincoln

The city of Lincoln provided us with bike counter data. It contains weekly count data for trails in Lincoln. We use the bike counter data for correlation study in Lincoln, which is useful to confirm that Strava cycling patterns are matching the cycling population patterns.

2.1.6 Census Data

Census data provides demographics information about states, counties, or even zip codes. In this project, we are interested in population, income, education level, and unemployment rates in certain areas. We utilize this data in building machine learning models to analyze spatial, temporal, and demographics effects on cycling.

2.2 Data Preparation

The data preparation task requires the utilization of several software programs. PostgreSQL (63), QGIS (64), and R (65) are standard tools in Data Science, Machine Learning, and Geographic Information System (GIS) Analysis. In Figure 1, we show a brief explanation of our process for efficient extraction and analyzing data from different sources.

2.2.1 SQL

SQL is a standard programming language for creating and manipulating databases. We use PostgreSQL to update Strava data into our database, which makes it easier to deal with big data. Strava data comes in hundreds of tables. Using PostgreSQL, we reduced the number of tables used to almost 30 tables, including three years of Strava data and other sources. Moreover, other software tools like R and QGIS provide the ability to access the database in SQL, making it a more convenient environment to process data efficiently. Furthermore, SQL stores big data without affecting the processor speed, unlike using R, putting so much load on the computer memory.

2.2.2 QGIS

QGIS is a potent tool for spatial analysis and mapping. Also, it is handy for combining spatial data. Mainly, when the spatial data was collected from different resources, making it impossible to use SQL to connect the information correctly. QGIS has many free to use plugins, including algorithms to solve complex spatial problems.

2.2.3 R

R is a free software environment for statistical computing and graphics. R is used for statistical analysis and generating professional plots. Moreover, R has many libraries that support building many machine learning models. In this project, we use R for data handling and management, machine learning tasks, and visualization.

2.3 Correlation Analysis

Using crowdsourced GPS data provided by cycling applications like Strava for cycling infrastructure planning has always concerned researchers. Because of the bias introduced in the

data, the data does not include all the cyclist's populations. However, the advantages of using Strava Metro data beats its disadvantages, as proven in the upcoming sections.

In this section, we are displaying the correlation between Strava data and bike counter data installed in eight different sites in Lincoln and Omaha, which are the biggest cities in Nebraska. Figure 2 represents an overview of the correlation study procedure.

When two sets of data are strongly linked together, they have a high correlation. Correlation is positive when the values of two data sets increase together, and it is negative when one value decreases as the other increases. Thus, when studying correlation, both data sets are considered linearly dependent. To measure the relationship, we use the correlation coefficient ranging between -1 and 1. When the correlation coefficient equals 1, it means both variables increase together, and we can conclude that both variables are perfectly associated (66).

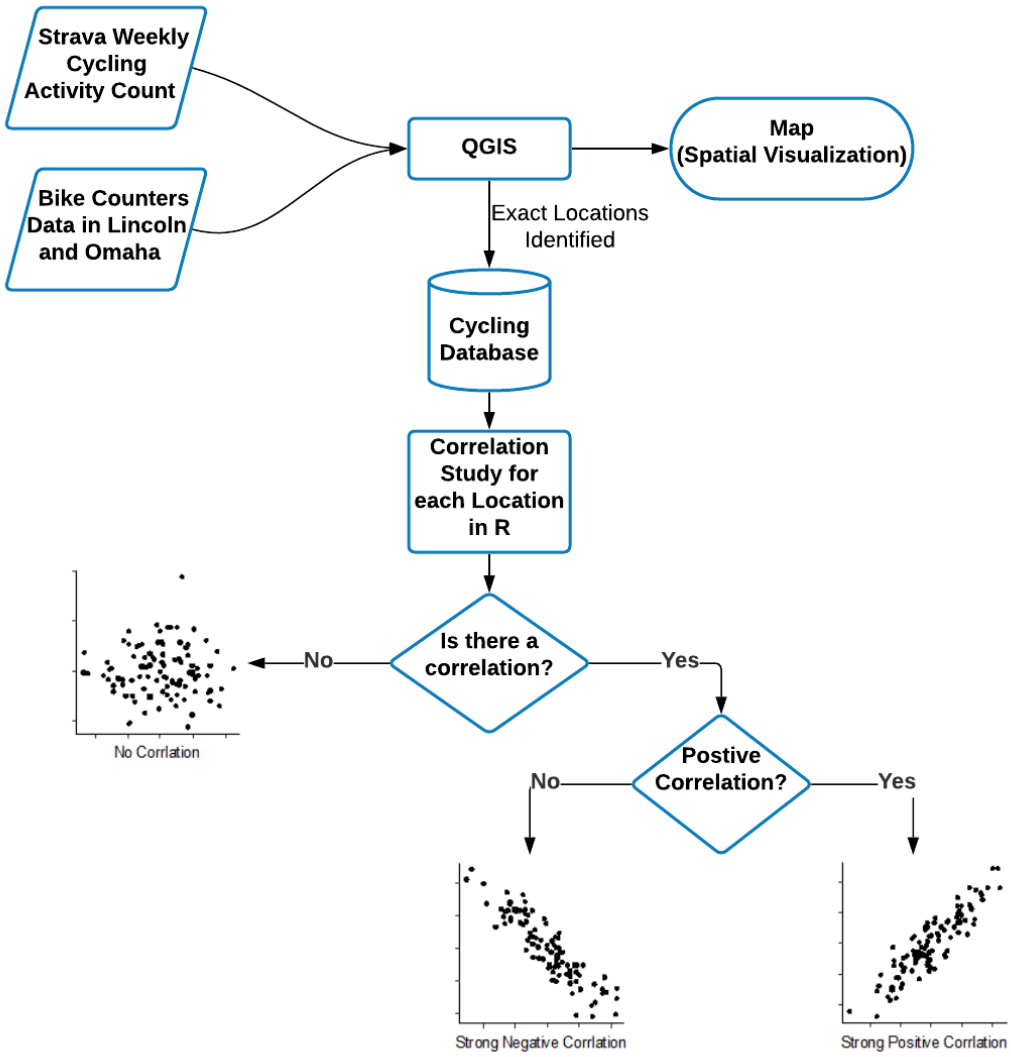


Figure 2: Flow chart of the correlation study between bike counters data and Strava data.

Similarly, when the correlation coefficient equals -1, it means that one variable increases the other one decreases. Thus, we can conclude that both variables are perfectly inversely associated. However, when the correlation coefficient equals zero, it means there is no correlation between the variables (66). The figure below shows a visual illustration of correlation relationships

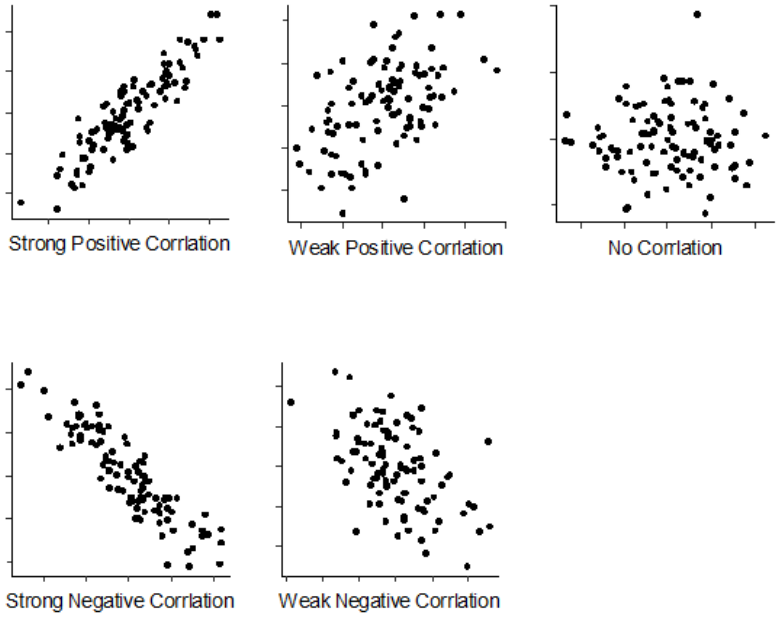


Figure 3: Visual illustration of correlation strength.

For the most part, correlation assumptions should be considered before studying the relationship between two variables. Altogether, both variables should be continuous and normally distributed. Furthermore, for a complete correlation study, we investigate the linearity between both variables.

2.3.1 Correlation Study Procedure

The most used correlation technique is Pearson's correlation. However, the method requires both variables to satisfy several assumptions such as continuity, linearity, homoscedasticity, outliers, and normal distribution. If the data does not satisfy one of these assumptions, statisticians recommend using Spearman's rank correlation or Kendall's Tau (66). In this project, we use Spearman's rank correlation. In the following, we show the Spearman's correlation formula (rho) (66).

$$\text{rho} = \frac{\sum(x - \bar{x})\sum(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

where x and y are the ranks of the first and second variables respectively, $\bar{x}$ and $\bar{y}$ are the mean of each variable observations.

In most cases, normality can be checked visually using the quantile-quantile plot (Q-Q plot) of each variable independent from the other. If the data falls in a straight diagonal line, we consider the data normally distributed. However, a statistically driven approach is to use the Shapiro Wilk test (67).

To perform the correlation between Strava data and counter data extracted from counting stations, we use the data provided by MAPA and the City of Lincoln in eight different sites. Strava data is linked with the exact edges that pass through the counting stations. Figure 4 shows a map view of the edges with counting stations we use to perform the correlation.

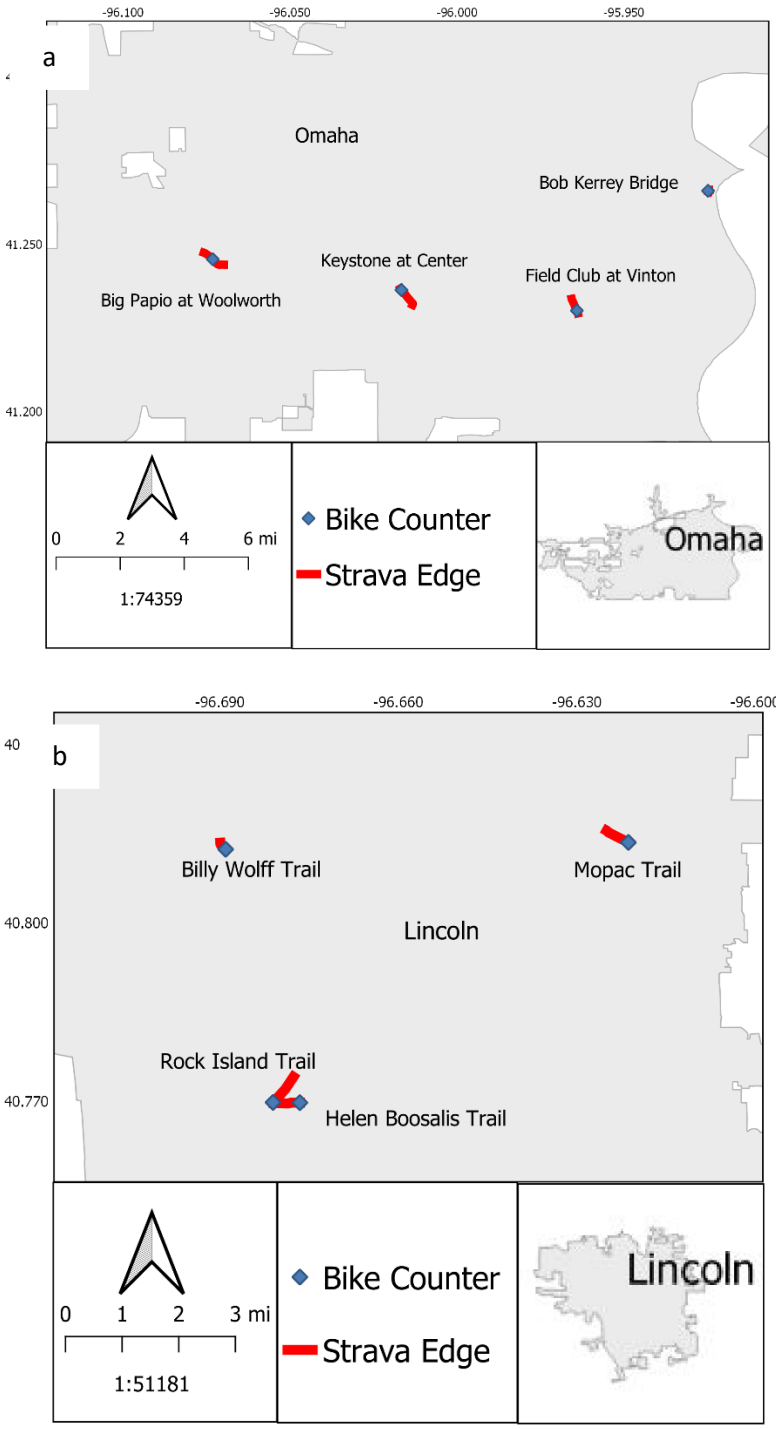


Figure 4: Bike Counter locations with Strava edges used in the correlation study. a) Omaha b) Lincoln

Because the Strava app builds the data on an hourly, monthly, and yearly basis, we aggregate the hourly data from Strava to a weekly basis to fit the weekly counter data from Lincoln. Afterward, we combine Strava and counter data in a separate dataset for each location.

We start with the Shapiro-Wilk test to check the normality of the data. The Shapiro-Wilk test Null hypothesis states that the data is normally distributed. Otherwise, the alternative hypothesis states that the data is not normally distributed (67). After performing the test, we found that Strava data was not normal. Therefore, we use the Spearman method to study the correlation. Moreover, this method does not assume linearity, which gives more flexibility in conducting the correlation.

Finally, we summarize the results of the correlation in figures and tables in the following sections. Moreover, we provide additional information, such as linear regression between Strava and counter data. Also, we show the percentage ratio between the number of trips from Strava and counting stations to support the correlation analysis. To evaluate the fitness of regression, we use the adjusted R2 parameter, which measures how close is the fitted data to the actual data. R2 in regression ranges between 0 and 1, if the R2 equals 1, then the fitted model perfectly fits the actual data (66).

2.3.2 Omaha Correlation Study

We acquired the installed bike counter data for trails distributed in Omaha from MAPA. We located the nearest Strava edge to the counter using QGIS. After we join the datasets, we perform the correlation procedure, as mentioned in the previous section. The four streets included in the bike counters data are:

- Keystone Trail
- Big Papio Trail
- Bob Kerry Bridge
- Field Club Trail

We study each site separately and provide the results in Figure 5 and Table 1.

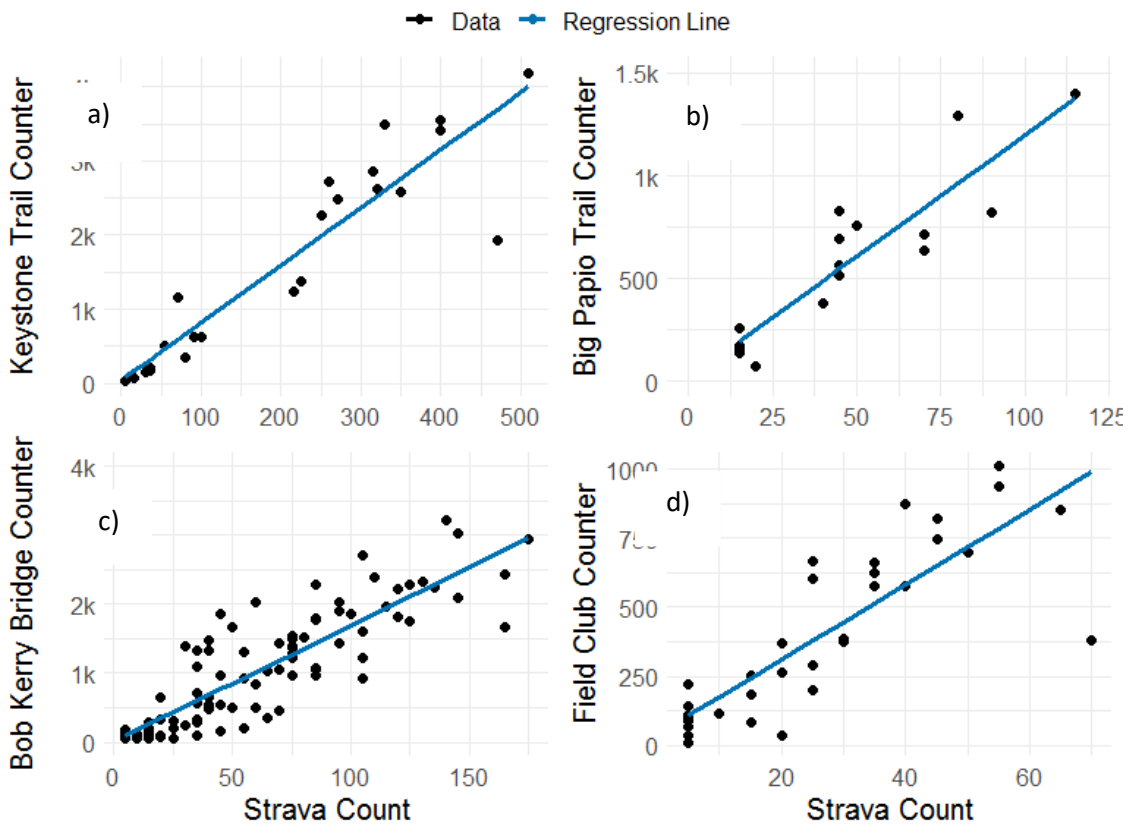


Figure 5: The relationship between Strava data and bike counter data in four different locations in Omaha. A) Keystone Trial, b) Big Papio Trail, c) Bob Kerry Bridge, d) Field Club Trail.

Table 1: Summary of Counter – Strava Correlation for several locations in Omaha.

Counter Location	Spearman Coefficient	Ratio %	Adjusted R^2
Keystone Trail	0.93	14.9	0.85
Big Papio Trail	0.87	9.6	0.80
Bob Kerry Bridge	0.87	9.1	0.73
Field Club	0.88	9.5	0.70

Figure 5 shows the linear relationship between Strava and ground counters installed in four different sites in Omaha, which provides a visual demonstration of the relationship between both variables. Also, it shows that Strava counts increase with bike counters counts. In addition to Figure 5, Table 4.1 shows a strong linear relationship between Strava and ground counter data from the city of Omaha. For instance, the measured correlation coefficient for all sites indicates a strong relationship between both variables because it is higher than 0.7. Moreover, the linear regression's R2 shows satisfactory results implying a highly positive correlation between Strava and counter data. Therefore, the correlation study proves that an increase in Strava cyclist activities is a projection of an increase in cycling activities of all cyclists. By using Strava data, we can understand and predict cycling activities in Omaha.

2.3.3 Lincoln Correlation Study

We acquire the counter data from the City of Lincoln. Like Omaha, we follow the same procedure of extracting and connecting the counter data to match the Strava data. The correlation study of Lincoln sites includes the following:

- Billy Wolf Trail
- Rock Island Trail
- Boosalis Trail
- Mopac Trail

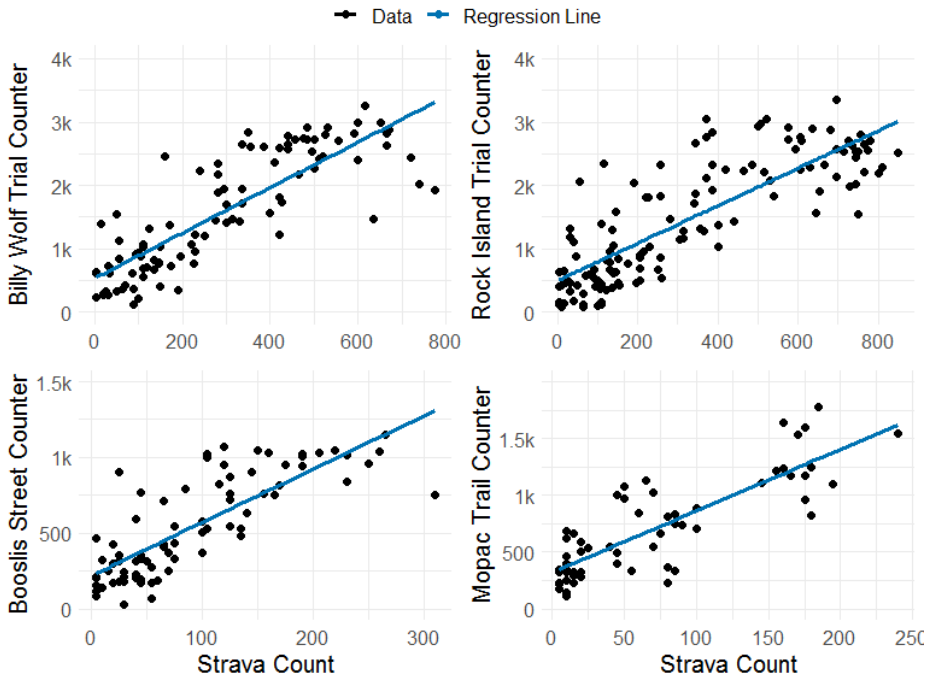


Figure 6: The relationship between Strava data and bike counter data in four different locations in Lincoln. a) Billy Wolf Trial, b) Rock Island Trail, c) Boosalis Street, d) Mopac Trail.

Table 2: Summary of Counter - Strava Correlation for several locations in Lincoln.

Counter Location	Spearman Coefficient	Ratio %	Adjusted R^2
Billy Wolf Trail	0.85	18.4	0.70
Rock Island Trail	0.81	22.7	0.66
Boosalis Street	0.80	15.5	0.64
Mopac Trail Club	0.80	9.2	0.68

From Figure 6, there is a linear correlation between Strava count and Lincoln city counts. Spearman's correlation coefficient ranges between 0.80 and 0.85 for these locations. However, we notice a higher percentage ratio of Strava trips in Lincoln compared to Omaha. For instance, Figure

7 shows a comparison between two locations that have almost the same cycling counts provided by bike counters.

Figure 7: Comparison between a) Bob Kerry Bridge Strava counts b) Rock Island Trail Strava counts.

Strava counts are higher in Rock Island Trail, which shows that Strava users form a higher percentage of the cycling community in Lincoln compared to Omaha. However, Omaha correlation studies show a stronger linear relationship between Strava and counter data. For example, we notice that the curve flattens with higher Strava counts at Rock Island Trail. This deviation from the linear regression line suggests that Strava users in Lincoln might form a higher portion of the cycling community with a higher number of activities. Figure 8 shows the percentage ratio of Strava activities in Bob Kerry Bridge and Rock Island Trail.

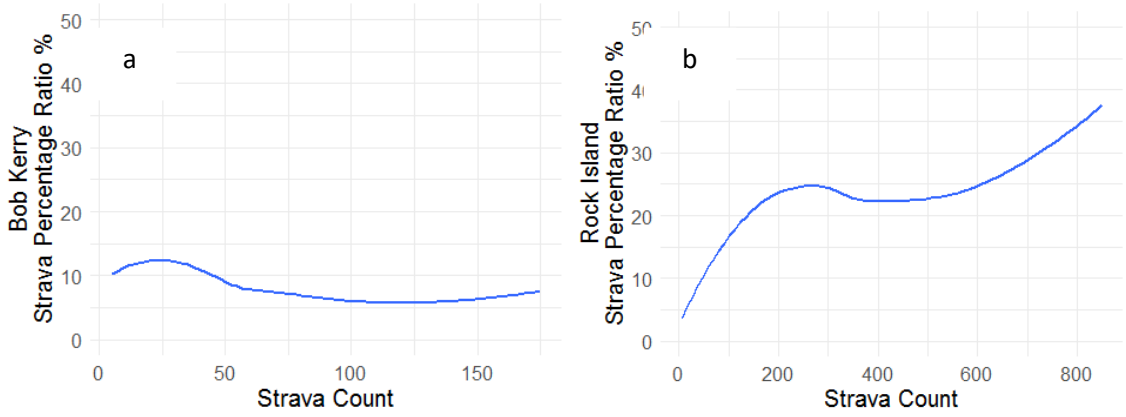


Figure 8: Percentage of Strava activities in a) Bob Kerry Bridge, b) Rock Island Trail.

Figure 8 shows that Strava activities form a consistent ratio of overall cycling activities in Omaha. Also, it indicates that the Strava sample in Omaha is smaller than the Lincoln sample. However, a smaller sample does not mean less representability of cycling patterns. This consistency of the sample ratio provides a robust linear model in Omaha with a higher correlation coefficient. On the other hand, we are not stating that the correlation study indicates the ineligibility of Strava data in

Lincoln. The correlation results in Lincoln show a high association between both variables. However, using Strava data in Lincoln should be implemented with extreme carefulness. Consequently, the correlation results show that Strava data is a potent tool to study cycling patterns. Also, it indicates that Strava data counts tend to increase when bike counter counts increase.

2.3.4 Summary of Findings

After conducting the correlation study, we found that all the bike counter data that are available in Omaha and Lincoln correlates with its counterpart in Strava data. Altogether, the correlation coefficient is above 0.7 in all studied sites. Moreover, Strava cycling data are following the same patterns of the overall cycling activities. Hence, we conclude that Strava data is a representable sample of the cycling population.

The correlation results in this section boost our confidence to use Strava data for cycling temporal and spatial patterns analysis. Crowdsourced GPS data like Strava offers an enormous amount of detailed spatial and temporal data, which is cost-effective and time efficient. Unlike relying on counter data, it only serves specific locations. Nevertheless, adding more counter data is essential for calibrating Strava data. For example, having more bike counters distributed around the State of Nebraska, the correlation study would be more useful for conducting a more in-depth spatial and temporal analysis.

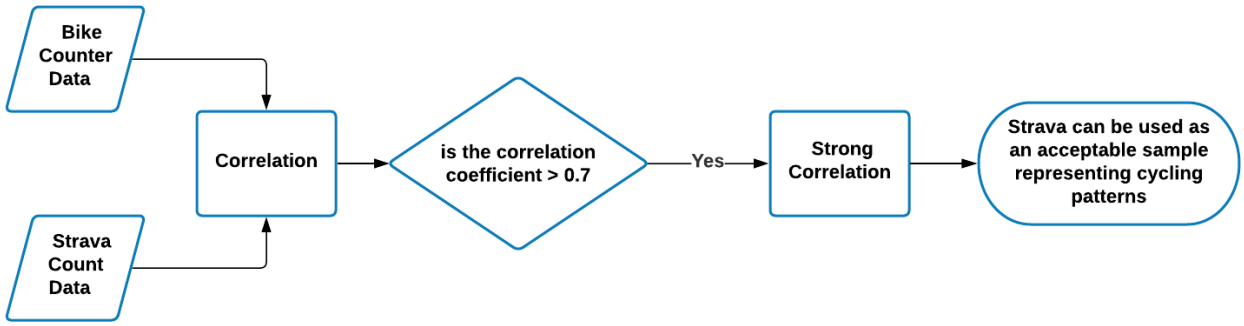


Figure 9: Correlation Study summary

2.4 Temporal Analysis of Cycling Data

There is a great interest in improving the cycling infrastructure in Nebraska by analyzing cycling data. Toward this goal, we acquired the Strava Metro data that covers the state of Nebraska from January 2017 to December 2019. The Strava app provides data into four categories edges, nodes,

origin-destination polygons, and shapefiles that contain OSM spatial attributes to create maps using GIS software. Edges are street segments between nodes. In other words, a set of edges form a street. Nodes are the intersections between edges, and origin-destination polygons divide a space into smaller areas. Every category of Strava data is divided into yearly, monthly, and hourly data. The goal of this section is to understand cycling behaviors in context with time-related factors. We use Strava hourly data to understand the effect of working hours and day of the week on the hourly cycling activities. Also, we use monthly data to study the weather effect on cycling.

Figure 10 shows a flow chart of the temporal analysis procedure. We use Strava hourly data to visualize the cycling patterns in Nebraska for each hour. Also, we use it to provide a study of the cycling activities on weekdays and weekends in terms of recreational and commute cycling. Furthermore, we use Strava monthly data for cycling seasonality analysis and to extract months with the highest activities. Finally, we use the monthly data to understand the weather effect on cycling by identifying the weather parameter that has the most effect on cycling.

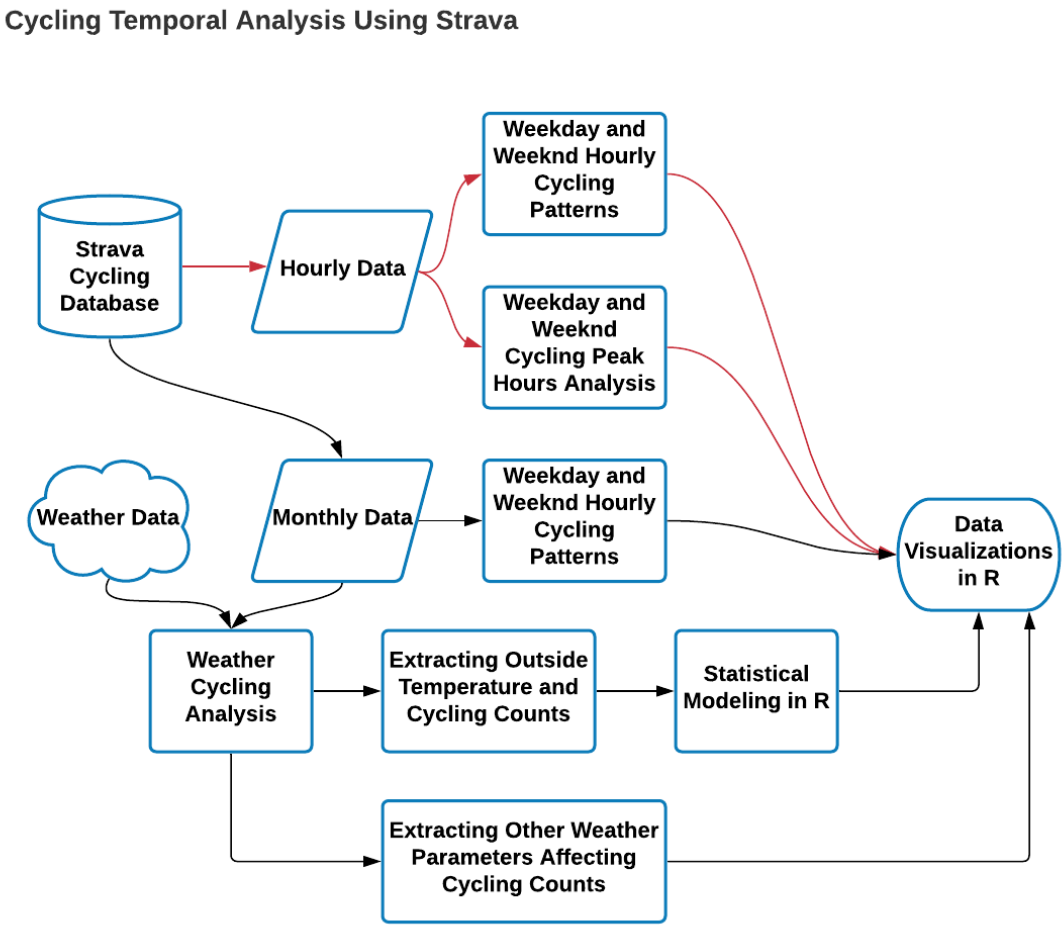


Figure 10: Flow chart of the cycling temporal analysis procedure.

2.4.1 Hourly Data

The hourly data contains total and commute cycling activities for every hour throughout the year. Additionally, for each trip, cycling purposes are divided into commuting and recreational activities. The Strava application distinguishes the commuting activities from the recreational activities using the start and the end places. Strava categorizes a commuting trip if the trip starts in a place and ends in another place. The difference between the total activity count and the commuting-trip count is the recreational cycling activity count.

The following figure shows Strava total cycling activities in 2017, 2018, and 2019. We aggregated the overall activity for each hour to extract peak hours of the day. The hourly data records contain every edge that had at least three cyclists per hour.

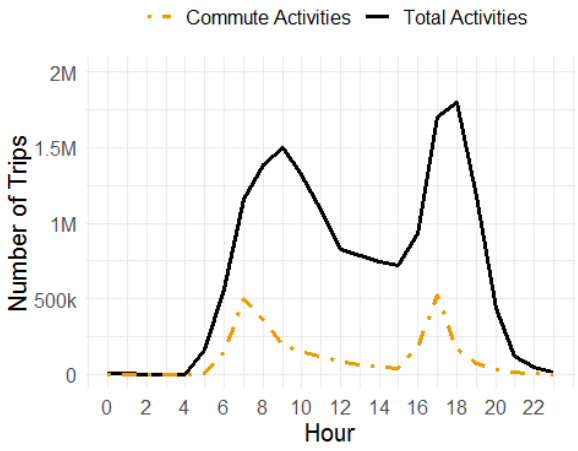


Figure 11: Total and commute hourly cycling activities in Nebraska for three years.

We calculate the number of trips in Figure 11 based on edges. Because Strava wants to protect the privacy of their users, it provides the number of trips per edge. In other words, the number of trips in the figure is the number of edges crossed, and one cycling trip can create from tens to thousands of trips in the data.

Figure 11 confirms that there are two cycling peaks in the morning and evening. In the morning, peak hours range between (8 am – 10 am). Similarly, we show that evening cycling hours have more cycling activities that range between (5 pm – 7 pm). Because of the time flexibility after work, evening peak hours have more cycling activities. Hence, it shows that people have more time to practice their cycling activities after work hours compared to the time they have in the morning. Additionally, the figure indicates that commute peak hours are consistent with total cycling activity peak hours, which provides further evidence of the representability of the data.

Traffic analyses focus on peak hours because it represents the most critical period of operations. However, the peak hour is not a constant value from day to day or from season to season. Therefore, it is deceiving to look at peak hours without including weekdays and weekends. To identify the cycling peak hours on weekdays and weekends, we accumulated the hourly data over three years, as shown in Figure 5.3.

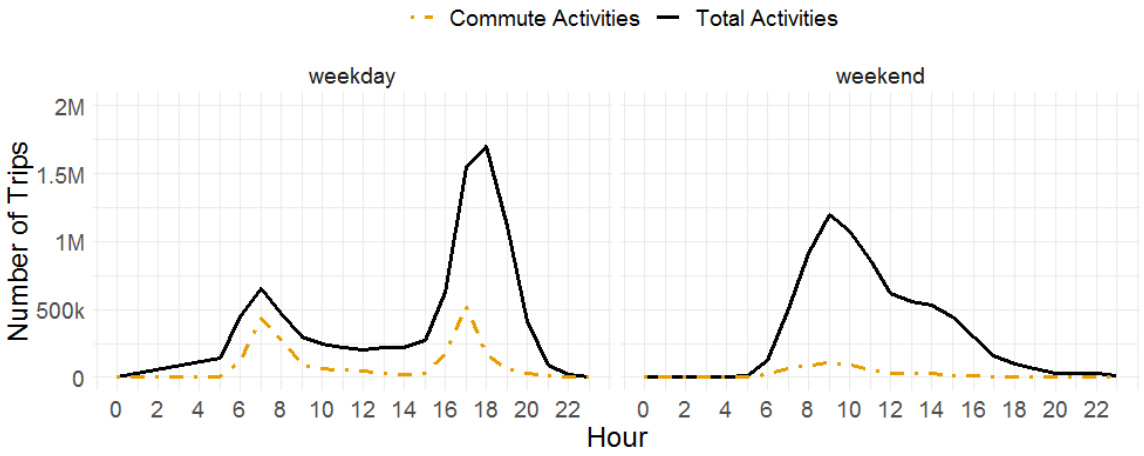


Figure 12: Strava hourly cycling activities categorized based on weekdays and weekends.

Peak hours on weekdays are different from those on weekends. Also, Figure 12 shows there are two sets of peak hours on weekdays. The first small peak hour is in the early morning between (6 am - 8 am) before working hours, and the second higher peak is from 4 pm and 7 pm after work hours. Between 9 am and 4 pm, the cycling activities decrease significantly because most people are working within these hours. Moreover, commute peaks have almost the same cycling activities indicating that commute cyclists usually use bikes to transport between home and workplace. Also, people are more likely to practice their recreational weekday cycling activities early in the morning before going to work or after work hours.

On weekends, commuting almost disappears, but recreational cycling activities increase. The considerable increase in recreational activities indicates that people prefer cycling after working hours or weekends because of time flexibility. Furthermore, weekend recreational cycling is more likely to occur within the day between 9 am and 4 pm following a sensible pattern.

Figure 13 demonstrates the recreational and commute cycling peak hours in three years. Also, we divide cycling activities depending on the day of the week into weekdays and weekends. Moreover, we show two peak hours for each day, one in the morning and the second is in the afternoon.

Figure 13 supports the observations obtained from Figure 12. The weekday peak hour analysis shows that cyclists follow the same patterns for three years; commute peak hours seem to entirely disappear on weekdays between (9 am – 3 pm). Also, recreational peak hours follow the same pattern with less concentration. On the other hand, weekend commute peak hours do not have a specific pattern compared to the recreational peaks, which are dense between 9 am and 4 pm.

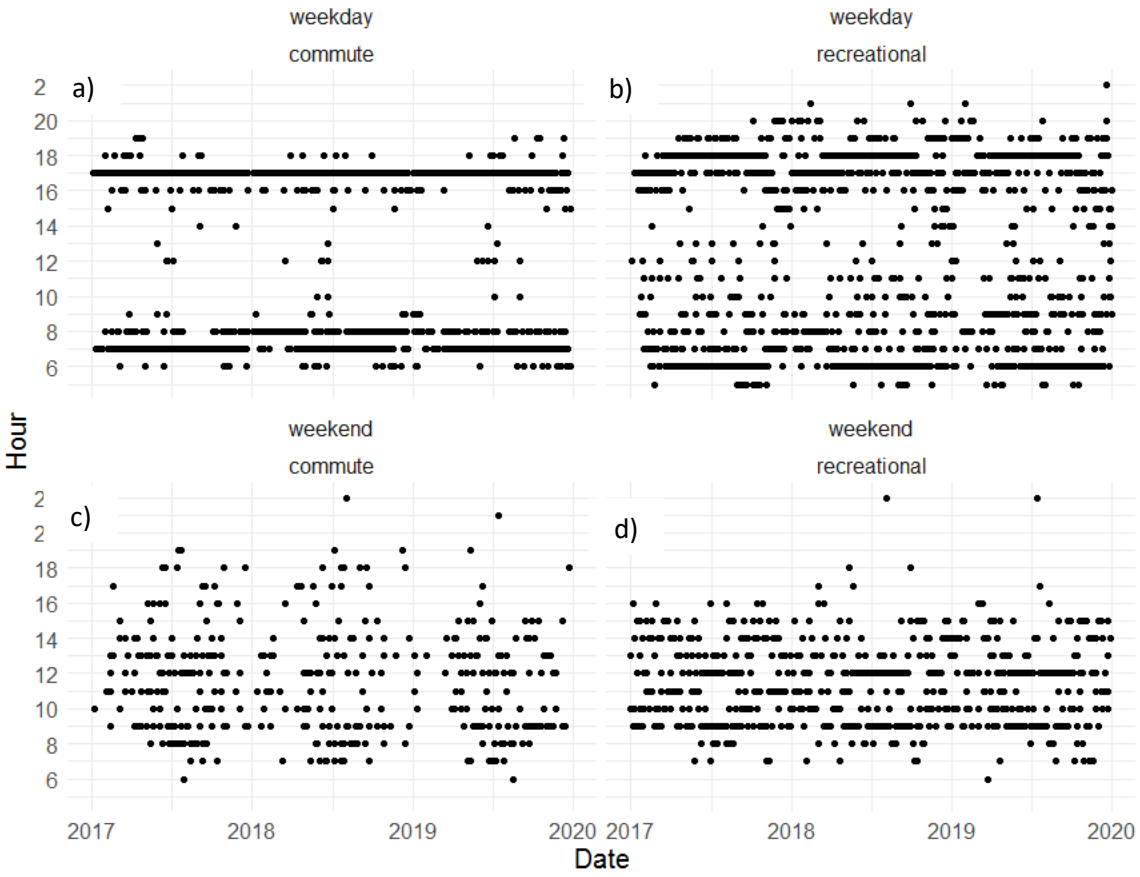


Figure 13: Cycling peak hours in Nebraska from 2017 to 2019 based on the day of week and trip purpose. a) commute weekday hours, b) recreational weekday hours, c) commute weekend, d) recreational weekend.

In Figure 14, we provide an insight into the difference between recreational and commute trips, show the cycling activities every day for three years between 2017 and 2019. The figure shows four categories based on the cycling purpose and the day of the week.

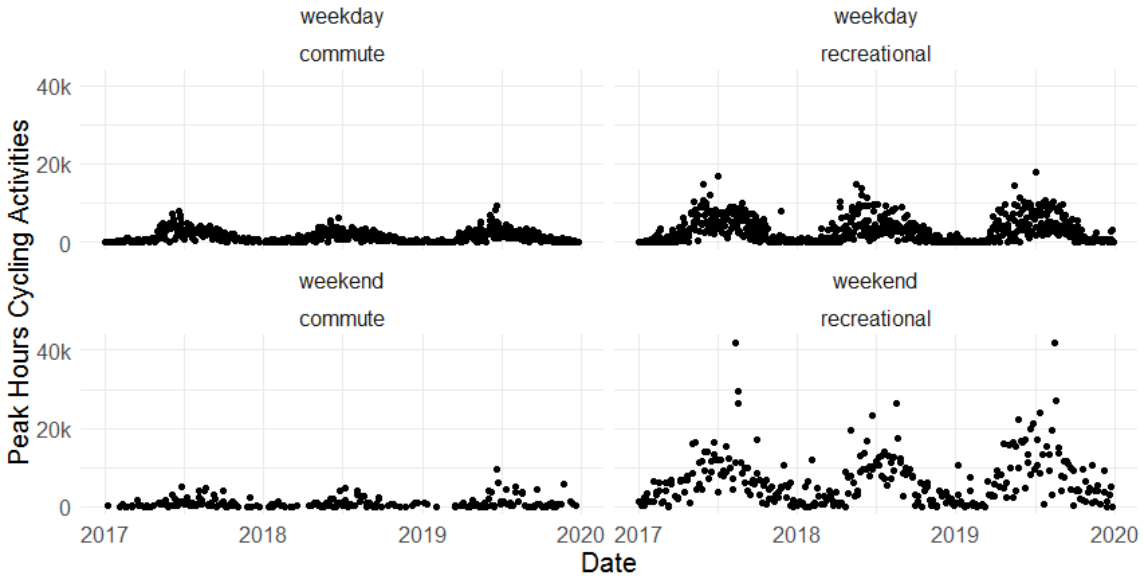


Figure 14: Number of cycling activities at peak hours every day between 2017 and 2019.

The number of recreational cycling trips tends to be more than commute cycling. Because of the nature of commute trip with limited time, cyclists are choosing the fastest route. Therefore, they are crossing fewer edges, indicating a shorter distance trip. Also, the figure indicates that recreational cycling involves crossing more edges riding for longer distances. Thus, it takes more time. Moreover, the number of trips in recreational activities tends to be higher at weekends, which shows another evidence of the relationship between recreational cycling and time availability.

2.4.2 Monthly Data

Strava monthly cycling activities are constructed similarly as the hourly data. The data consists of edges, nodes, and OD. However, they deliver cycling activities per month on edge level. Also, because of the company's privacy policy mentioned in section 2, the monthly data is more reliable because of the broader time frame. Therefore, we rely on utilizing the monthly data more frequently compared to the hourly data.

In Figure 15, we aggregated total cycling activities in Nebraska for three consecutive years between 2017 and 2019.

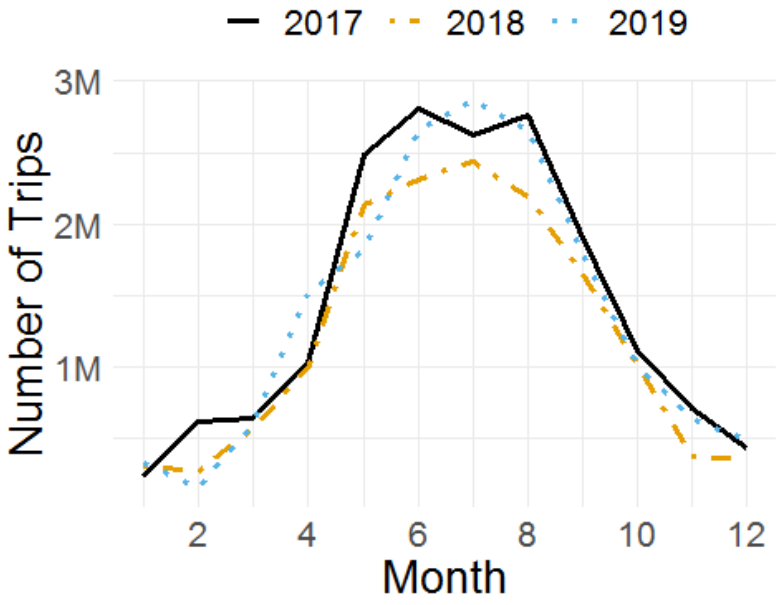


Figure 15: Monthly cycling trips for three years between 2017 and 2019.

2.4.3 Weather and Cycling

The weather has a crucial role in shaping cycling patterns. The outside temperature is the most dominant weather parameter influencing cycling. Generally, when the temperature rises in spring and summer, cycling activities increase significantly. The figure below shows monthly cycling patterns and outside temperature for three consecutive years.

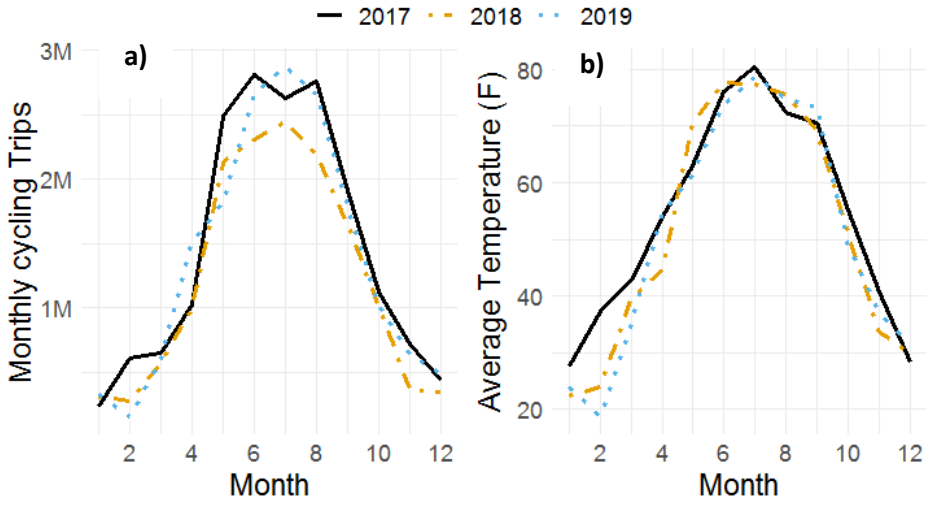


Figure 16: a) Aggregated Monthly cycling activities, b) Monthly average temperature.

Figure 16 indicates that cycling is directly affected by weather temperature. The aggregated monthly cycling patterns follow temperature patterns affected by changing seasons. Also, Nebraska usually experiences harsh winters where temperatures fall below zero. Moreover, the

region encounters snowstorms and wind chills that create impossible circumstances for biking in certain days of the year. For instance, looking at the temperature graph, we can see that winter in 2018 and 2019 had a lower average temperature, which projected in lower cycling activities in these years.

To understand the cycling temperature relationship, we fit a regression model between temperature and monthly cycling activities. We generate a polynomial regression between temperature and number monthly cycling activities. We describe the polynomial regression in the following equation (66):

$$y = C_0 + C_1 x + C_2 x^2 + \ldots\ldots + C_n x^n$$

Where n is the polynomial degree, linear regression is polynomial regression with a degree of 1. C is a set of coefficients, and y is the dependent variable we want to predict, in this case, it is the number of cycling activities. The figure below provides a brief explanation of the procedure we follow to build the polynomial regression.

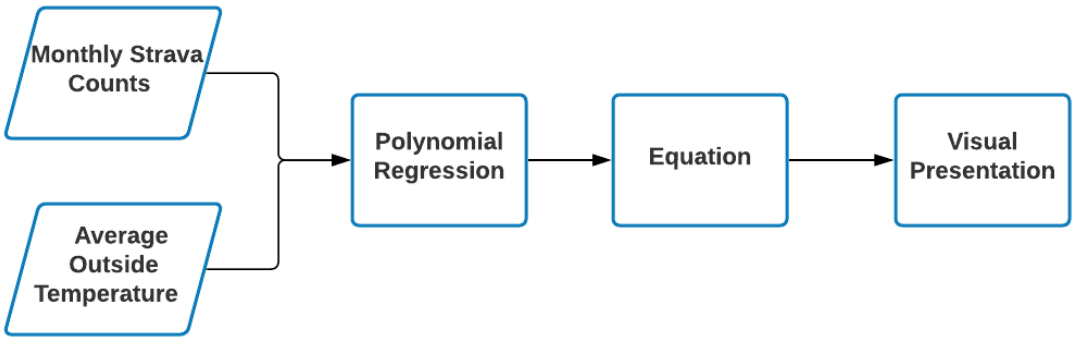


Figure 17: Flow chart of the polynomial regression procedure.

The polynomial regression equation that describes the cycling temperature relationship:

$$\text{Number of Cycling Trips} = 427.65 \times (\text{Temperature})^2$$

The equation shows a quadratic relationship between temperature and the number of monthly cycling activities obtained from Strava. Moreover, Figure 18 shows a visual representation of the relationship between temperature and cycling.

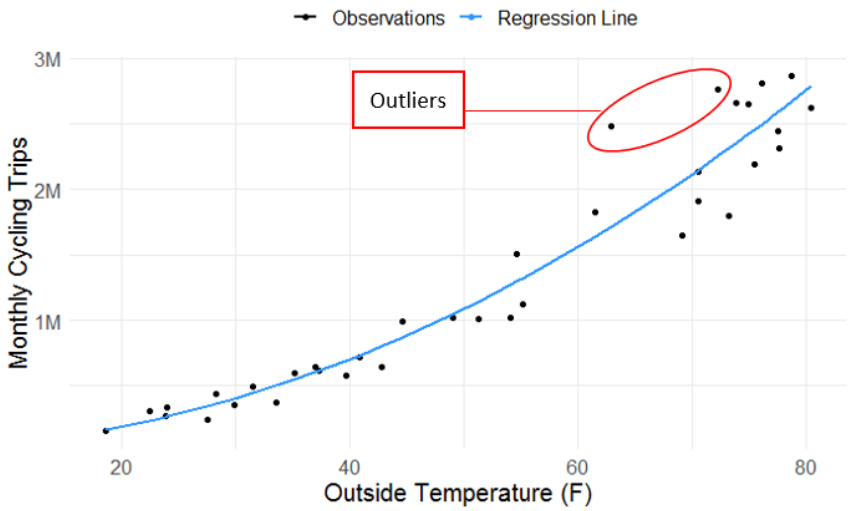


Figure 18: The relationship between monthly cycling activities and outside temperature for three years between 2017 and 2019.

Figure 18 shows that there is a directly quadratic relationship between both variables. Using the polynomial regression provides minimized error compared to the linear regression. R2 for this regression model is 0.92. Also, we measure the accuracy of the model predictions using the mean absolute percentage error, which equals 14% for this model. Consequently, the regression model suggests that there is a strong relationship between temperature and cycling, which signifies that temperature and season variation is a primary factor influencing cycling in Nebraska. Another explanation could be that bicycling is weather-dependent in Nebraska because there are few incentives to encourage cycling. Therefore, any disincentives (like weather) have a much more substantial impact.

Other weather parameters might disturb cycling activities, and the temperature is not the only factor in the variation of cycling counts. In Figure 19, we show two outlier observations obtained by plotting the residuals of the model predictions compared to the actual data. The outlier points are the cycling activities in May 2017 and August 2017. In May, the average temperature difference does not explain the high number of cycling activities in 2017. Though, the difference in foggy days and thunderstorms clearly shows that May 2017 encountered fewer weather fluctuations compared to 2018 and 2019, which discouraged cycling activities in these periods.

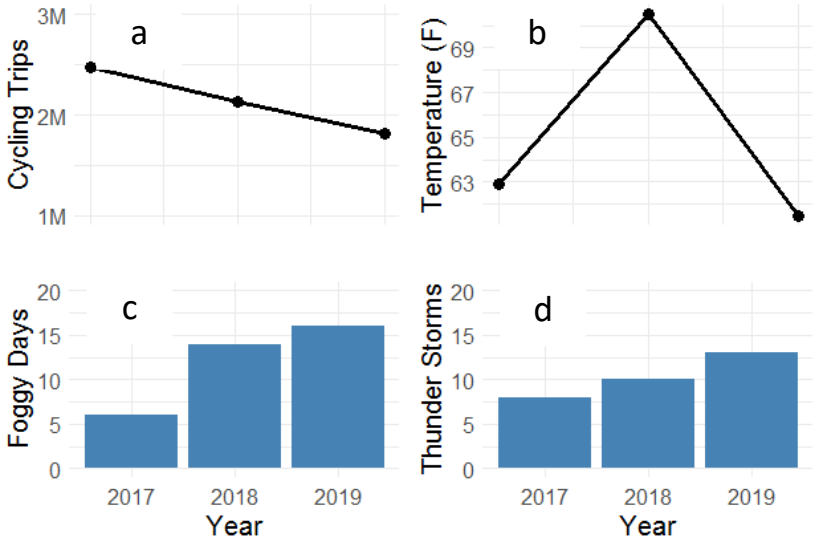


Figure 19: May cycling and weather relationship. a) May cycling activities in three years, b) average outside temperature. c) the number of foggy days, d) the number of days with thunderstorms.

In August 2017, the number of cycling activities was higher in 2017 and 2019 compared to 2018, despite the lower average temperature in 2017. The more moderate cycling activities in 2018 could be explained by the high temperature making the weather hotter and inconvenient for cycling. Also, other weather parameters can affect cyclists, such as fog and rain. The figure below shows a decrease in cycling activities in 2018, which encountered higher rainfall rates and more foggy days.

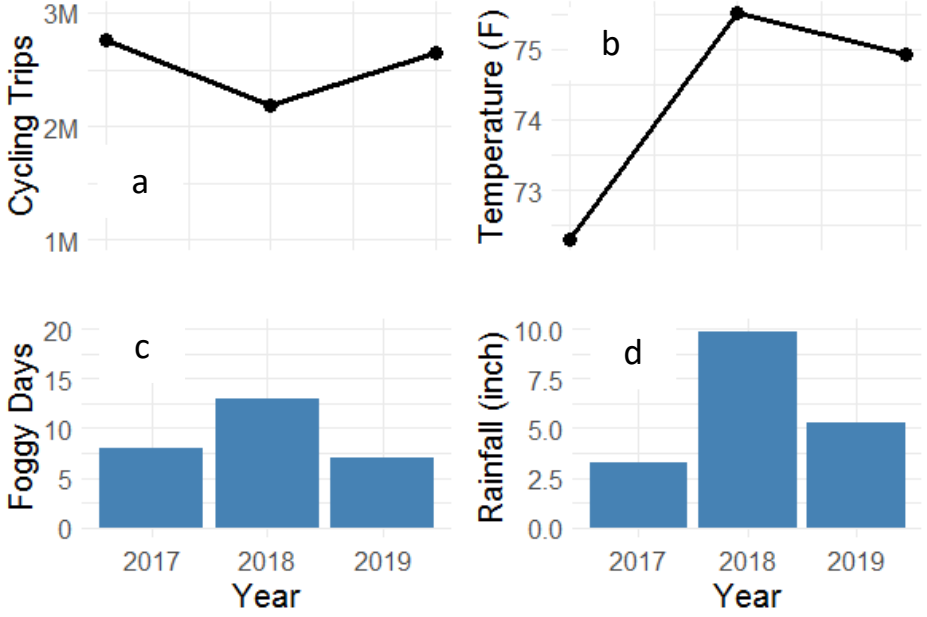


Figure 20: Comparison of August cycling activities and the relationship with weather. a) cycling activities in three years, b) Average outside temperature. c) the number of foggy days, d) rainfall precipitation in August.

2.4.4 Section Summary

In this section, we showed Strava temporal cycling analysis in Nebraska. The study included three years of data from 2017 to 2019, which we divided into hourly and monthly data. After conducting the research, we conclude that the hourly data from Strava follows usual traffic patterns, and it showed that commute cycling increases in the hours before and after work. Also, the data showed that recreational cycling increases significantly at weekends and after work hours on weekdays.
Moreover, daily peak hours analysis showed that morning and evening commute peak hours tend to be before and after work hours. However, it showed no distinctive patterns for commute weekends. Additionally, recreational cycling had a more scattered pattern on weekdays. In comparison, it showed a high intensity of peak hours in the middle of the day during weekends. The peak hours analysis showed that recreational cycling tends to be longer in time and distance. Unlike commuting cycling that aims to minimize trip time inducing shorter distances.
Finally, the combined monthly cycling and weather analysis showed that the temperature has a primary effect on cycling. The polynomial regression demonstrated the quadratic relationship between cycling and temperature. However, we showed that temperature is not the only weather parameter causing cycling variance. As an example, we showed that thunderstorms, fog, and rain also have a precise effect by decreasing cycling activities.
In conclusion, this section provides a piece of evidence to utilize Strava data in cycling analysis. In addition to correlation analysis, the temporal analysis shows that Strava follows the expected traffic patterns.

2.5 Spatial Analysis of Cycling Data

Temporal analysis is beneficial to understand the bigger picture of cycling patterns and behaviors. However, it does not provide any urban planning insight. In this section, we dive deep into the types of cycling infrastructure that cyclists use in Nebraska. Figure 21 shows a flow chart of the procedure we follow to extract and analyze cycling infrastructure effect on cycling.
The goal of this section is to investigate the effect of several infrastructure types on cycling activities. Also, we draw comparisons between existing infrastructures and how it affects cycling activities.

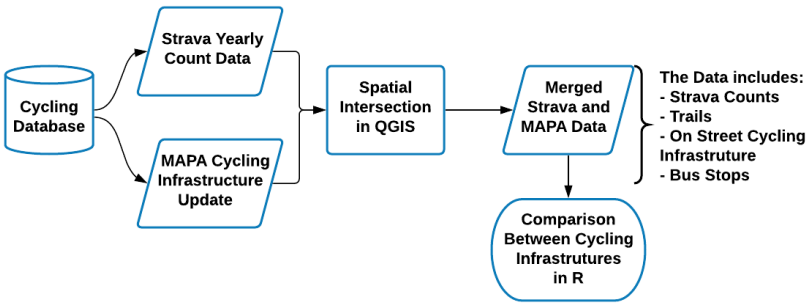


Figure 21: Flow chart of the spatial analysis operations.

We use yearly and monthly Strava data for spatial analysis. Additionally, Strava delivers the Nebraska OSM data for spatial demonstration using GIS software. After joining Strava with its spatial counterpart, we combine Strava and MAPA datasets spatially using QGIS. This process might be executed manually or by using the intersection algorithm in QGIS. Finally, we use the combined outcome from QGIS in R, which offers many statistical tools for data analysis.

2.5.1 Exploration of Several Spatial Factors Affecting Cycling

Cycling infrastructure is usually installed to build organized streets that serve all kinds of transportation methods and provide safety for cyclists. Therefore, safe, and convenient cycling routes encourage people to practice cycling.

In Omaha, cycling infrastructure consists of trails and on-street infrastructure including bike lanes, signed bicycle routes, and shared lane markers (sharrows). Furthermore, we add streets with bus stops to the analysis because buses in Omaha offers to carry the bikers and their bicycles on board. In Figure 22, we show a map view of the spatial location of trails and on-street bike infrastructure in Omaha. We specifically use the city of Omaha because MAPA provided the data for us. As a planning agency in charge of designing streets and cycling routes, the agency is the most trusted source to get this data.

Figure 22.a shows the on-street infrastructure in the Greater Omaha Area. The infrastructure is concentrated in the city center near major universities. Near Papillion, which is a small city next to Omaha, there are bike lanes installed in 2017 and 2018. Furthermore, additional bike lanes in Bellevue near Bellevue University. However, most of these bike lanes were installed before 2017. Because we only have the Strava data from 2017 to 2019, we cannot investigate the changes in cycling activities after installing the bike lanes in this location.

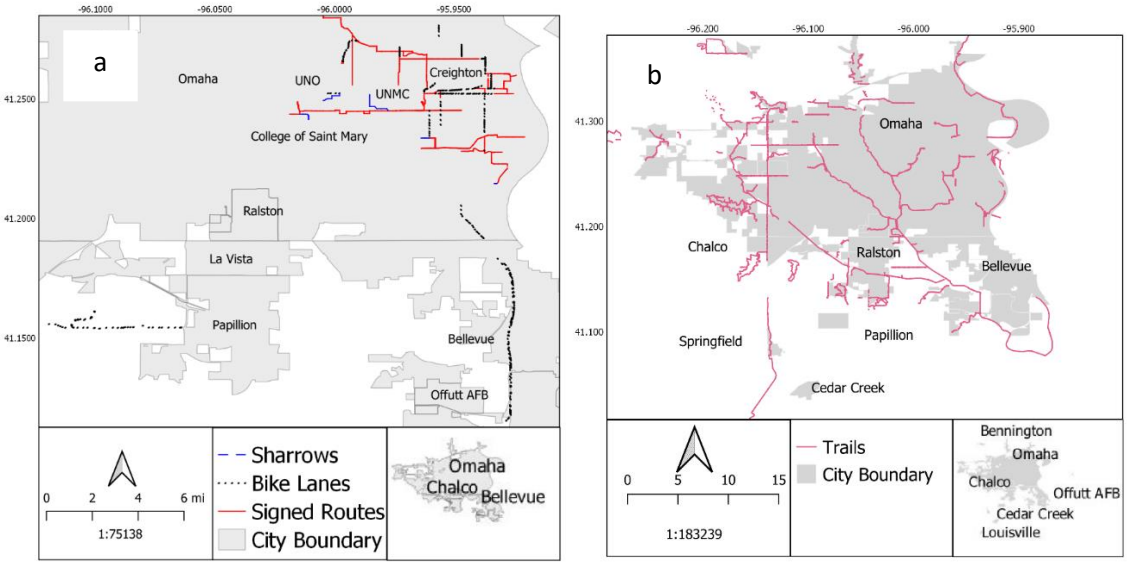


Figure 22: Map showing the location of a) On-street infrastructure, b) Trails in Omaha Area.

We aggregate the average cycling activities for three years based on the infrastructure type. Figure 23 shows a comparison between the average total and commute cycling activities and the infrastructure type. Because of the discrepancy in the number of edges between infrastructure types, we use the average rather than using the total sum. Thus, we calculate the average by dividing the total Strava count of cycling activities by the number of edges for each type of infrastructure.

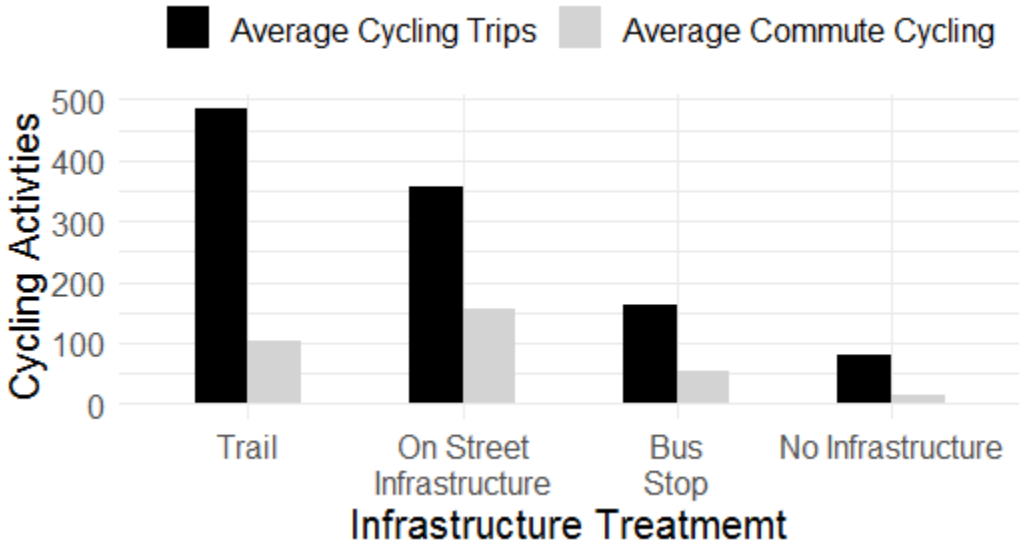


Figure 23: Comparison between several cycling infrastructures based on the total and commute cycling activities in the Greater Omaha Area.

We can see that cyclists prefer trails and on-street cycling infrastructure shoeing cyclists prefer designated biking infrastructure rather than having no infrastructure. Furthermore, commute analysis shows that on-street cycling infrastructure is the most preferred by cyclists. We explain this observation by the fact that these infrastructures are usually installed in streets with higher activities like downtown or near university campuses.

On the other hand, total cycling activities, including commute and recreational cycling, are higher on trails. These trails are usually isolated from street grids that accommodate automobile drivers. Thus, the cycling trails become the perfectly safe place for recreational cycling apart from cars and other disturbances and the second most used routes for commute purposes. Therefore, trails offer safety and convenience for cycling commuters even with longer commute distances.

We do not consider bus stops as cycling infrastructures. However, public buses in Omaha provides the ability for the cyclist to hang their bicycles outside the bus. Figure 24 shows an example of a bicycle rack used on Omaha buses to transport cyclists and bikes. Therefore, Figure 23 shows that roads with bus stops experience higher cycling activities than roads with no infrastructure, which indicates that installing a bike rack on buses might be useful for cyclists.

Figure 24: Bicycle rack on Omaha Metro buses.

2.5.2 On-Street Cycling Infrastructure Treatments

On-street cycling infrastructure treatment is mostly linked with bike lanes. These bike lanes are portions of the street exclusive for cyclists. Bike lanes are solid white marked lines that are usually installed on roads or pavements. Likewise, sharrows are placed on the streets. Unlike bike lanes, sharrows do not give exclusive portions of the road for cycling purposes, and it only indicates that bicyclists are permitted on this lane. Another type is bicycle signage or signed routes, which is the least sophisticated cycling infrastructure. As it only informs cyclists and drivers that cyclists are permitted using roadside signs. The City of Omaha has installed all three types throughout the years. In this section, we show the effect of each cycling infrastructure on the cycling trips recorded by Strava users between 2017 and 2019. Figure 25 shows an example picture of each type.

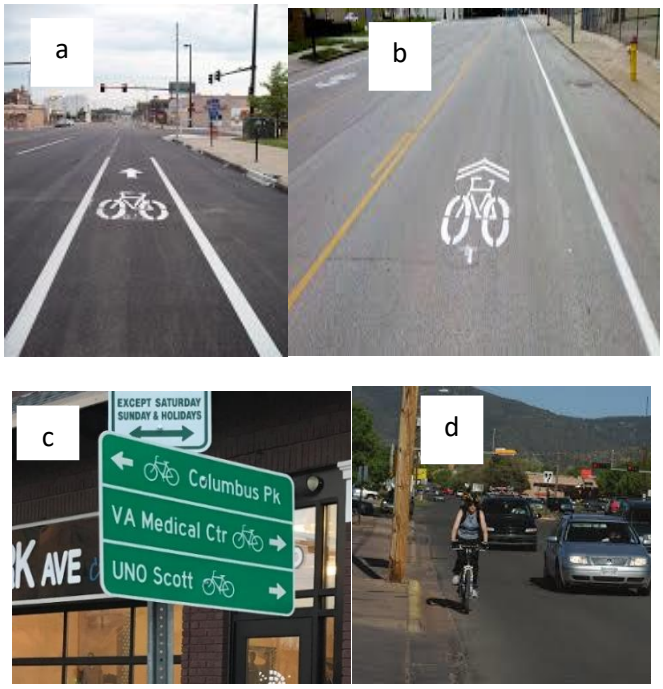


Figure 25: An example of on-street infrastructures in Omaha. a) Bike lane, b) Sharrows, c) Signed Routes, and d) No infrastructure.

Figure 26 shows a flow chart of the procedure followed throughout this section. We manually connect Strava data with MAPA data containing information about the On-street infrastructure using QGIS, including the type and completion or removal date. Subsequently, we use the most dynamic location, which comprises diverse types of infrastructure to create comparisons leading to conclusions regarding the convenience and activity of each infrastructure.

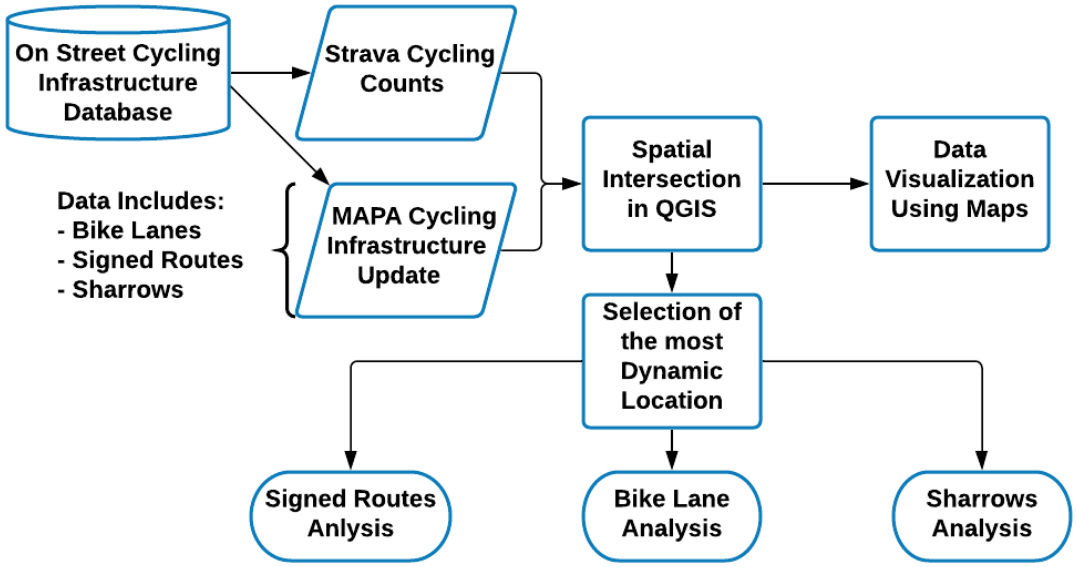


Figure 26: Flow chart of the on-street cycling infrastructure analysis.

Figure 27 shows the latitude and the longitude of several cycling infrastructures in Omaha. We separate them into three locations to minimize the variance in population and center of activities effect on the number of cycling trips. Location 1 is located near the city center of Omaha, which makes it more likely to encounter higher cycling activities. Moreover, Location 1 has different types of cycling infrastructure, which makes it a better candidate for comparison studies between the different infrastructures. Additionally, this specific location experienced many changes throughout the study timespan by adding or removing cycling infrastructures. Therefore, we provide a more in-depth investigation of this specific area.

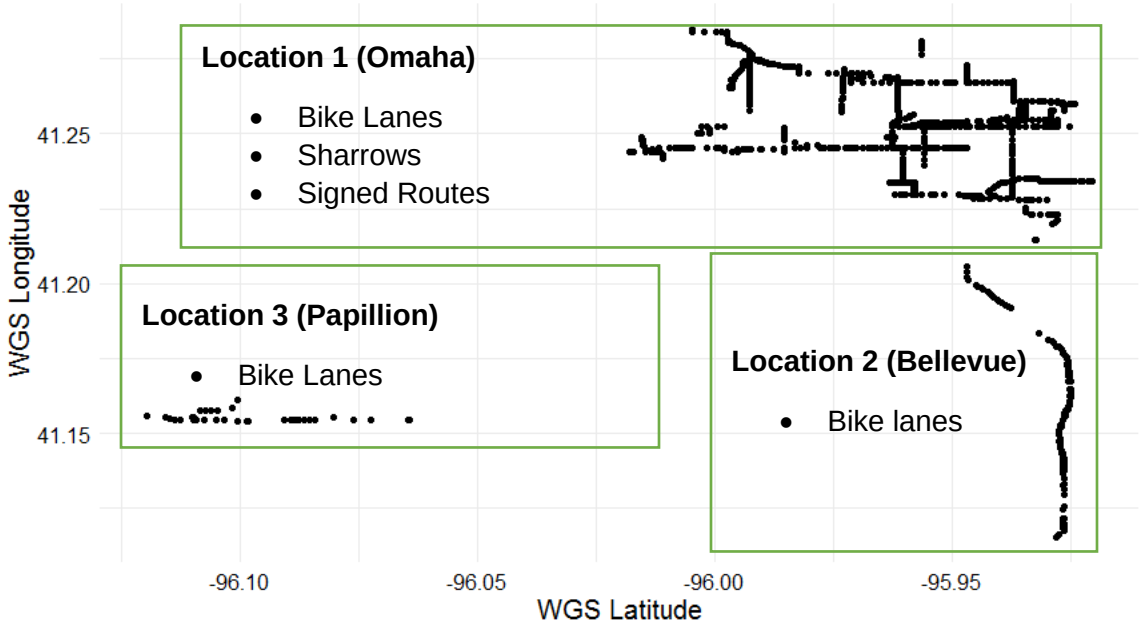


Figure 27: Dividing of on-street cycling infrastructure in the Omaha area into three locations.

This area of study has three main cycling infrastructure types, bike lanes, sharrows, and signed routes. Most of the bike lanes and sharrows were added before 2017. On the other hand, signed routes were mounted in March 2019. Unlike bike lanes, some of the sharrows were removed and replaced by signed routes. In Figure 28, we show how this area evolved between 2017 and 2019. We can see in 2019, most of the sharrows were replaced by signed routes, which provides an opportunity to study the effect of removing sharrows. Also, it provides another chance to explore the impact of the signed route and compare both cycling infrastructures.

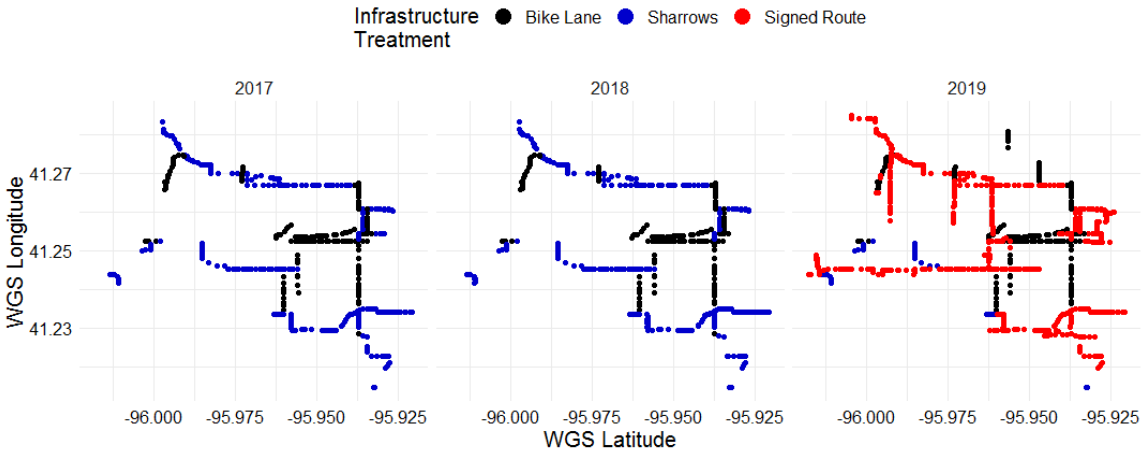


Figure 28: The evolution of cycling infrastructure in Omaha between 2017 and 2019.

To provide a more detailed analysis, we use the monthly data from Strava. Figure 29 below shows the monthly average cycling trips regardless of the cycling infrastructure type in Location 1. The cycling activities show a slight increase in 2019, which might be considered because of adding

bicycle signs across many streets. Because of the discrepancy in the number of streets depending on the type of infrastructure. It should be noted that we use average cycling activities, rather than using the total number of cycling activities. For instance, the total shows a higher number of activities for the most installed infrastructure. However, we calculate the monthly average, which provides the number of cycling activities divided by the number of edges. Hence, we can compare the effect of each infrastructure type.

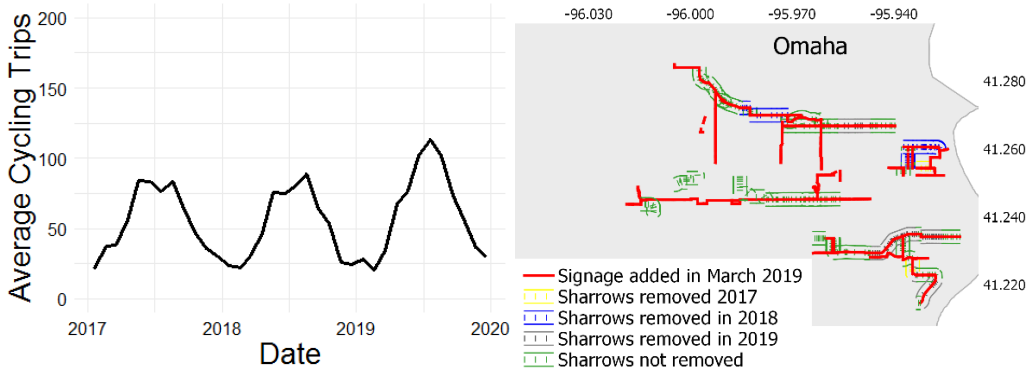


Figure 29: The average cycling trips across edges with different cycling infrastructures in Location1.

Here, we show the effect of each infrastructure type on cycling activities. In Figure 28, we show red lines indicating the addition of signed routes in 2019. Some of these streets had sharrows, which were removed between 2017 and 2019. However, some sharrows are still existing. Figure 6.10 shows the average cycling trips on streets that had sharrows with signed routes added in March 2019. Figure 29 shows that the monthly average cycling trips were almost the same in 2017 and 2018. In contrast, there is an increase in cycling activities in 2019, which shows that signed routes could be the reason for increasing cycling activities. For instance, the number of cycling trips stayed the same when sharrows were removed in 2017. This observation suggests that sharrows have no significant effect on cycling.

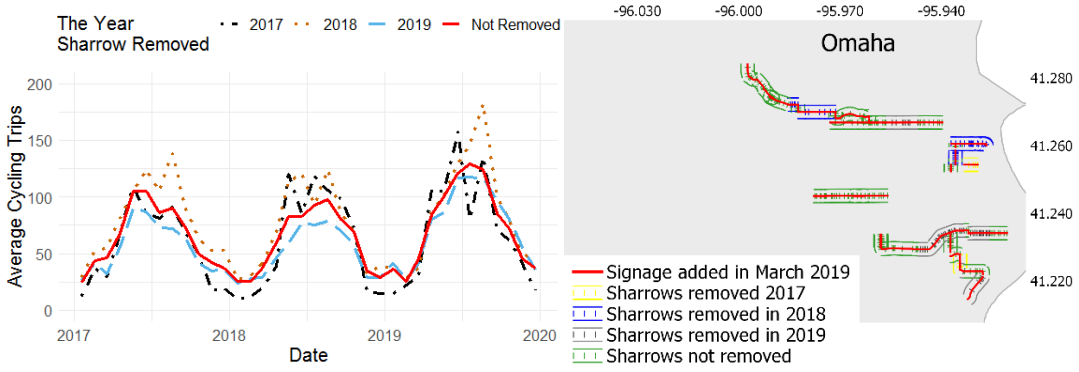


Figure 30: Monthly cycling activities on streets that had sharrows and signed routes added in March 2019.

Next, we investigate the cycling activities on streets with only sharrows. Figure 31 shows the monthly cycling activities in three years on sharrows and no signage added in 2019. We can see that sharrows removal did not affect the cycling activities. As shown in the figure, the cycling activities were already low on these sharrows, which shows their existence did not have an added value to the cycling activities.

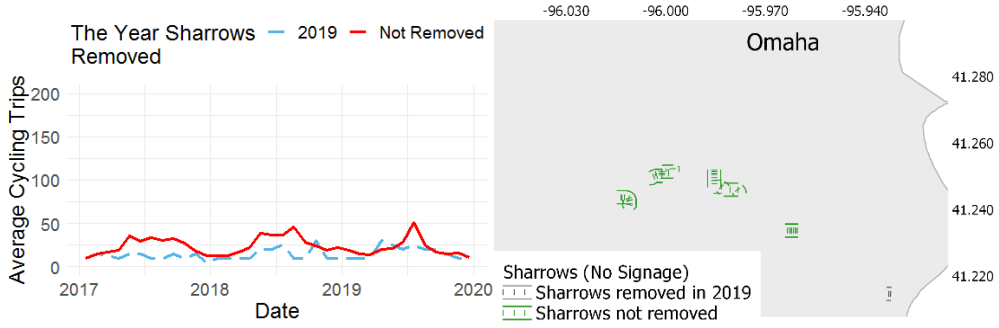


Figure 31: Monthly average cycling activities on sharrows.

In Figure 28, we show that some streets did not have any cycling infrastructure in 2017 and 2018. However, the red dots indicate that the signed routes where added later in 2019. We demonstrate the cycling activities ion these streets in Figure 32, which shows a higher monthly average of cycling trips in 2019 compared to 2018 and 2017.

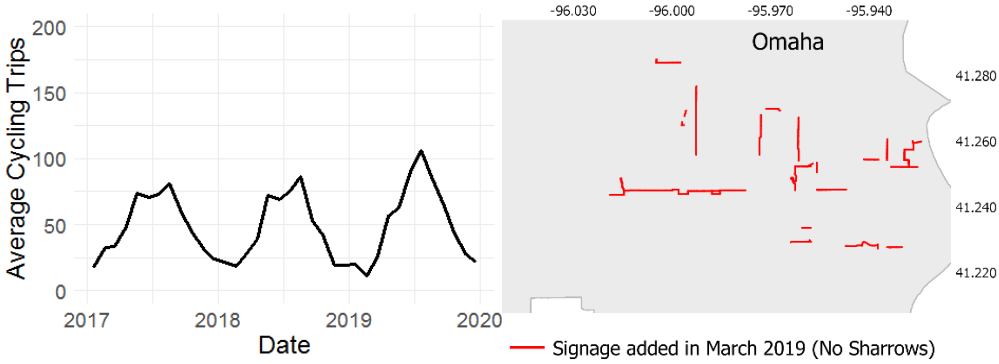


Figure 32: Average monthly cycling activities in streets with signed routes added in March 2019.

The use of signed routes might explain this increase. Nonetheless, in the following section, we support these results by conducting a complete analysis of signed routes and the spatial factors affecting cycling in this area.

2.5.3 Signed Routes

In this section, we provide a more predictive oriented analysis of the signed routes and cycling activities. To demonstrate the used data, we show in Figure 33 by using four types of data to construct the model. The modeled data consists of cycling count data from Strava, street

characteristics from OSM, and monthly weather data. To add more dimensions to the model, we include the demographic data that corresponds to the location of added signed routes. Demographics data, extracted on the zip code level, includes population, gender, employment rate, number of houses, and median age.

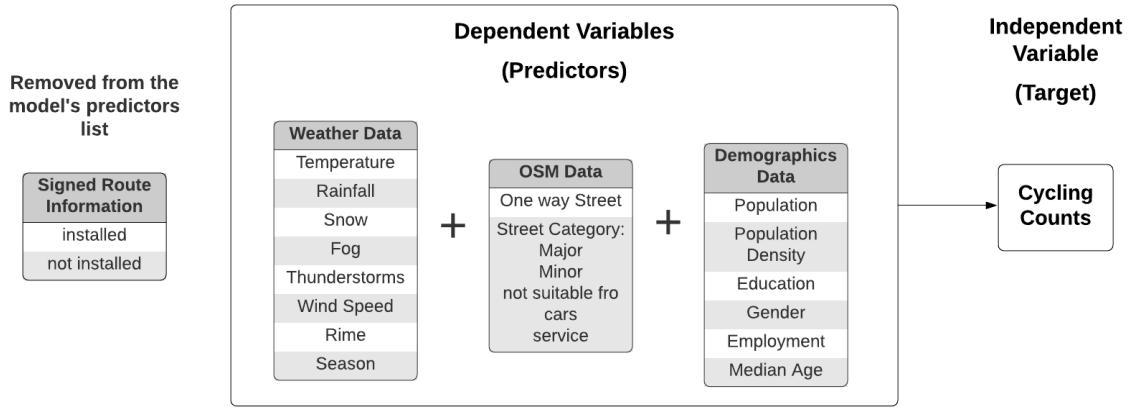


Figure 33: Demonstration of the data used to predict the monthly cycling activities on signed routes.

In Figure 34, we show a flow chart presenting the process we follow to build the machine learning model and extract results. We use the integrated monthly data to split into a training dataset containing the 2017 data and a validation set containing 2018 data. Also, we use the validation set for error measurement to compare different machine learning models. Finally, we use the 2019 data for testing the best model and provide the most factors affecting cycling in this area.

It is important to note that we hide information regarding the installation of signed bicycle routes from the model. If the trained model predicts the validation set more accurately than it predicts the 2019 data, we can conclude that signed bicycle routes affect the cycling activities. Therefore, we can compare two scenarios of installing or not installing signed routes. The following table provides a further explanation of the comparison method.

Table 3: the comparison method used to understand the effect of signed routes on cycling activities.

Comparison between predicted and actual cycling activities in 2019	Interpretation
Predicted activities are less than actual activities	Signed routes triggered an increase in cycling activities
Predicted activities are more than actual activities	Signed routes caused a decrease in cycling activities
Predicted activities equal actual activities	Signed routes did not have any effect on cycling activities

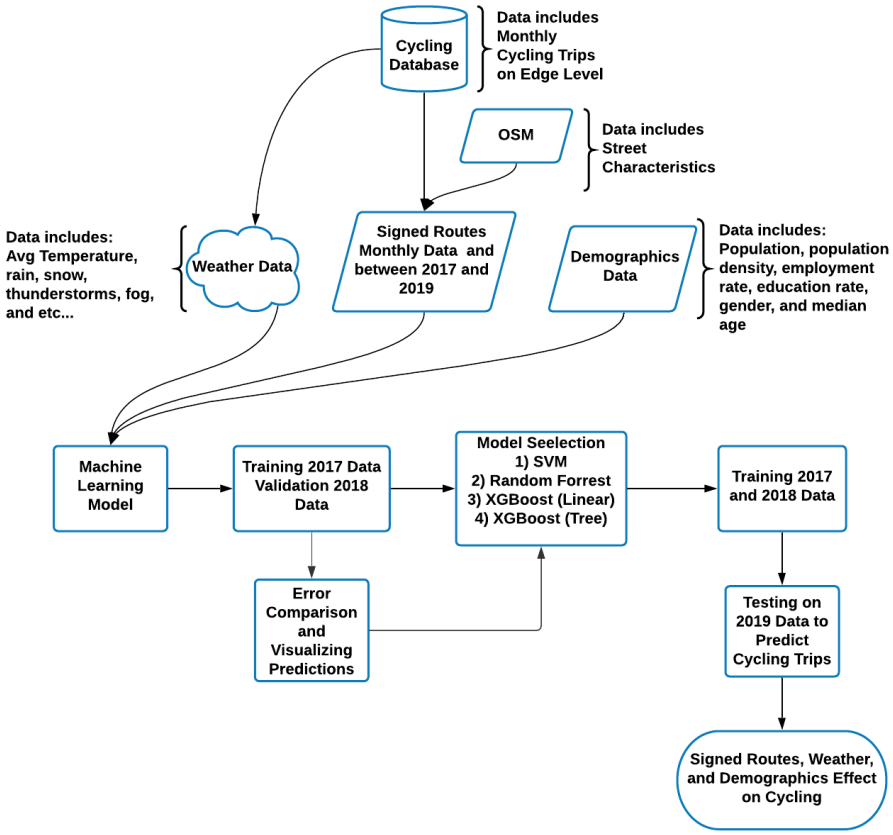


Figure 34: Flow chart demonstration of the signed routes analysis using Machine Learning.

2.5.3.1 Data Preparation and Machine Learning Overview

In this section, we provide an overview of the machine learning process. In the beginning, we preprocess the data by identifying the variables in the dataset into numerical and categorical values.

For the most part, categorical variables require more attention. For instance, If the categorical variable consists of three or more distinct values, we need to look at the nature of the variable. If it is ordinal, then we can substitute by binary numbers. However, if the categorical variable consists of non-ordinal values, we need to create dummy variables in which we add extra columns to the dataset that corresponds to each value. However, we avoid creating dummy variables for a large dataset that contains IDs or a vast amount of different values, because it is computationally expensive. Figure 35 shows an additional example explaining the categorical variable processing.

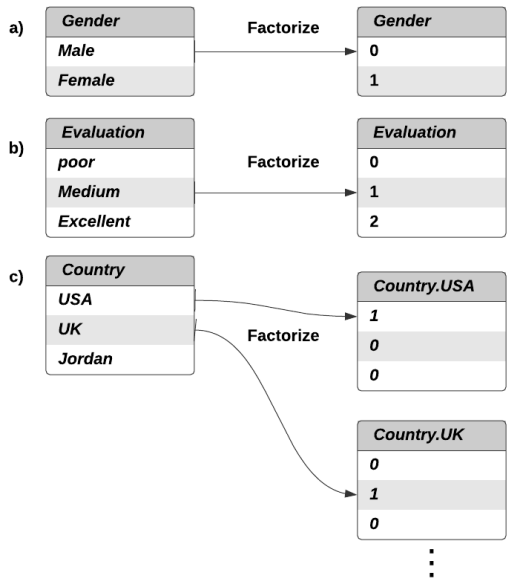


Figure 35: An example of categorical variable handling, when it includes a) two values, b) three or more ordinal distinct values, c) three or more non-ordinal distinct values.

Consequently, the data is ready to enter the machine learning process after completing the preprocessing phase. In Figure 34, we split the data into a training and testing set and use the weather, OSM data, and demographics data to predict the Strava cycling counts by training the model on the 2017 data and predicting the 2018 cycling counts. We use different machine learning models to extract the best model, which we feed to the final training data that includes the 2017 and 2018 data. Thus, we can create a comparison between and 2018 and 2019 predictions to understand the effect of signed routes. For instance, if the model predicts 2018 data accurately and predicts fewer cycling activities in 2019. Therefore, we can conclude that sign routes increased cycling activities, and it had positive on the cycling infrastructure. Finally, we use the best model to extract the crucial parameters that influence cycling.

2.5.3.2 Machine learning Model: Training and Testing.

We perform machine learning in R using the CARET library, which is a library designated for solving machine learning problems. The library has more than 200 machine learning algorithms used to create predictive models. Also, it includes feature selection, model parameter tuning, and variable importance extraction (68).

Generally, we avoid aggregation of the data spatially, and we use monthly edge data from Strava in the analysis. However, later in this section, we aggregate the data by taking the accumulative sum for comparison and visualization purposes.

The first model consists of Strava, OSM, and weather data. We train the 2017 data using k-fold cross-validation, which leaves out a portion of the training data to test the model performance (69). Subsequently, we teach the model using four different well-known machine learning algorithms,

Support Vector Machines (SVM), Random Forrest, XGBoost (Linear Booster), and XGBoost (Tree Booster) (70, 71, 72). To extract the best performance, we tune each model separately to minimize mainly the Root Mean Squared Error (RMSE) in addition to the Mean Squared Error (MAE). The following Figure 36 shows the trained model performance for each machine learning algorithm.

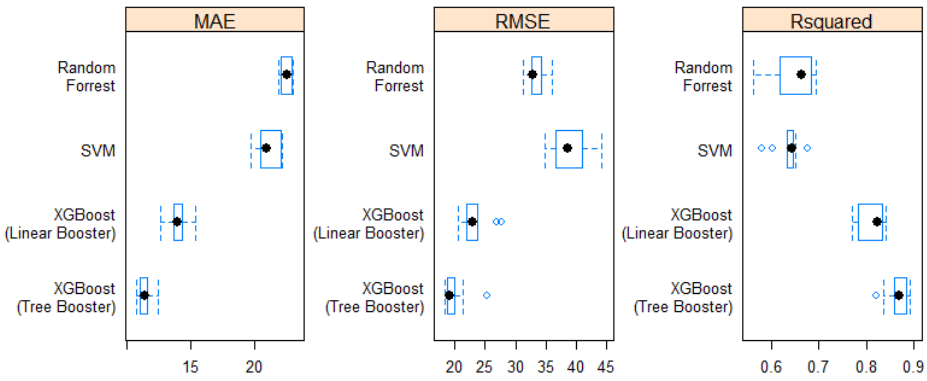


Figure 36: Regression model performance metric for several algorithms.

Figure 36 shows that XGBoost worked better for the training set resulting in minimized error compared to other models. Also, it provided the highest R2 value. Furthermore, the Tree Booster model produces less error value compared to the Linear Booster.

Nevertheless, we do not fully trust the trained model performance results because some models might be overfitting if it is trained rigorously. For example, if the training error was so much lower than the testing error, this might indicate that the model is overfitting.

Thus, we test the trained model on the 2018 data and measure the performance for each model. In Figure 37, we show error analysis of the models' predictions compared to the actual cycling activities in 2018

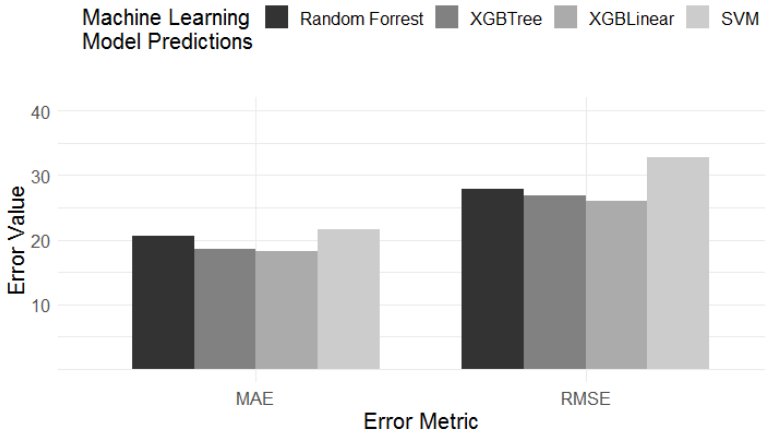


Figure 37: Predictions error measurement for several machine-learned models.

Unlike the training set performance results, which showed XGBTree had the best performance, the error analysis for the 2018 validation set showed that XGBLinear performs slightly better than XGBTree. Moreover, it provides the least error among all predictive models used in this study. For a better understanding of the results, in Figure 38, we show every model prediction compared to the monthly accumulative sum of cycling activities in the entire spatial area.

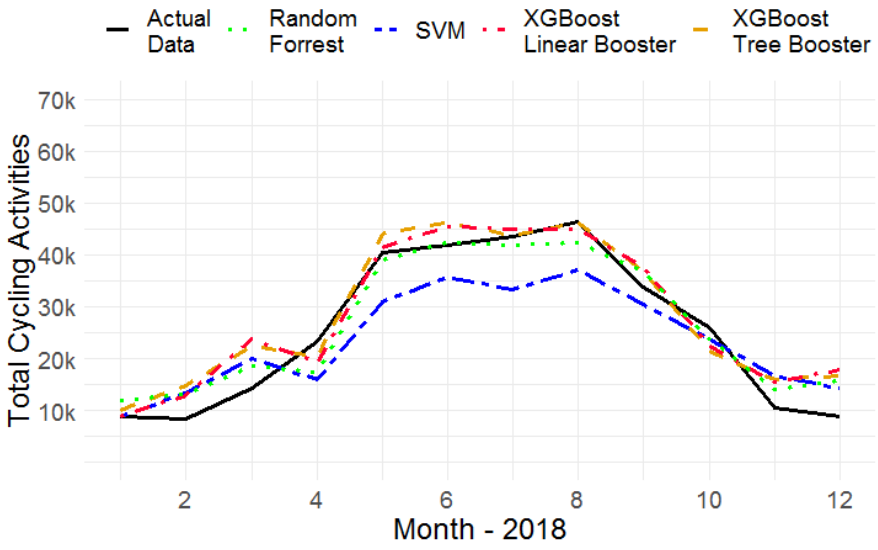


Figure 38: The actual and predicted cycling activities in 2018.

Figure 38 shows the actual and predicted number of monthly aggregated cycling trips in 2018. From Figures 37 and 38, we can see that XGBLinear is the best model in predicting the cycling activities. Also, it predicts higher activity months between April and October more accurately than other models. Moreover, the model predictions for 2018 validation set provides an accurate model that can be compared with 2019 predictions as a benchmark reference to understand the effect of adding signage in this area.

To ensure that the model has enough data for training, we use a second model that utilizes the data of 2017 and 2018 as a training set. We use the outcome of models' comparison from Figure 37, which showed that XGBLinear was the best model for this type of data.

XGBoost is an optimized tree boosting method that is widely used recently. As we mentioned before, we are using the CARET Library, which provides two versions of XGBoost (Linear Booster and Tree Booster) under the name of "Extreme Gradient Boosting." In the same library, we can tune the parameters to provide the best performance possible. Tuning a machine learning model is a repetitive process executed until extracting the parameters of the model to minimize the error. Table 6.2 shows the results of the tuning process and the function of each parameter in the XGBLinear model.

Table 4: Tuning parameters of the XGBLinear model to predict cycling activities in 2019.

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

Now, we execute the model with the best performance, which shows a slight decrease in the error compared to the model that uses only 2017 data.

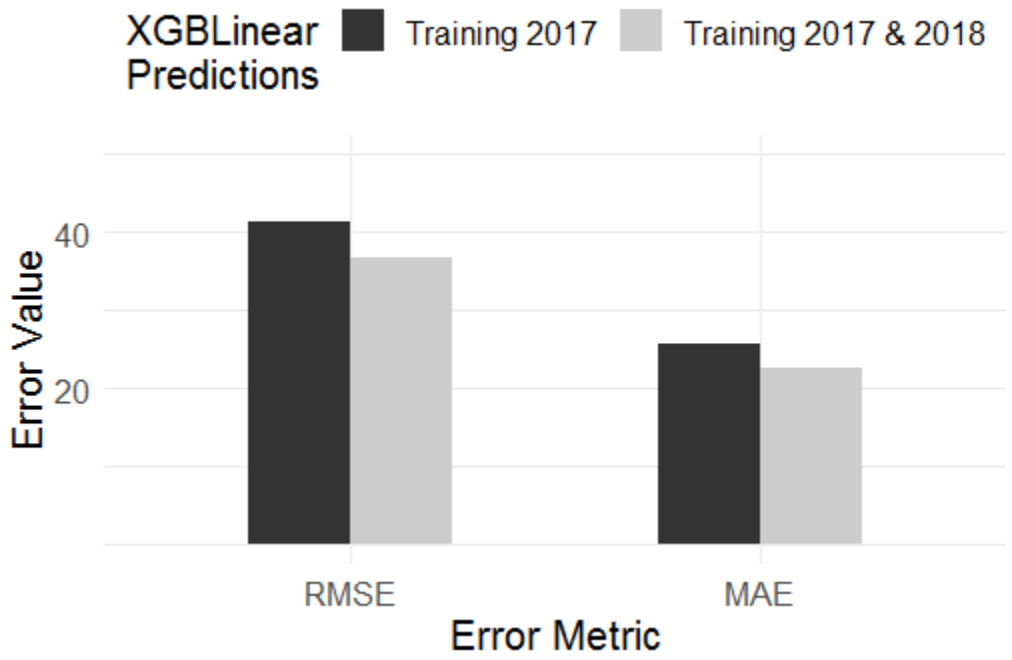


Figure 39: Error comparison between both models based on data utilization.

In Figure 39, we show an error comparison between both models. The second model that trains in 2017 and 2018 data predicted the number of cycling trips with better accuracy. Also, it shows that adding more data to the machine learning model improves the accuracy of the model.

In Figure 40, we show the actual and predicted cycling activities in 2019. The actual data shows an increase in cycling activities in 2019 compared to 2017 and 2018. The model failed to predict the rise in cycling activities in April 2019 after installing the signed routes. Also, it was unable to predict the peak in the summer of 2019. Hence, the machine learning analysis provides an insight into the cycling activities if there were no signed routes, which shows that cycling activities would be lower.

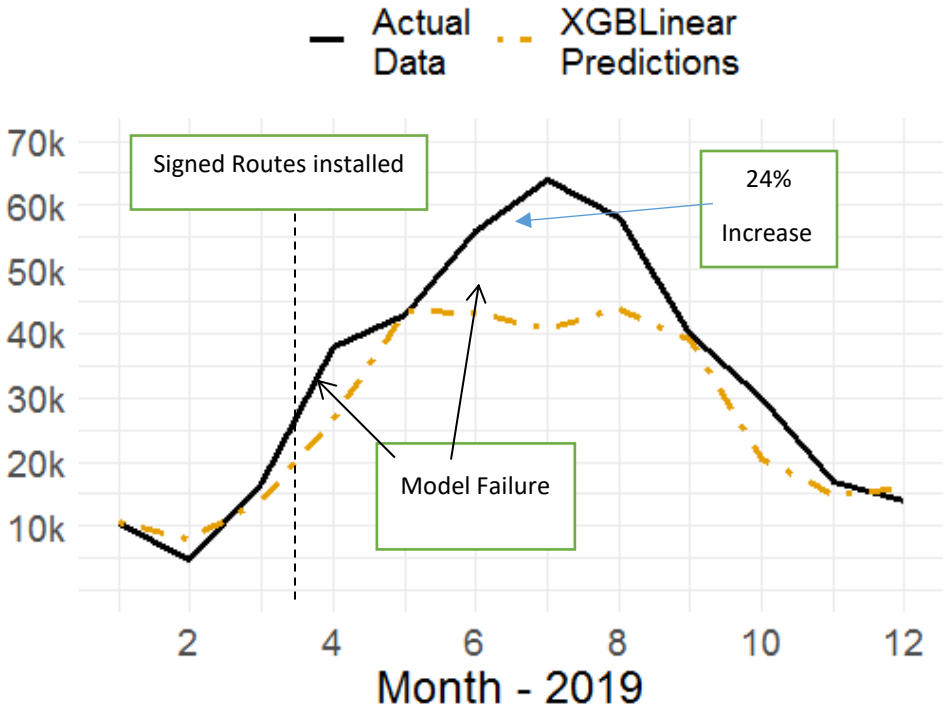


Figure 40: Aggregated cycling trips in 2019 compared to the aggregated predictions of the XGBLinear model, which show projected cycling activities without signed routes.

The analysis provides evidence that cycling trips increased when signed routes were installed. Also, we can see an almost 24% increase in cycling trips in peak months compared to the projected cycling trips provided by the model. As a result, the signed routes had a positive impact on cycling activities, and people are responding by taking these routes.

2.5.3.3 The Most Crucial Factors Affecting the Machine Learning Model

In addition to signed routes, other weather and spatial parameters affect cycling. In Figure 41, we show that temperature is the most crucial factor in predicting cycling activities followed by the population. Additionally, the figure indicates that thunderstorms and rain play a role in decreasing cycling. Also, wind speed is more critical in the winter season because of the wind chill factor, which makes it dangerous to bike in Nebraska. Moreover, it is better to distinguish between street types in cycling analysis. For instance, we can see that separating between major, minor roads, one-way streets affect the model's predictions.

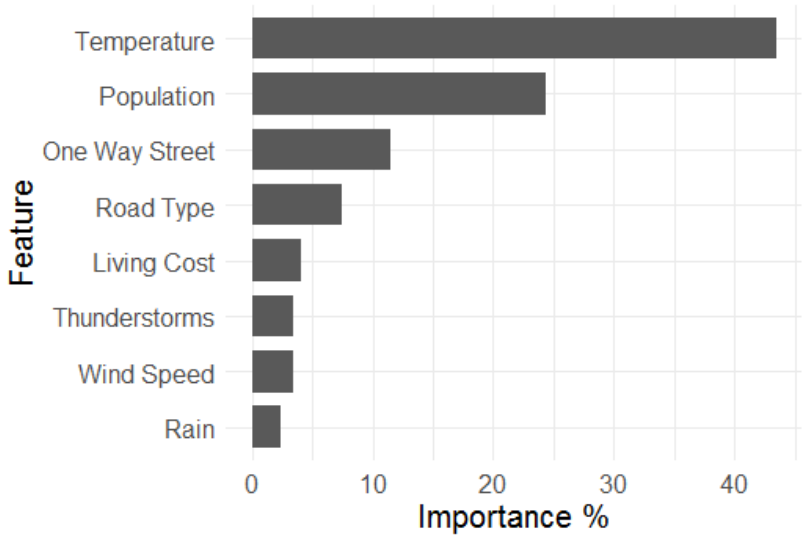


Figure 41: The temporal and spatial factors affecting cycling.

In Figure 41, we showed that weather temperature and population are the most crucial factors to determine cycling activities. Here, we show the effect of removing these predictors from the model. Likewise, we use XGBLinear to predict the cycling activities, which resulted in the least error between all trained models. In Figure 42, we show the error analysis of the model results after removing the temperature and population data from the model matrix.

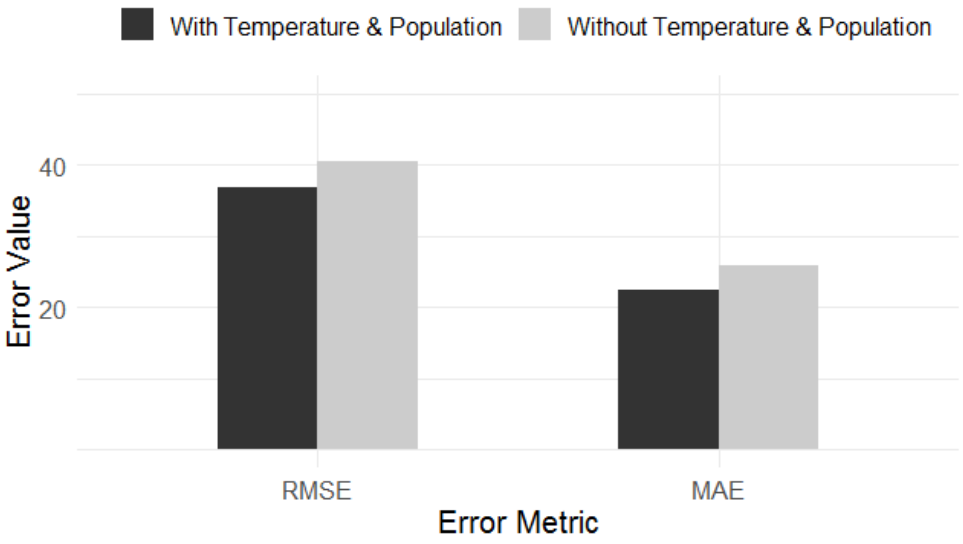


Figure 42: The effect of temperature and population on the predictions error.

Figure 42 shows that error increases after removing the temperature and population data from the model. Also, it confirms the variable importance results shown in Figure 41. Moreover, Figures 41 and 42 show that cycling cannot be analyzed without studying temporal and spatial factors, which influence cycling depending on the location.

2.5.3.4 Summary of Findings

In this section, we provide an in-depth analysis of signed bicycle routes and its effect on cycling activities in Omaha. The signed routes were added in March 2019. The preliminary analysis in Figures 29, 30, and 32 shows an increase in the average monthly activities on those streets. The increase strongly indicates the influence of signed routes on the number of activities. Moreover, we show further analysis by building a machine learning model based on weather, OSM, and demographics data to predict the cycling activities in 2018 and 2019. The model shows accurate results in predicting the 2018 activity because there were no signed routes added at that time. Also, the accurate results boost our confidence in the data used in this model applied to extract the most crucial temporal and spatial factors affecting cycling. Afterward, we use the same combination of data to predict the cycling activities in 2019 without including the signed routes in the model. The results showed that the expected cycling activities without signed routes are lower than the actual events proving that cycling activities increased because of the installation of signed routes.

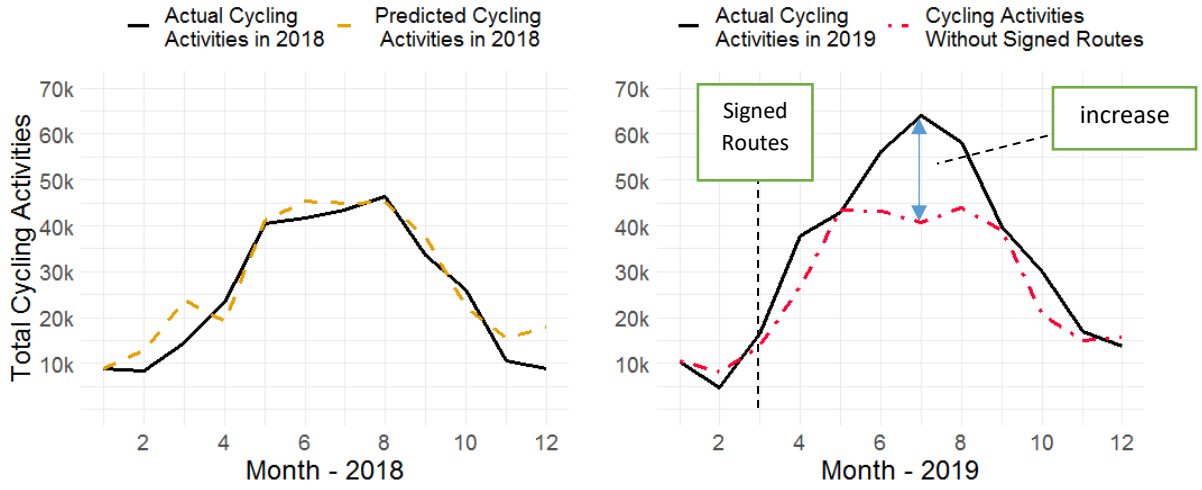


Figure 43: Comparison between 2018 and 2019 cycling activity predictions. a) the accurate predictions of 2018 cycling activities. b) the increase in cycling activities after installing signed routes.

Finally, we show another comparison between 2018 and 2019 cycling activities in Location 1. Figure 44 shows the percentage of yearly increase in the number of cycling activities in Location 1 through 2019.

Figure 44 represents the data on the edge level, showing the yearly increase or decrease in cycling activities. We can see that most of the edges encountered an increase of more than 25% in 2019. In addition to signed routes, the figure shows that other cycling infrastructure types like bike lanes had an increase in activities. A minimal number of edges located on the verge of the cycling infrastructure area experienced a decrease in cycling compared to 2018. Also, many of these edges are sharrows, which shows that sharrows had an insignificant effect on cycling infrastructure.

The results of this section provide an in-depth analysis showing the importance of signed routes in increasing cycling activities in Omaha. Also, it shows that signs are a powerful guiding tool for cyclists.

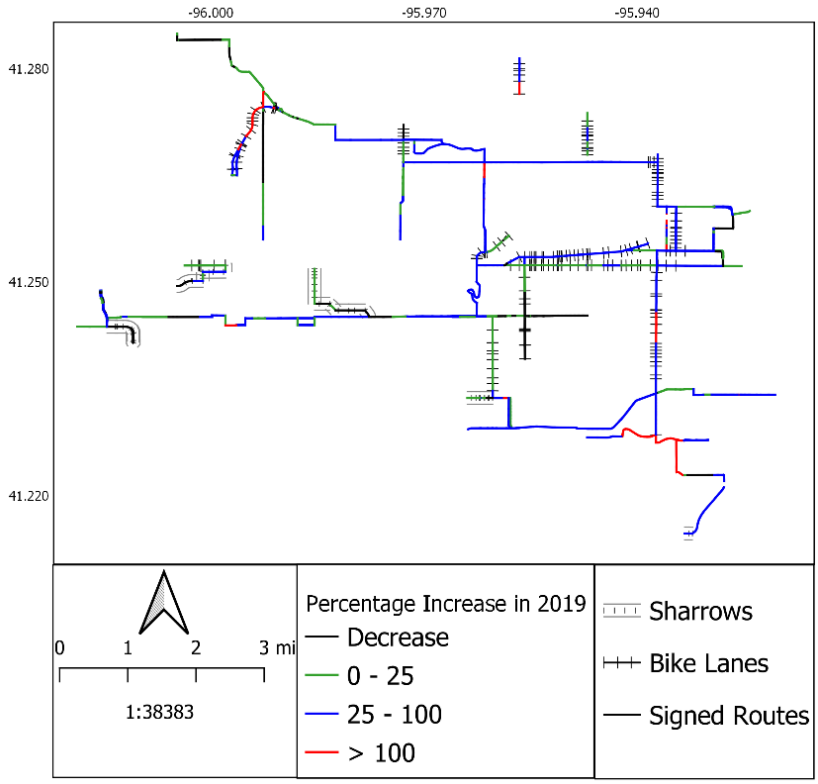


Figure 44: The percentage increase in yearly cycling activities in 2019 on edge level.

2.5.4 Bike Lanes

In this section, we study cycling activity on bike lanes in several locations across Nebraska. However, we limit our study on bike lanes added between 2017 to 2019 because of the time frame of Strava data available. Moreover, we provide a quick study of the effect of signed routes on cycling traffic across streets with bike lanes in Omaha.

2.5.4.1 Omaha Bike Lanes

The lack of bike lanes is apparent in Omaha. However, recently, city planners shifted their approach to install more bike lanes in Omaha streets in the next years. We study the influence of the bike lanes and how it changed the cycling patterns in Omaha. Figure 45 shows the already existed bike lanes in Omaha. Most of the bike lanes in Location 1 existed before 2017, which is out of the Strava data timeframe. Alternatively, there is one street in this location that had bike lanes after July 2017.

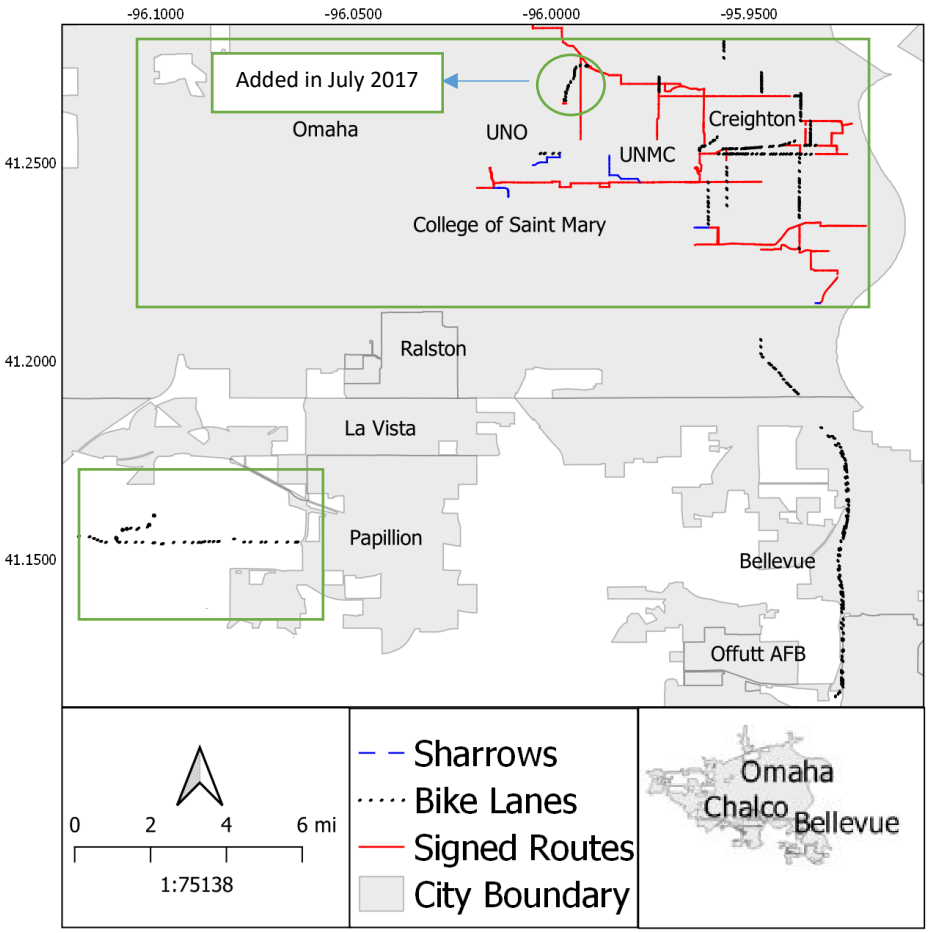


Figure 45: Bike lane locations in Omaha.

In Figure 46, we show the cycling activities on bike lanes in Location 1. The data shows that cycling has not changed between 2017 and 2018. The solid red line suggests that bike lanes installed before 2017 had no change until 2019 when the signed routes were added. Thus, we imply that signs affected the whole area, forming an increase in cycling activities by connecting existing infrastructure. Moreover, the bike lanes installed in July 2017 did not experience any rise of the cycling activities in 2018. However, the average cycling trips almost doubled in 2019, showing another evidence of the importance of having signed bicycle routes. On the other hand, we are not diminishing the effect of bike lanes, but adding bicycle signs to the bike lanes provides a sense of guidance for cyclists. Therefore, it results in increased cycling activities, awareness, and safety.

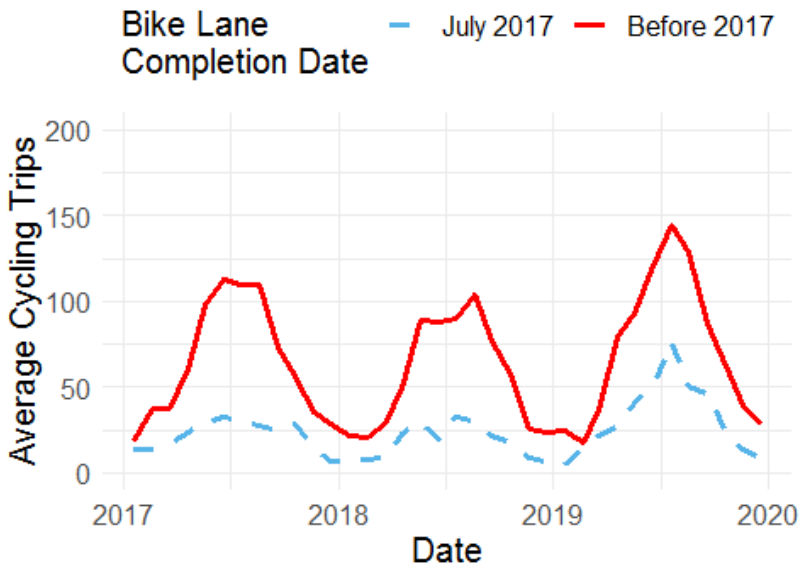


Figure 46: Bike lane average cycling activities in Location 1.

Figure 45 shows two groups of bike lanes in Location 2 near Papillion divided based on the completion year. We study the bike lanes added after 2017 to understand the validity of having bike lanes outside cities on cycling activities. However, we use the yearly data to demonstrate the results to avoid losing cycling counts because of the small area and the distance from the city center. As shown in Figure 47, adding the bike lanes increased the yearly average from almost zero in 2017 and 2018 to more than 400 in 2019. Likewise, we can observe an increase in the bike lanes installed in 2017 through 2018 and 2019, which shows the power of bike lanes and how it increases cycling activities.

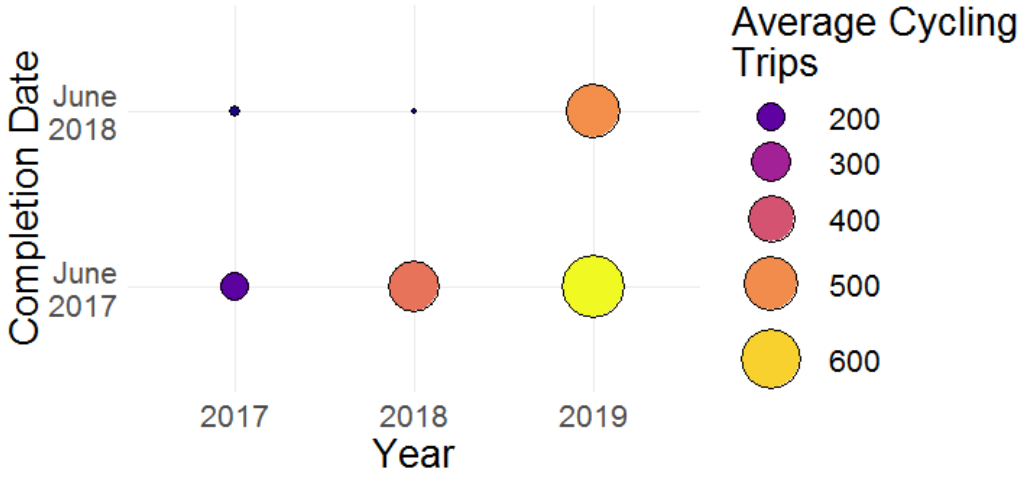


Figure 47: Average yearly cycling activities on bike lanes added after 2017.

2.5.4.2 Lincoln Bike Lanes

The city of Lincoln added bike lanes to 13th street in October 2018 (73). As shown in Figure 48, the city of Lincoln shrank the space for automobile transportation to add bike lanes on both sides of the specified street segments. Bike lanes are known to provide a more convenient space for commute cycling. To get the most benefits of a bike lane, planners should understand the cycling patterns and the routes cyclists usually use to commute.

To evaluate the bike lane spot, we study the monthly cycling counts from Strava during 2017 and 2019. As a result of these bike lanes, cycling trips increased significantly after installing the bike lanes. As shown in Figure 49, the cycling peak tripled in 2019. This increase showing the positive effect of adding bike lanes in this specific area.

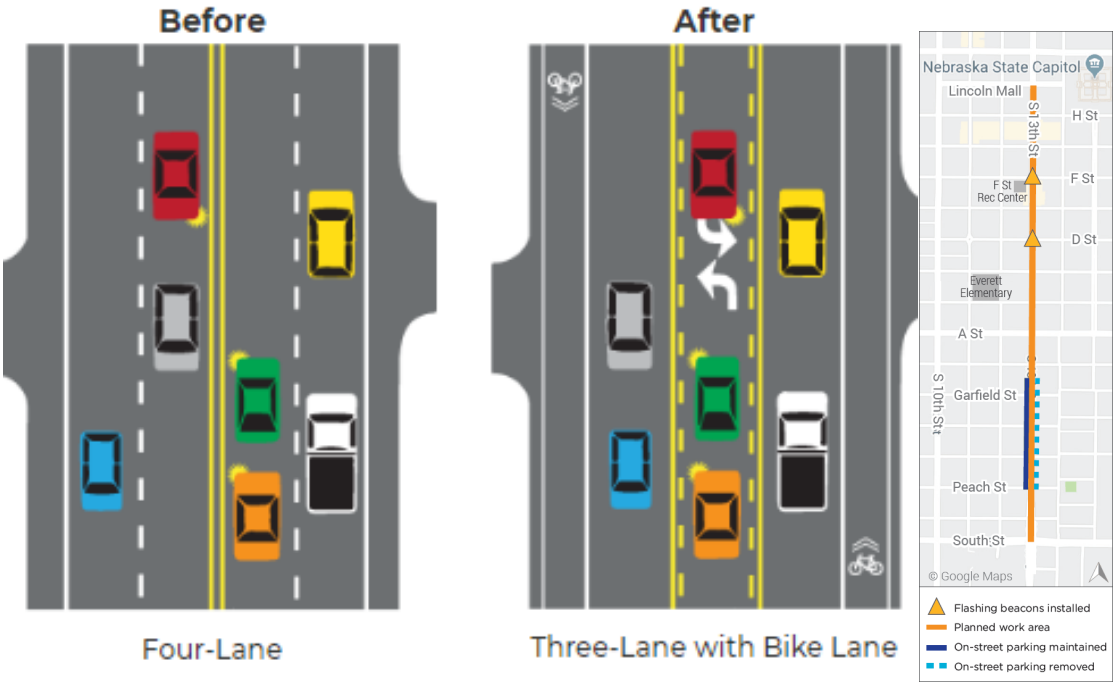


Figure 48: Bike lanes added to 13th Street in Lincoln (73).

Moreover, to prove that bike lanes tend to serve commute purposes. In Figure 50, we show that the number of commute trips is profoundly affected by installing bike lanes, unlike recreational trips, which was slightly affected by the addition of bike lanes.

The increasing cycling and commute trip show that bike lane is a dominant factor in reshaping cycling patterns. Also, city planners should understand the cycling patterns before installing a bike lane.

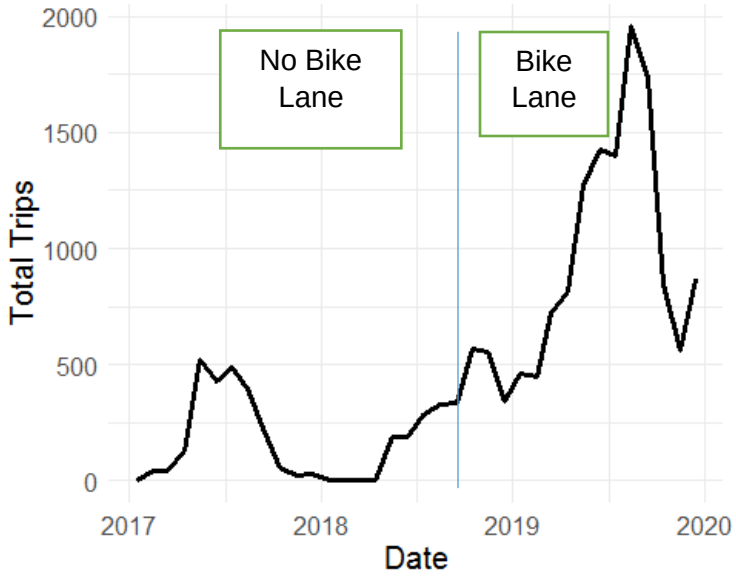


Figure 49: Bike lane effect on cycling activities of the 13th street in Lincoln.

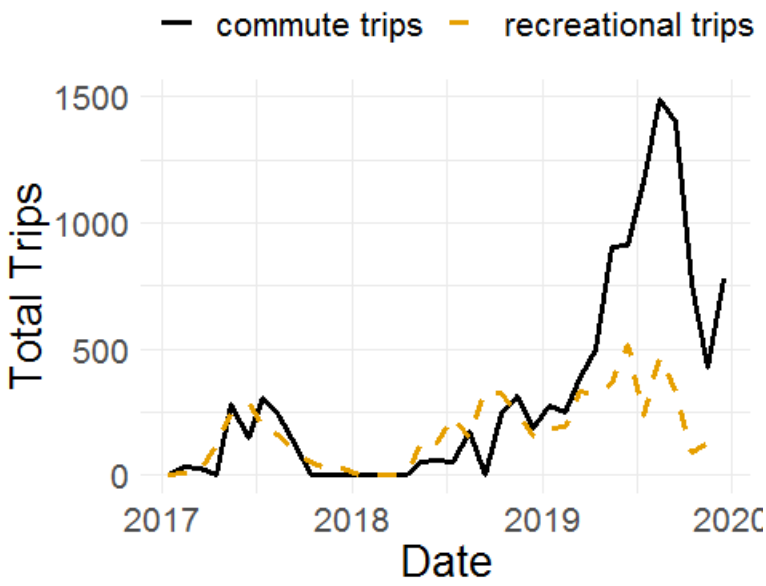


Figure 50: Bike Lane effect on commute and recreational cycling.

2.5.4.5 Trails

A trail is a separate bike or pedestrian pathway from motorized traffic. Unlike bike lanes or sharrows, trails are explicitly isolated infrastructure designated for cyclists and pedestrians. As an example of a multi-purpose trail, Figure 51 shows a picture of the Zorinsky trail in Omaha.

Figure 51: An example of a multi-purpose cycling trail.

Usually, trails cater to recreational activities between cycling, walking, and running because they provide the safest and most convenient option among all other infrastructures. Also, the isolated nature of trails makes it less likely to encounter commute activities as the cycling trip requires longer distance compared to bike lanes. Figure 52 shows the difference between trails and on-street infrastructure commute and recreational cycling percentage in Omaha. Cycling nature on trails tends to be oriented towards recreational cycling. However, almost 25% of cycling activities are commute. This observation shows that trails might encounter commute activities if built to connect essential spots around the city.

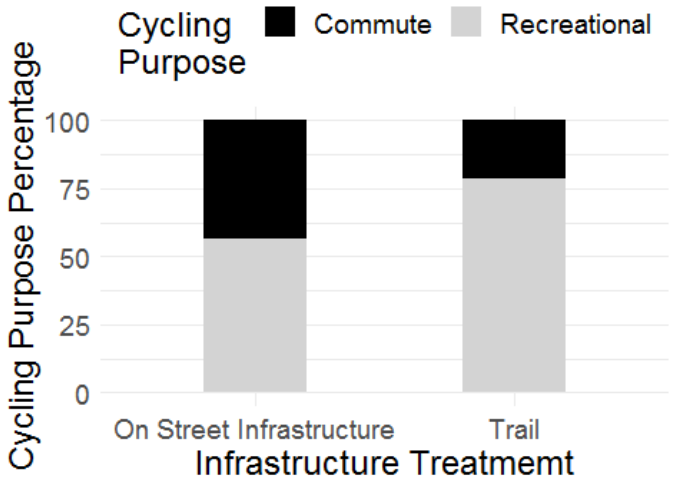


Figure 52: Comparison between trails and on street infrastructure in terms of cycling purposes.

In Nebraska, trails are prevalent among cyclists explaining why the most used edges in Nebraska are identified as trails. For instance, Figure 53 shows the top five cycling trails in Nebraska

compared to the top five non-trail cycling streets. Without a doubt, trails have more cycling activity than other cycling infrastructure.

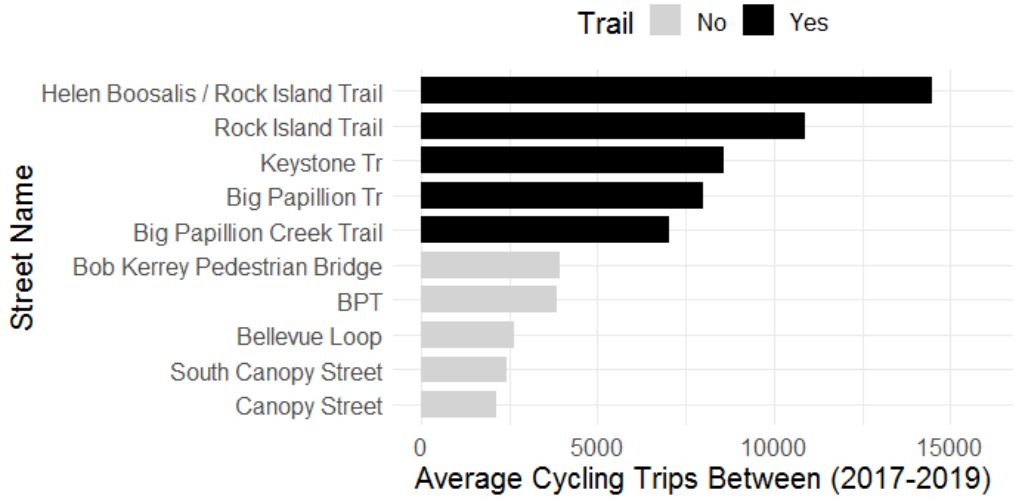


Figure 53: Comparison between the top five cycling trails and the top five non-trail streets in terms of average cycling activities.

Moreover, Figure 54 shows average monthly activities of cycling on trails compared to the overall average of other streets. There is a massive gap between trails and other roads, and it is evident that trails are very favorable to cyclists in Nebraska.

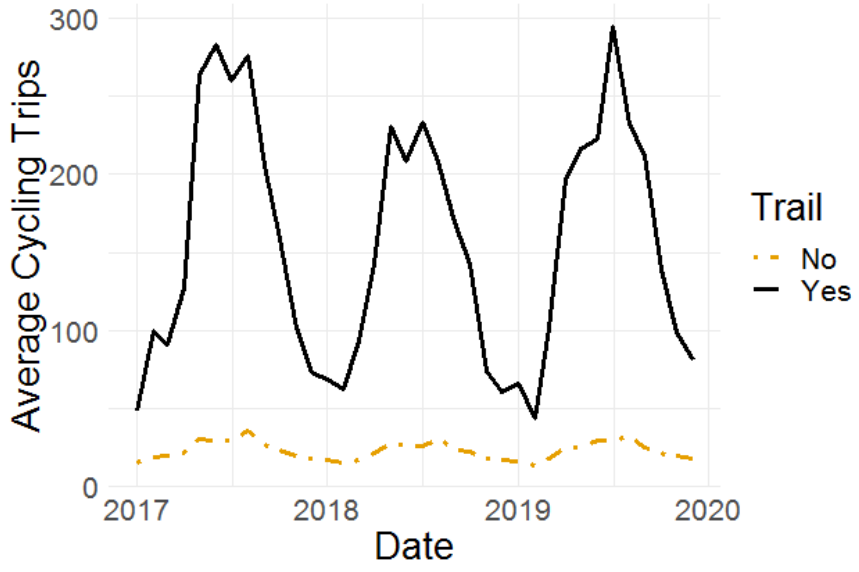


Figure 54: Average monthly cycling activity comparison between trails and other streets in three years between (2017 – 2019).

2.5.4.6 Section Conclusions

This section illustrated a comparison between cycling infrastructures available in Nebraska, showing the effect of each type of infrastructure on the commute and recreational cycling activities. Here, we provide conclusions obtained from this section.

In Nebraska, Trails are the highest bike traffics among all streets in the state. Moreover, the most used trails serve many cycling activities more than any other road in Nebraska. Because of the safety and convenience trails provide, people prefer recreational cycling on trails. Moreover, in Omaha, trails are the second most used infrastructure for commute cycling regardless of being isolated from the grid. Therefore, building more trails in Nebraska will accommodate the whole spectrum of cycling activities.

The commute cycling analysis in Omaha showed that on-street infrastructure is the most popular for commuters, including bike lanes, signed routes, and sharrows. The combined study of Strava and MAPA data showed that sharrows do not have any significance in increasing cycling traffics. On the other hand, Figures 29, 30, and 32 showed an increase in cycling activities on streets where signed routes were installed in 2019.

Moreover, the machine learning model predictions in this section demonstrated the effect of signed routes on cycling. The model showed forecasts of the number of cycling activities with no signs, and it proved that cycling activities increased by almost 24% after installing the signed routes. Furthermore, the comparison of yearly cycling activities between 2018 and 2019, shown in Figure 44, provided further evidence of the effectiveness of signed routes in increasing cycling activities. The figure results showed that most of the edges in Location 1 experienced an increase of more than 25% in cycling activities in 2019. Therefore, we conclude that signed bicycle routes provide guidance and safety for cyclists. Thus, it increases the cycling activities.

In Omaha, the bike lane analysis showed that cycling increased after the installation of signed routes. Also, it showed a considerable increase in the average activities in Location 2, which suggests that people are responding positively to the added infrastructure. Similarly, In Lincoln, we studied the impact of adding bike lanes on 13th street. The analysis showed that cycling activities almost tripled on this specific street. The increase was mainly in commute cycling, which provides a shred of additional evidence towards the importance of bike lanes for commute cycling.

In summary, further installations of bike lanes and signed routes will help establish cycling as a transport method. By increasing the number of cycling infrastructure in Nebraska, people will feel safer. Therefore, it will encourage them to practice cycling habits.

Conclusion

In this project, we acquired cycling data from Strava. To provide a comprehensive study of cycling traffics and patterns, we collected data from different sources. Furthermore, we stored it in a database that includes all cycling related data to utilize the data efficiently.

In section 2.3, we studied the correlation between Strava counts and bike counters in Nebraska. The results showed a strong relationship between Strava and bike counters. Also, by visualizing both datasets, we showed that Strava counts follow the same patterns of bike counter data. In conclusion, the correlation study proved that Strava is a representable dataset of the cycling population, and it is useful for temporal and spatial analysis of cycling traffics in Nebraska.

In section 2.4, we studied the temporal cycling activities. The hourly analysis showed that cycling increases in the hours before and after work on weekdays. Also, peak hours analysis showed that peak cycling hours are highly concentrated before and after working hours on weekdays. On the other hand, on weekends, commute cycling almost disappears, unlike recreational cycling, which increases in the middle of the day. Moreover, recreational peak hours on weekends do not follow a distinctive pattern distributed between 9 am and 4 pm. Furthermore, we showed that commute cycling tends to be shorter in time and distance compared to recreational cycling.

The monthly analysis showed that cycling is profoundly affected by weather conditions. The outside temperature is the primary factor affecting cycling activities. For this reason, cycling tends to increase in warmer weather and significantly decreases in cold weather. In conclusion, the hourly and monthly analysis of Strava data combined with the correlation study provides substantial evidence that Strava data is a robust tool for cycling analysis and planning.

In section 2.5, we conducted a spatial analysis of cycling trips. The study showed that most of the cycling trips are concentrated on trails in Nebraska. We suggested introducing new trails that serve commute and recreational cycling sufficiently to increase cycling mobility. On-street infrastructure is the most suitable for commute cycling in Omaha. For instance, the results of the machine learning models indicated that adding signed routes increased cycling trips significantly.

Furthermore, signed routes are not an exclusive piece of infrastructure and can be used with other types of cycling infrastructure like bike lanes. We showed that cycling trips increased on the bike lane where signed routes are installed. Unlike sharrows, which did not affect cycling traffics because cycling activities did not decrease after removing this type of infrastructure.

Additionally, we showed a rise in cycling activities after installing the bike lanes in Lincoln. This increase indicates that cyclists prefer using bike lanes when they are available, which provides safety and convenience for cyclists.

In conclusion, cyclists are responding positively to the existing and added cycling infrastructure. Also, cycling activities tend to increase in designated cycling infrastructure. Consequently, we recommend adding more signed routes, bike lanes, and trails in Nebraska.

Recommendations and Future Work

Currently, we confirmed that cycling infrastructure has a significant role in increasing cycling activities. We assume this increase happens because people are feeling safer by cycling on a designated infrastructure. However, a complete analysis is required before confirming this hypothesis. Moreover, we recommend building an entire database for cycling, which includes cycling counts, spatial information, and weather data. This database will help researchers to provide more comprehensive studies and accurate results in the future. Finally, we recommend increasing the bike counter with widespread distribution around in Nebraska.

Signed routes can be used to connect existing infrastructure and to inform cyclists and vehicle drivers that cyclists are permitted to ride. The machine learning method we showed can be used to measure the effect of any changes on cycling infrastructure. In this research, we used this method to quantify the effect of signage on the number of cycling trips, which resulted in an increase in cycling trips after controlling for other factors. Moreover, bicycle signage is not an exclusive cycling infrastructure, it can be used with other types of infrastructure like trails and bike lanes, which would increase cycling trips across these infrastructures.

For future research, we recommend using the Strava data for identifying future bike counters locations in urban and rural areas. Strava data is very detailed (temporally and spatially) it can be extracted on an hourly, monthly, and yearly basis, which works well with temporary and permanent bike counters. Also, the data can be used to understand the effect of cycling infrastructure on neighboring streets and areas whether it would increase or decrease cycling activities. Finally, having this data would be beneficial for identifying possible cycling activities in rural areas, which will result in faster treatment of infrastructure in these areas.

References

1. Ogilvie, D., M. Egan, V. Hamilton, and M. Petticrew. Promoting walking and cycling as an alternative to using cars: systematic review. *BMJ*, 2004. 329(7469):763-6.

2. de Hartog, J. J., H. Boogaard, H. Nijland, and G. Hoek. Do the Health Benefits of Cycling Outweigh the Risks*?*. *Environmental Health Perspectives*, 2010. 118(8):1109-16.

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

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

5. Hong, J. D. P. McArthur, and M. Livingston. The evaluation of large cycling infrastructure investments in Glasgow using crowdsourced cycle data. *Transportation*, 2019. 1-14.

6. Pettit, C. J., S. N. Liesk, and S. Z. Leao. BIG BICYCLE DATA PROCESSING: FROM PERSONAL DATA TO URBAN APPLICATIONS. *ISPRS annals of the photogrammetry, remote sensing and spatial information sciences*, 2016. III-2:173-9.

7. Beura, S. K., K. V. Kumar, S. Suma, and P. K. Bhuyan. Service quality analysis of signalized intersections from the perspective of bicycling. *Journal of Transport & Health*, 2020. 16:100827.

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

9. Kriit, H. K., J. S. Williams, L. Lindholm, B. Forsberg, and N. Sommar. Health economic assessment of a scenario to promote bicycling as active transport in Stockholm. *BMJ*, 2019. 9(9):e030466.

10. Aldred, R., A. Goodman, J. Gulliver, and J. Woodcock. Cycling injury risk in London: A case-control study exploring the impact of cycle volumes, motor vehicle volumes, and road characteristics including speed limits. *Accident Analysis and Prevention*, 2018. 117:75-84.

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

12. Loidl, M., R. Wendel, and B. Zagel. Spatial analysis and modelling of bicycle accidents and safety threats. *International Cycling Saftey Conference*, 2015.

13. Meuleners, L. B., M. Stevenson, M. Fraser, J. Oxley, G. Rose, and M. Johnson. Safer cycling and the urban road environment: A case control study. *Accident Analysis and Prevention*, 2019. 129:342-9.

14. Nabavi Niaki, M. S., N. Saunier, and L. F. Miranda-Moreno. Is that move safe? Case study of cyclist movements at intersections with cycling discontinuities. *Accident Analysis and Prevention*, 2019. 131:239-47.

15. Pucher, J., J. Dill, and S. Handy. Infrastructure, programs, and policies to increase bicycling: An international review. *Preventive Medicine*, 2009. 50:S106-25.

16. Saha, D., P. Alluri, A. Gan, and W. Wu. Spatial analysis of macro-level bicycle crashes using the class of conditional autoregressive models. *Accident Analysis and Prevention*, 2018. 118:166-77.

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

18. Zhao, J., J. Wang, Z. Xing, X. Luan, and Y. Jiang. 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*, 2018. 111:119-35.

19. Rahman, M. S., M. Abdel-Aty, S. Hasan, and Q. Cai. Applying machine learning approaches to analyze the vulnerable road-users' crashes at statewide traffic analysis zones. *Journal of Safety Research*, 2019. 70:275-88.

20. Sanders, R. L., A. Frackelton, S. Gardner, R. Schneider, and M. Hintze. Ballpark Method for Estimating Pedestrian and Bicyclist Exposure in Seattle, Washington: Potential Option for Resource-Constrained Cities in an Age of Big Data. *Transportation Research Record*, 2017. Volume 2605: https://doi.org/10.3141/2605-03.

21. Saad, M., M. Abdel-Aty, J. Lee, and Q. Cai. Bicycle Safety Analysis at Intersections from Crowdsourced Data. *Transportation Research Record*, 2019. Volume 2673: https://doi.org/10.1177/0361198119836764.

22. Dong, C., A. J. Khattak, C. Shao, and K. Xie. Exploring the factors contribute to the injury severities of vulnerable roadway user involved crashes. *International Journal of Injury Control and Safety Promotion: Traffic Safety Analysis*, 2019. 26(3):302-14.

23. Vanparijs, J., L. Int Panis, R. Meeusen, and B. de Geus. Exposure measurement in bicycle safety analysis: A review of the literature. *Accident Analysis and Prevention*, 2015. 84:9-19.

24. Chen, C., J. C. Anderson, H. Wang, Y. Wang, R. Vogt, and S. Hernandez. How bicycle level of traffic stress correlate with reported cyclist accidents injury severities: A geospatial and mixed logit analysis. *Accident Analysis and Prevention*, 2017. 108:234-44.

25. Raihan, M. A. Improved Methods for Network Screening and Countermeasure Selection for Highway Improvements. PhD Dissertation, 2018.

26. Wang, H., Y. Wang, M. B. Lowry, C. Chen, and Z. Pu. Bicycle Safety Analysis: Crowdsourcing Bicycle Travel Data to Estimate Risk Exposure and Create Safety Performance Functions. PhD Dissertation, 2016.

27. Cai, Q., M. Abdel-Aty, J. Lee, M. Saad, and S. Castro. Replication Data for: An Assessment of Traffic Safety between Drivers and Bicyclists Based on Roadway Cross-Section Designs and Countermeasures Using Simulation.Transportation Research Board Annual Conference, Washington, D.C., 2018. https://doi.org/10.7910/DVN/FFSR6R.

28. DiGioia, J., K. E. Watkins, Y. Xu, M. Rodgers, and R. Guensler. Safety impacts of bicycle infrastructure: A critical review. *Journal of Safety Research*, 2017. 61:105-19.

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

30. Buehler, R. Bikeway networks. *Transport reviews*, 2016. 36(1):9-27.

31. Bopp, M., D. Si, and D. Piatkowski. *Bicycling for Transportation*. Elsevier, United States, 2018.

32. Ferenchak, N. N., and W. Marshall. The relative (in) effectiveness of bicycle sharrows on ridership and safety outcomes. Transportation Research Board, Washington, D.C., 2016.

33. Wall, S. P., D. C. Lee, S. G. Frangos, M. Sethi, J. H. Heyer, and P. Ayoung-Chee, et al. The Effect of Sharrows, Painted Bicycle Lanes and Physically Protected Paths on the Severity of Bicycle Injuries Caused by Motor Vehicles. *Safety*, 2016. 2(4):26.

34. Speck, J. Do Not Use Sharrows as Cycling Facilities. In Walkable City Rules. *Springer*, 2018. p. 146-7.

35. Abraham, J. E., S. McMillan, A. T. Brownlee, J. D. Hunt. Investigation of cycling sensitivities. Transportation Research Board Annual Conference, Washington, D.C., 2002.

36. Sener, I., C. Bhat, N. Eluru. An analysis of bicycle route choice preferences in Texas, US. *Transportation*. 2009. 36(5):511-39.

37. Pritchard, R. Revealed Preference Methods for Studying Bicycle Route Choice—A Systematic Review. *International journal of environmental research and public health*. 2018. 15(3):470.

38. Hall, W., N. Shadbolt, T. Tiropanis, K. O Hara, and T. Davies. *Open data and charities*, 2012.

39. Romanillos. G., Z. M. Austwick, D. Ettema, and J. De Kruijf. Big Data and Cycling. *Transport Reviews: Cycling As Transport*. 2016. 36(1):114-33.

40. Karoline N. Helleland, and Julie F. H. Stokstad. BIG CYCLIST DATA. Master Thesis. Aalborg University, 2017.

41. Rogers, S., and N. P. Papanikolopoulos. Counting bicycles using computer vision. *IEEE*, 2000.

42. Watkins, K., R. Ammanamanchi, J. LaMondia, and C.A. Le Dantec. Comparison of smartphone-based cyclist GPS data sources. Transportation Research Board, Washington, D.C., 2016.

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

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

45. Jestico, B., T. Nelson, and M. Winters. Mapping ridership using crowdsourced cycling data. *Journal of Transport Geography*, 2016. 52:90-7.

46. Boss, D., T. Nelson, M. Winters, and C. J. Ferster. Using crowdsourced data to monitor change in spatial patterns of bicycle ridership. *Journal of Transport & Health*, 2018. 9:226-33.

47. Heesch, K. C., and M. Langdon. The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour. *Health promotion journal of Australia*. 2016. 27(3):222-9.

48. Strava Metro Data Analysis Summary. Colorado Department of Transportation, 2018.

49. Villamagna, A., L. Getts, R. Young. Active Transportation Accounting: Developing Metrics for Project Prioritization. Publication *FHWA-NH-RD-26962R*, U.S. Department of Transportation, 2019.

50. LaMondia. J., and K. Watkins. Using crowdsourcing to prioritize bicycle route network improvements. Publication 01641227, U.S. Department of Transportation, 2019.

51. Izadpanahi P, Leao Z S, Lieske SN, Pettit CJ. Factors motivating bicycling in Sydney: analysing crowdsourced data. Proceedings of 33rd PLEA International Conference, 2017.

52. Sultan, J., G. Ben-Haim, J. Haunert, and S. Dalyot. Extracting spatial patterns in bicycle routes from crowdsourced data. *Transactions in GIS*, 2017. 21(6):1321-40.

53. Orellana, D., and M. L. Guerrero. 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*, 2019. (7):1314-30.

54. Romanillos, G., and J. Gutiérrez. Cyclists do better. Analyzing urban cycling operating speeds and accessibility. *International Journal of Sustainable Transportation*, 2020. 14(6):448-64.

55. McArthur, D. P., and J. Hong. Visualising where commuting cyclists travel using crowdsourced data. *Journal of Transport Geography*. 74:233-41.

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

57. Haworth. J. INVESTIGATING THE POTENTIAL OF ACTIVITY TRACKING APP DATA TO ESTIMATE CYCLE FLOWS IN URBAN AREAS. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016. XLI-B2:515-9.

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

59. Zhao., J., C. Guo, R. Zhang, D. Guo, and M. Palmer. Impacts of weather on cycling and walking on twin trails in Seattle. *Transportation Research Part D*. 2019. 77:573-88.

60. Haklay, M., and P. Weber. OpenStreetMap: User-Generated Street Maps. *MPRV*. 2008. 7(4):12-8.

61. OpenStreetMap. https://www.openstreetmap.org/about. Accessed Dec 5, 2019.

62. MAPA. https://mapacog.org/about/what-is-mapa/. Accessed Jan 17, 2020.

63. Stonebraker, M., L. A. Rowe. The design of POSTGRES. *SIGMOD record*. 1986. 15(2):340-55.

64. QGIS Development Team. QGIS geographic information system. Open source geospatial foundation project. 2016.

65. Team RC. R: A language and environment for statistical computing. 2013.

66. Neter, J., W. Wasserman, and M. H. Kutner. Applied linear statistical models. 2. ed ed. Homewood, Ill: Irwin; 1985.

67. Shapiro, S. S., M. B. Wilk. An Analysis of Variance Test for Normality (Complete Samples). *Biometrika*, 1965. 52(3/4):591.

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

69. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1974 Jan 1,;36(2):111-47.

70. Chen, T., and C. Guestrin. XGBoost. *ACM*, 2016.

71. Awad, M., and R. Khanna. Support vector regression. in Efficient learning machines. Springer, 2015. p. 67-80.

72. Liaw, A., and M. Wiener. Classification and Regression by RandomForest. Forest. 23. 2001.

73. South 13th Street Improvement Project. Available from: https://lincoln.ne.gov/city/ltu/projects/13th/lincoln-mall-south/.