TRCLC 2017-03
June 13, 2019

Integrating Crowdsourced Data with Traditionally Collected Data to Enhance Estimation of Bicycle Exposure Measure

FINAL REPORT

Valerian Kwigizile, Jun-Seok Oh, and Keneth Kwayu
Western Michigan University

Transportation Research Center
for Livable Communities
Western Michigan University

Technical Report Documentation Page

1. Report No. TRCLC 2017-03	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A
4. Title and Subtitle Integrating Crowdsourced Data with Traditionally Collected Data to Enhance Estimation of Bicycle Exposure Measure		**5. Report Date** June 13, 2019 **6. Performing Organization Code:** N/A
7. Author(s) Valerian Kwigizile, Jun-Seok Oh, and Keneth Kwayu		**8. Performing Org. Report No.** N/A
9. Performing Organization Name and Address Western Michigan University 1903 West Michigan Avenue Kalamazoo, MI 49008		**10. Work Unit No. (TRAIS)** N/A **11. Contract No.** TRCLC 2017-03
12. Sponsoring Agency Name and Address Transportation Research Center for Livable Communities (TRCLC) 1903 W. Michigan Ave., Kalamazoo, MI 49008-5316		**13. Type of Report & Period Covered:** Final Report 8/15/2017 – 5/01/2019 **14. Sponsoring Agency Code:** N/A
15. Supplementary Notes		

16. Abstract

Although many transportation agencies have invested substantially in efforts to improve cycling environments, limitations on methodologies used to estimate bicycle exposure (i.e., volume) continue to impact the decision process. Traditional methods for measuring bicycle volume have been proven to be challenging and costly, especially when planning for non-motorized facilities at network level. One of the potential supplement to traditional methods is to use crowdsourced cycling data. This study explored the potential of incorporating crowdsourced data in estimation methods to improve the spatial-temporal estimation of bicycle exposure. Different probabilistic and machine learning models were tested, including the Negative Binomial (NB) model, Random Forest (RF), Support Vector Machines (SVM), Artificial Neural Network (ANN) and K-Nearest Neighbors (KNN). In terms of prediction, the Random Forest model was found to have a better prediction capability. The addition of Strava counts, which had an average observed penetration rate of 7 percent, improved the RF model significantly by increasing its ability to explain variations in hourly bicycle volume from 65 percent (R-Sqrd = 0.65) to 71 percent (R-Sqrd = 0.71). The study also conducted a simulation study to assess the change in model performance based on different simulated Strava penetration rates and found that a unit change in the percent of simulated Strava penetration rate has a very significant influence on the model's prediction performance. The products of this study can assist planners to make informed decisions currently or in the future by providing them with a reliable method for estimating bicycle exposure.

17. Key Words Bicycle exposure, crowdsourced data, machine learning, modeling	**18. Distribution Statement** No restrictions. This document is available to the public through the USDOT website.

19. Security Classification - report Unclassified	**20. Security Classification - page** Unclassified	**21. No. of Pages** 117	**22. Price** N/A

Disclaimer

The contents of this report reflect the views of the authors, who are solely responsible for the facts and the accuracy of the information presented herein. This publication is disseminated under the sponsorship of the U.S. Department of Transportation's University Transportation Centers Program, in the interest of information exchange. This report does not necessarily reflect the official views or policies of the U.S. government, or the Transportation Research Center for Livable Communities, who assume no liability for the contents or use thereof. This report does not represent standards, specifications, or regulations.

Acknowledgments

1. This research was funded by the US Department of Transportation through the Transportation Research Center for Livable Communities (TRCLC), a Tier 1 University Transportation Center.
2. The data used in this study was, in part, provided by Strava under a limited research license.
3. The authors wish to acknowledge the support from the City of Ann Arbor, City of Grand Rapids, DTE Energy, and the Kent County Parks Department for their assistance during field data collection.

Table of Contents

List of Figures

List of Tables

1 Introduction and Background

1.1 Research background and problem statement

Cycling is increasingly becoming an important mode choice for leisure and work trips (McKenzie, 2014; Statistica, 2016). The benefits of cycling go beyond moving people from origin to destination to improving riders' health and reduce motor vehicle congestion and greenhouse emissions (Cupples and Ridley, 2008). Although many transportation agencies have put more efforts on improving cycling environments, limitations on methodologies used to estimate bicyclist exposure (i.e., volume) impact the decision process. Measuring bicycle exposure is very important for planning bicycle systems as well as ensuring safety of such systems. Traditional methods for measuring bicycle volume have been proven to be challenging and costly. For example, while more transportation agencies are installing permanent count stations (Griffin et al., 2014) which provide excellent data on ridership, these count stations lack spatial details (Jestico et al., 2016). With limited data collected manually or using sensors, several researchers have attempted to develop models for estimating bicycle volume (Buckland and Jones, 2008; Griswold et al., 2011; Molino et al., 2009; Oh et al., 2013). However, such models are less accurate due to limitations in spatial coverage and detail of the data collected manually or using sensors.

Crowdsourced data of cycling activities can be a good source of bicycle exposure measure. Crowdsourced data are collected using GPS-enabled smartphones through fitness apps allowing cyclists to track their routes (Jestico et al., 2016). Data collected using fitness apps have the potential to supplement other data collected through traditional methods to provide spatially detailed data for estimating bicycle exposure. However, comprehensive research on how to integrate crowdsourced data with traditional data is lacking. Understanding opportunities and limitations associated with crowdsourced data is necessary to guide integration of the data. For example, one major limitation of crowdsourced data is the sample bias since those being counted have to opt-in to the program and have to own a smartphone and remember to use the app on each trip (Ryus et al., 2016). As a result, the volume collected through crowdsourcing

can be used to establish minimum and potentially biased volumes at the location. In order to adjust this volume to total volume, additional information such as the proportion of cyclists using the app is needed.

1.2 Research Goals and Objectives

The primary goal of this study was to explore opportunities and document limitations associated with integration of crowdsourced cycling data with data collected using traditional methods to accurately estimate the bicyclists' exposure measure. Specifically, the research had the following main objectives:

[1] Correlate/compare crowdsourced data volumes with manual counts (ground truth) and relate it to the proportion of cyclists using the crowdsourcing technology.

[2] Use crowdsourced data to correlate and estimate bicycle volumes with infrastructure and demographics characteristics.

1.3 Structure of the report

Figure 1.1 shows the general workflow of this research that was used to achieve the research goals and objectives. The workflow summarizes how each of the collected data was used in the analysis. The research begins by reviewing relevant past studies in Chapter 2. The site selection process and types of data that were collected is covered in Chapter 3. The analyses of survey data and the estimation of bicycle exposure are covered in Chapter 4 and Chapter 5, respectively. Conclusions and recommendations are documented in Chapter 6.

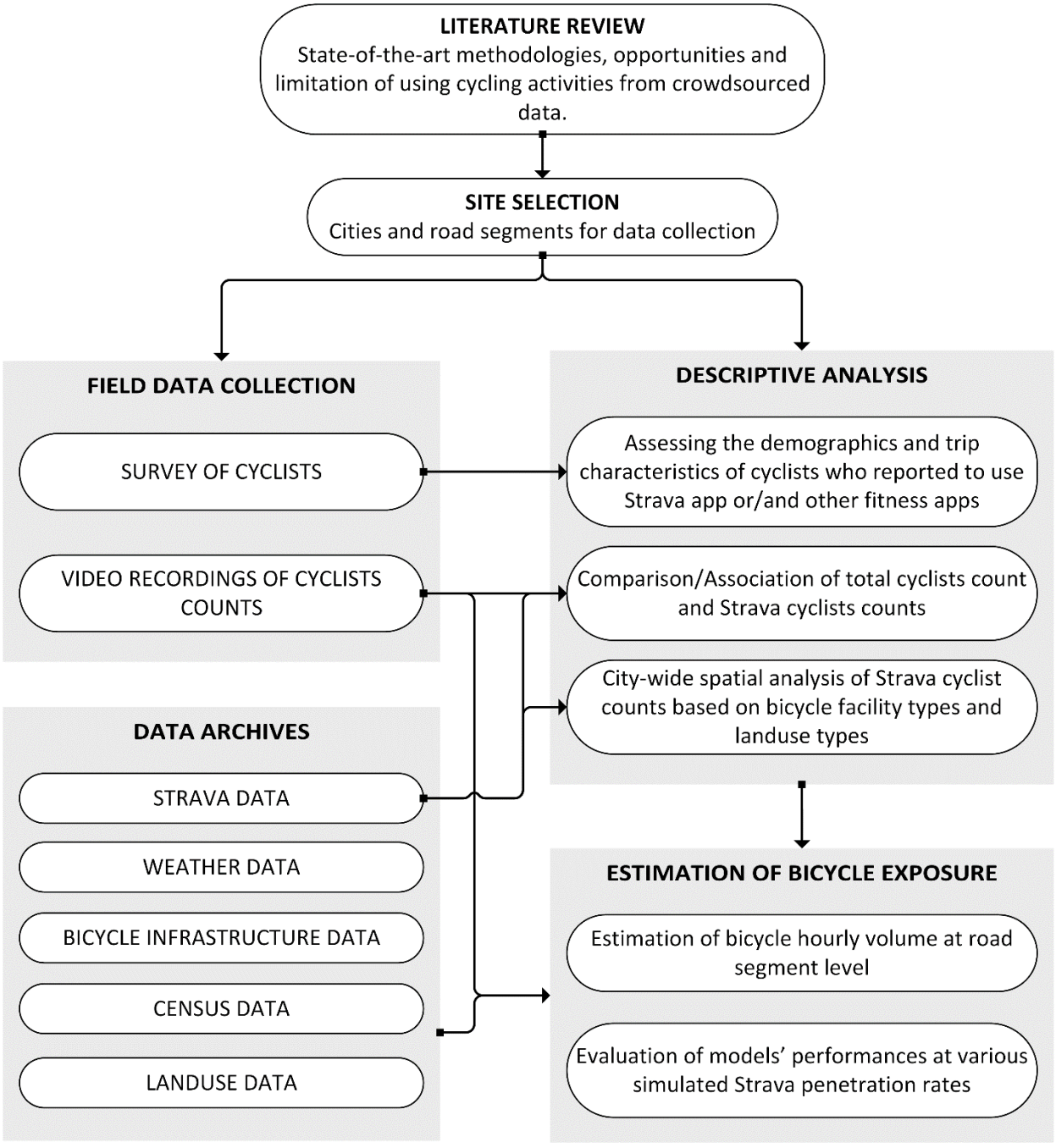


Figure 1.1 Research organization

2 Literature Review

2.1 Crowdsourcing as a term

The term crowdsourcing was coined by Jeff Howe as an act of an institution to outsource some of its functions to a large network of people through an open call (Howe, 2006). Crowdsourcing as a term varies across different disciplines based on its application. Estellés-Arolas and González-Ladrón-De-Guevara (2012) came up with an integrated definition of crowdsourcing based on past literature. They recommended a definition of crowdsourcing as "type of participative online activity in which an individual, an institution, a non-profit organization, or company proposes to a group of individuals of varying knowledge, heterogeneity, and number, via a flexible open call, the voluntary undertaking of a task." The use of crowdsourcing for a given field is mostly online-based, uses open call, clearly define the crowd and the goal of the task. Crowdsourcing has now attracted a lot of attention as it provides rapid and cheaper means of collecting information from a dispersed group of people (Misra et al., 2014). It can be used by organizations to collect data that were expensive to obtain using internal organization expertise. Also, it has been used to obtain information of superior quality and quantity than those provided by professions in the industry (Barbier et al., 2012).

In the planning of active transportation, crowdsourcing offers a wide range of perspectives, data timeliness and direct communication between planners, stakeholders and public at large. Crowdsourcing can be integrated into various active transportation projects such as bicycle master plan, bicycle share maintenance and planning, pedestrian master planning, mobility element performance measures, non-motorized access improvement, bicycle and pedestrian circulation studies, bicycle sharing programs and campus plans. There are various crowdsourcing data tools and sources that can be incorporated in active transportation initiatives. These data sources may take a form of big data, open data, and civic technologies (Smith, 2015).

Crowdsourced data in active transportation exist in different types which are in-situ data, thematic data, thumbtack data and spatial inventory data. Figure 2.1 provides example of crowdsourcing platforms. In-situ data such as Strava and Moves offers real-time geospatial information. Thematic data are usually aggregated in a given geographical area such as American Community Survey and National Household Travel

Survey. Thumbtack data offers points locations on a map each with a given attribute. Spatial inventory data such as OpenStreetMap and Cyclopath comprise of the digital representation of ground features (Smith, 2015).

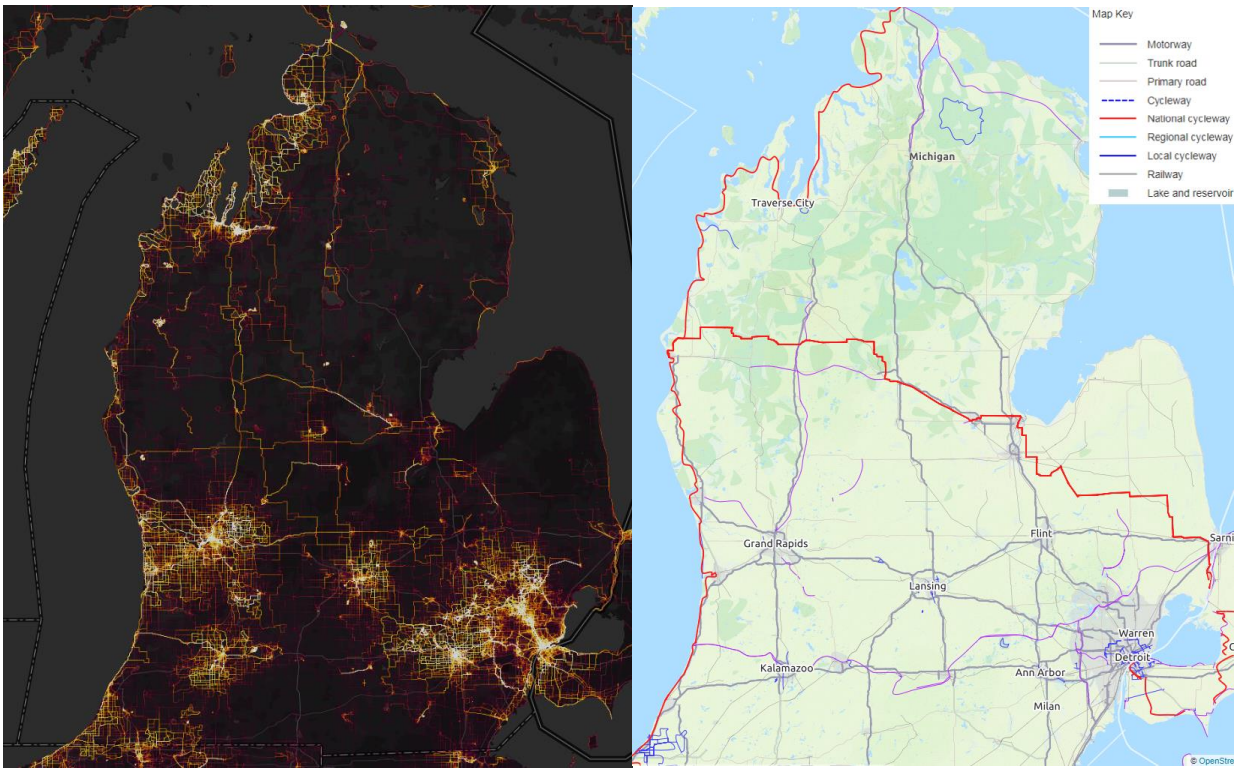


Figure 2.1: Example of crowdsourcing platforms -Strava (Left) and OpenStreetMap (Right)

Sources (www.strava.com/heatmap, www.openstreetmap.org, Accessed April 2018)

2.2 Advantages of crowdsourced data in active transportation

The introduction of crowdsourced data in active transportation planning has offered a unique platform for collecting non-motorized activities. The crowdsourced data can be used as an effective way of offsetting limitations which are inherent in the traditional data collection techniques. Conventional data collection strategies lack spatial and temporal granularity when estimating the bicycle demand (Conrow et al., 2018a; Musakwa and Selala, 2016). Data collected from retrospective surveys and other conventional methods are usually available in an aggregated format and therefore difficult to conduct spatial-temporal analysis (Leao et al., 2017). The conventional data sources have also been

found to be labor intensive which increases the overall cost of acquiring and evaluating the data (Sanders et al., 2017). Also, a time lag has been observed from collection of bicyclists' spatial-temporal activities to dissemination of information to the targeted audiences (Whitfield et al., 2016). Unlike traditional method, crowdsourced information can be disseminated to the intended audience with a high degree of spatial granularity (Whitfield et al., 2016).

In the planning for active transportation, crowdsourced data can provide real-time monitoring of cyclists in time and space (Jestico et al., 2016; Selala and Musakwa, 2016). This offers myriads of proactive approaches in combating high-risk areas for cyclists and making proactive decisions on where and what type of facilities need improvements. This could ultimately maximize the potential benefits in terms of increased ridership, safety and comfort (Blanc and Figliozzi, 2016). Crowdsourced data can also be used to develop robust spatial-temporal exposure measures which are essential in quantifying risks that face cyclists at different roadway locations (Sanders et al., 2017).

The online web-based crowdsourcing tool has made community outreach programs much easier than ever before as it has the ability to collect user inputs in a large spatial extent and within a short period of time (Piatkowski et al., 2015). For example, Delaware Valley Regional Planning Commission (DVRPC) used crowdsourced data in the process of establishing bicycle level of service in Mercer County, New Jersey (Krykewycz et al., 2012). This parameter was essential in formulating a new bicycle master plan. The initial bikeability dataset was improved successfully by using a web-based public and stakeholder outreach program.

2.3 The use of crowdsourced cycling data in studying cyclists' behavior

Researchers have used crowdsourced cycling data to study different aspects of bicyclists' travel behavior. Crowdsourcing cycling activity data is increasingly becoming cheaper than conventional data collection means as the result of the proliferation of smartphone use and low-cost GPS devices. Most of the studies that have used crowdsourced cycling data have focused on the spatial-temporal analysis of cyclist activities at a given geographical unit. Some studies have gone further by integrating crowdsourced data with the traditional counts and retrospective surveys to estimate bicycle demand with the

smallest spatial unit being a road intersection or a road segment. Presented herein is the review of studies that have explored various ways in which crowdsourced data can be used to study bicyclist behavior.

Tracking bicycle activities in space and time is essential for active transportation planners as they can allocate limited resources where they are mostly needed. Crowdsourcing of cycling activities using volunteered geographic information (VGI) have opened cheaper and rapid alternative for tracking cyclists' activities in a larger spatial scale. VGI involve the use of digital tools such as smartphones and web-based applications to collect, analyze and share spatial information volunteered by individuals (Ferster et al., 2018). Sultan et al. (2015) explored the use VGI data-OpenStreetMap (OSM)- for analyzing the bicycle road network in Amsterdam, Netherlands and Osnabrück, German. The open street map contains spatial data on routes where cyclists are allowed or prohibited to cycle. The authors were able to compute the percentage share of bicycle usage for each road type based on total road length. The results showed that majority of the cyclists used roadways that were designated for pedestrians and motorized users. The study demonstrated the usability of VGI from OpenStreetMap despite its known setbacks such as lack of completeness and homogeneity.

In another study, Norman et al (2019) used a fitness tracking application-MapMyFitness to understand how the trail within the reserves are used by mountain bikers, runners and walkers. The VGI from MapMyFitness app was used to supplement existing route data to predict relative popularity and percent composition of mountain biking, running and walking for each trail. The results showed that mountain biking was a more popular activity on the trails followed by walking. Also, more recreational activities (biking, walking and running) occurred more frequently during weekends compared to weekdays.

The cycling data obtain from VGI smartphones apps can also be used to monitor aberrant riding behaviors. Dhakal et al., (2018) used the data obtained from a smartphone application-CyclePhilly to investigate factors that influence wrong-way riding of cyclists in Philadelphia. The commuters riding trips were associated with high likelihood of wrong-way riding while road with bicycle facilities such as shared lane markings, and buffered

lane reduces the wrong-way riding behavior. Further, the odds of wrong-way riding behaviors were high for longer trips which had more likelihood for wrong-way riding.

In Sydney, Australia, Leao et al (2017) investigated factors that promote bicycle ridership using crowdsourced cycling tracking application-RideLog. The RideLog app comprises of crowdsourced information such as road slope, cycling infrastructure, the proximity of the location to parks or coasts and commercial centers. Variables that were found to significantly increase ridership was the proximity of cycling tracks to parks and coastal areas and proximity of the location to the commercial centers.

In San Francisco, California, Hood et al (2011) developed a GPS-based bicycle route choice model. Infrequent bicycle users were observed to prefer riding on a road with bicycle lane while steep slopes were avoided by women and cyclists on commute trips.

In Dresden, German, Fröhlich et al., (2016) developed an app-BikeNow which uses traffic management system to inform cyclists on next green light phases of traffic lights along the track of cyclists and provides a suggestion to a cyclist to either keep or change the current speed. Such kind of information to the cyclists is expected to reduce waiting time at red traffic lights and increase overall riding comfort. In return, cyclists will be motivated to share their spatial-temporal data. The data can then be used in measurement of quality of bicycle traffic and bicycle infrastracture, and identification of high risk intersections.

Several studies have utilized Strava Metro as the crowdsourced data source for studying spatial-temporal bicyclist activities. Strava provides an anonymized and aggregated data package known as Strava Metro. The Strava Metro data has been established to help community in making informed decisions in construction, modification and maintenance of bicycle and pedestrian facilities. Heesch and Langdon (2016) utilized Strava data to evaluate the change in cycling behavior associated with infrastructural changes. The study found that GPS tracking data obtained from Strava app can be used to quantify the short-term changes in cycling near the area where there was an infrastructure improvement. Also, the authors observed a large variation in number of cyclists using the Strava app at different locations. Therefore the comparison of results across multiple locations was not recommended as it could have led to errorneous conclusions. Strava data can be triangulated with other available sources used for

monitoring cycling activities to adjust for the differential use of Strava app across multiple locations.

In another study, Musakwa and Selala (2016) used Strava data to study cycling trends and patterns in Johannesburg, South Africa. Strava Metro was preferred data source as it was available for the whole city. The analysis involved studying the trends of bicycle trips per month and further subdividing the bicycle trips by hour. Also, comparison of trends was conducted by the recreational and commuting trip purposes. A similar study was conducted by Griffin and Jiao (2015) in Travis County, Texas. Strava Metro data was used to determine typical areas that bicyclists will likely ride for fitness purpose based on residential and employment density, land use, presence or absence of bicycle facility and road terrain. Strava data was found to provide useful information that can be integrated into the multi-modal planning and health assessment studies.

In Oregon state, USA, Strava data was been used to understand cycling patterns at the macro level by seasons, day of the week, time of the day and trip purpose (Brandway et al., 2014). At the street level, the disaggregate analysis of bicycle patterns provided insights on where, why and when cyclists ride. For example, by using Strava Metro data, planners identified locations where cyclists prefer to take shortcuts and cyclist's stress level while sharing a space with other road users.

Jestico et al. (2016) used crowdsourced data to quantify and map spatial and temporal variations of cyclists' activities in Victoria, British Columbia. The analysis encompassed the comparison of am-peak and pm-peak manual cyclist counts and crowdsourced cyclists' counts. Strava data was found to have a good representation of cyclists' activities in urban areas and can be used to supplement traditional cyclist counts in estimating cyclists' demand.

A study by Boss et al. (2018) used Strava data to monitor the spatial change of cycling pattern before and after the construction of new infrastructure. By using the Strava data, the authors were able to detect the change in ridership at locations where the construction or temporary closure of cycling infrastructure occurred and spillover effect on the nearby locations. The results demonstrated the usability of Strava in performing city-wide analysis due to infrastructural changes.

McArthur and Hong (2019) used Strava Metro commute trips and Origin-Destination (OD) table of commute trips to understand the pattern of cyclists activities in Glasgow city, Scotland. The Strava link flow was used to locate the most popular links used by cyclists. The OpenStreetMap and Strava OD link flow were used to estimate the expected link flows based on the All-or-Nothing shortest route trip assignment approach. Using the difference between observed and modeled flows, it was possible to identify the most popular routes and unpopular routes. The unpopular routes are the ones that had lower observed flow that would have been expected if the commuters were opting for the shortest routes.

Several studies have compared crowdsourced data with conventional bicycle counts in studying bicycle trends. These types of studies primarily focus on identifying and quantifying samples' representativeness of crowdsourced data. Conrow et al. (2018) compared the spatial pattern of crowdsourced bicycle count data and conventional bicycle count data for the city of Greater Sydney, Australia. The study investigated the representativeness of crowdsourced data on bicycle ridership. Both data sources were found to have higher ridership proportion near the central business districts and other areas where bicycle infrastructures were likely be present.

Revealed preference reported by users are more reliable than stated preference in active transportation planning. Reported state references collected through survey and other conventional means usually have uncertainty on whether correct responses are provided by the participants answering hypothetical questions (Assemi et al., 2015). Surveys have been used widely to obtain bicyclists' opinions based on their prior experiences when using a given roadway or bicycle facility. Crowdsourcing can be used as the platform for gathering real-time operational and safety concerns related to non-motorized facilities. A mobile or web-based online application can allow road users to express their opinions and concerns in real time.

Assemi et al (2015) evaluated the usability of crowdsourced tool to capture revealed preference. The authors used Amazon Mechanic Turk and Advanced Travel Logging Application for Smartphones II (ATLAS) as the crowdsourcing platforms. Participants were asked to state their trip purposes and GPS locations of their trips. The

preliminary results indicated the possibility of using crowdsourcing platforms for collecting revealed preferences from the population.

A more practical study was conducted by Blanc and Figliozzi (2016) using a smartphone application (ORcycle) developed by the Oregon Department of Transportation to gather information pertaining to ridership and revealed preferences. Bicyclists were asked to score different roadway bicycle facilities based on the level comfort and safety that they experience while riding on those facilities. The data was then used to model the reported cyclists' comfort level. Factors that were found to decrease cyclists' stress include the provision of bicycle boulevard and separated bicycle path. The authors further demonstrated how these factors can be promoted to increase ridership.

Crowdsourced data which comprises of public ideas and preferences has also been used to investigate various ways of improving bike sharing systems. Piatkowski et al (2015) investigated the usability of community feedbacks via online crowdsourcing tool to improve the distribution of bicycle sharing stations. Four cities which have bicycle sharing programs that incorporate web-based community outreach were used in the study. The cities were Philadelphia, Pennsylvania; Chicago, Illinois; Cincinnati, Ohio; and Portland, Oregon. The number of proposed bike sharing stations via crowdsourcing was found to be significantly affected by travel mode to work, race, and ethnicity. Census block groups with a higher number of people cycling and walking were associated with higher number of proposed bike sharing stations. A similar study was also conducted in Cincinnati's Ohio (Afzalan and Sanchez, 2017). The study focused on the use of crowdsourced information to augment citizens' participation in selecting desired locations for the bike sharing stations. Further, the data was used to forecast expected demand, cost and revenue. Organizational factors that were found to affect the utility of crowdsourced information include the capability of organization to analyze crowdsourced data, the perception of planners about the value of the crowdsourced data and the extent to which organization facilitate citizens engagement.

In another study, Wu and Frias-Martinez (2015) used the crowdsourced approach to improve the accuracy of bike travel time provided by Google application for Washington D.C. bike sharing system. By accounting for slope and trip distance, they developed a crowdsourced predictive model that improved the accuracy of Google's biking time by 5

percent. The ground truth for the model validation was the log data collected each time a user rented a bicycle.

Crowdsourced data sources can be used to supplement traditional counts in the estimation of bicycle volume. Hochmair et al. (2016) estimated the commuting and non-commuting bicycle trips as bicycle kilometer traveled (BKT) using Strava Metro data as one of the explanatory variables. Strava Metro bicycle counts were found to be useful in estimating number of trips by purpose and time of day because of its spatial and temporal granularity. Weekday and weekend models that included Strava Metro data as one of the explanatory variable had better performance than models that excluded Strava Metro data. A similar approach was used by Sanders et al (2017) to estimate pedestrian exposure for the city of Seattle, Washington. The addition of Strava Metro data as one of the explanatory variables in the model increased the explanatory power of the model from 57 percent (Pseudo R^2=0.57) to 62 percent (Pseudo R^2=0.62). Further, it reduced the complexity of the model which is an essential factor for model transferability to other similar locations. In another study, Haworth (2016) estimated the bicycle flow in urban areas using Strava data with emphasis on investigating its representativeness and potential bias. The estimated flows using Strava data were compared with the London cycle census data (LCC) while controlling for road type, hour of the day, day of the week and presence of bicycle lane. The Strava Metro data was found to be a significant predictor of total cyclist flow with the estimated R-squared value of 0.7.

2.4 Integrating conventional and crowdsourced data sources

Most of crowdsourced data have inherent bias toward a specific portion of the biking population. A convenient way of offsetting such bias will be to fuse the crowdsourced data with other data sources. Griffin et al (2015) used conventional data obtained from bicycle count data, GPS survey data and Strava data to understand how effectively bicyclists can be monitored in real time using multiple data sources. Griffin's study quantified how GPS survey differs from Strava Metro data by trip purpose. It also identified the proportion of bicycle volumes for each land use type that was represented by each data source. Another study compared four U.S cities (Austin, Denver, Nashville, and San Francisco) at Census block level, on the number of active commuters in Strava Metro data and

number of active commuters that were reported in U.S Census Bureau's American Community Survey (Whitfield et al., 2016). Higher correlation between the datasets was observed in high population density areas.

Spatial-temporal distribution of bicyclists' activity data from GPS can be integrated with reported bicycle safety data to provide better and reliable estimates of bicyclist's risk levels. Strauss et al. (2015) mapped cyclists' activities extracted from GPS data in conjunction with cyclists' injuries in Montreal, Canada. Short-term bicycle counts obtained from conventional count data and long-term GPS data were combined to determine the bicycle crash risks at signalized intersections, non-signalized intersections and along the roadway segments.

2.5 Modeling techniques in fusing conventional and crowdsourced cyclist's activities

Crowdsourced data only represent a sample of people walking or cycling. Therefore, it can be fused with other data sources such as traditional manual bicyclists' counts or retrospective surveys to obtain results which will be generalizable to the total cyclists' population. There are several techniques for fusing multiple sources of spatial data. Lesiv et al. (2016) compared several methods for fusing spatial data namely nearest neighbor, naive bayes, logistic regression, geographically-weighted logistic regression (GWR), as well as classification and regression trees (CART). Minor difference in performance was observed across the methods with GWR showing a slightly better performance than the other methods.

Proulx and Pozdnukhov (2017) used geographically weighted data fusion (GWDF) technique whereby four datasets including Strava Metro were combined to provide a better estimate of bicycle flow at segment level. The most computationally intensive component of this study was homogenization of the datasets to have the same spatial and temporal scale since each dataset had its own spatial scale and temporal resolution. Each data source was compared with ground-truth bicycle counts. Strava Metro data had a better coefficient of determination with ground-truth bicycle counts compared to other data sources. As expected, the model with a better prediction performance was obtained after combining multiple data sources. The main assumption in the Proulx and

Pozdnukhov study was that each data source represented a specific segment of bicyclist users. This assumption aligns with the study by Watkins et al (2016) that compared Strava data with the agency-monitored smartphone (Cycle) data in Atlanta. It was found that the two data sources represented different population segments based on gender, age, percent of commute trips, trip lengths, and location of bike paths.

Different regression models have been used to incorporate Strava data as predictor variables in estimating bicycle volume. Hochmair et al. (2016) estimated bicycle trips for Miami-Dade county using Strava data as one of the predictor variable in the regression equation. The regression equation comprised of eigenvector spatial filter to account for spatial autocorrelation and biases in parameter estimation. Other sociodemographic and location-specific variables were also included in the model. Sanders et al. (2017) used Poisson regression with robust standard errors to estimate the average annual daily bicyclist volume (AADB) using annual Strava Metro counts as one of the exogenous variables. The mode had a decent performance with the Pseudo R-squared of 0.568.

Jestico et al. (2016) used generalized linear equations to establish the relationship between crowdsourced data from Strava Metro with the bicyclists' manual count data. A coefficient of determination increased from 0.40 to 0.58 when larger time windows were used for aggregating the data. In terms of representativeness, one Strava user was found to represent fifty-one riders in the total population. Cyclist volume was also predicted into low, medium and high level using Generalized Linear Model (GLM) with a Poisson distribution link. Strava data in conjunction with other explanatory variables such as slope, traffic speed, on-street parking and time of the year were used to predict the bicycle activity level.

An improved modeling approach is needed to quantify reported risk and perceived level of comfort experienced by cyclists when using the road network. Blanc and Figliozzi (2016) used ordinal logistic regression to model the cyclist's level of comfort as a function of facility type, trip characteristics and trip stressors. The data were obtained from the smartphone application (ORcycle) which was designed to collect statewide information about the bicyclists' safety data, crash data, travel and perceived level of comfort (Figliozzi and Blanc, 2015).

In another study, Jestico et al (2017) used a negative binomial regression model to investigate the relationship between bicyclist incidents reported through web-based crowdsourced approach (BikeMaps.org) with roadway infrastructure characteristics. Further, Branion-Calles et al. (2016) used logistic regression to compare the odds of crowdsourced near miss data relative to crowdsourced collision data. Furthermore, the comparison was made between the odds of crowdsourced collision data relative to collision data from the insurance reports. The results indicated higher odds of reporting crowdsourced near miss incidents than crowdsourced collision incidents for commute trips and at locations without bicycle facility. Also, high odds of crowdsourced collision reports, as opposed to collision data from insurance reports, were linked with peak traffic hours, midblock locations and routes with bicycle facility.

Also, different modeling techniques have been used to incorporate crowdsourced data to improve bike sharing systems. Piatkowski et al. (2015) used the hierarchical linear regression model to investigate the relationship between the proposed number of bike sharing stations collected via web-based crowdsourcing tool and sociodemographic characteristics of the community at census block group. Wu and Frias-Martinez (2015) used random forest and support vector machine to predict the crowdsourced biking time using Google biking time for the Capital bike share system in Washington D.C. The predictions were adjusted for slope and distance to improve the accuracy.

2.6 Challenges associated with the use of crowdsourced data

A major challenge of crowdsourced data that has been documented in literature is lack of sample's representativeness. It occurs when characteristics of the participants who are volunteering information do not represent the population characteristics. Therefore, the data obtained from crowdsourced data can be selective, ultimately introducing a bias when used by planners who are striving to make an equitable distribution of resources and services. The degree of bias in decision making and resource allocation using crowdsourced data will largely depend on the characteristics of the participants volunteering the information such as level of internet access and technological literacy, public awareness and how the participants were recruited (Blanc et al., 2016; Piatkowski et al., 2015). For example, Strava data have been reported to have inherent sample bias

towards cyclists who are recreational riders more than commuters or utilitarian riders. Also, Strava data has been found to be skewed towards male cyclists (Lee and Sener, 2019). Leao et al (2017) found that demographic information collected using RiderLog for crowdsourcing cycling data was biased towards urban populations. This aligns with Piatkowski et al (Piatkowski et al., 2015) study which found that crowdsourced data can be highly biased towards communities with certain socio-demographic characteristics. For example, in investigating bike sharing program it was found that the proposed bike sharing stations via crowdsourcing were biased towards white populations. In addition, Blanc et al., (2016) found that bicycle activity data collected from the smartphone applications underrepresents females, older adults and low-income population.

Privacy agreements between the participants and data vendors make it difficult to obtain individual information of cyclists such as age and gender from crowdsourced data. This limits its utilization and integration with other conventional data sources. Further, it limits the evaluation of sample representatives as most of the participants demographic characteristics are provided in aggregated format. In addition, most of the crowdsourced data may lack quality check before being disseminated to the planners and public at large (Ferster et al., 2018; Smith, 2015). This may result into inaccuracies and misinterpretation of the results. The GPS-based apps that are used to track cycling activities depend on different hardware configurations which can potentially introduce variability in data quality (Dhakal et al., 2018).

All these biases and limitations in crowdsourced data should be properly outlined to avoid misinterpretation of results (Boyd and Crawford, 2012). The spatial and temporal resolution of crowdsourced data cannot be leveraged effectively in active transportation planning without a proper understanding of sample's representativeness of cyclists' population. Combining crowdsourced data with other data sources will enhance sample representativeness and hence improve the data usability. The quality check and privacy skeptics of crowdsourced data can be addressed by using advanced analytical approaches for analyzing and processing crowdsourced information (Barbier et al., 2012). For instance, there has been a huge effort in anonymizing the data for privacy issues without distorting the raw information (Leao et al., 2017). The technological advancement

in data security will likely improve in the future and in turn boost public engagement in decision making through various crowdsourcing platforms.

3 Site Selection and Data Collection

This chapter describe the procedures and methods used to acquire both archived and new field data. First, the chapter covers the selection process of sites (i.e., roadway segments) used for field data collection. The chapter then expands by providing the descriptive statistics of all data types that were used in the estimation of bicycle exposure to better understand the pattern and distribution of the data. For each data type, challenges that were faced while collecting and processing the historical and field data are documented.

3.1 Study site selection

One of the major data collection tasks of this project was to collect the cyclists' activities on the selected roadway midblock segments. Inadequacy of spatial-temporal information of cyclist counts is a well-known and documented limitation. To circumvent the situation, the research team resorted to collect the field bicycle count data. The two cities in Michigan, namely Ann Arbor and Grand Rapids, were selected as the case study areas for the study. For each city, a careful and planned site selection process was conducted based on data that were available in all roadway locations. The selection process was mainly based on landuse, roadway type and bicycle facility information. The land use types were grouped into four major categories, which were commercial, institutional, residential and recreational. All the roadway segments were also categorized by a specific type of bicycle facility available along each roadway segment. The bicycle facilities were divided into four major groups, which were bike lane, shared lane marking, trail and no bicycle dedicated facility. In locations where cyclists had no dedicated bicycle facility, either sidewalk or roadway shoulder was available. Roadway type information that was used followed the national functional classification relevant to our study i.e., arterial, collector and local. Each roadway type represented a group of roadways with similar geometric and traffic volume characteristics. The available historical cyclists count information at some areas of Ann Arbor and Grand Rapids were considered during site selection to ensure that the selected sites were a good representation of the overall cyclist activities in each city. A total of 19 roadway segments were selected for field data

collection in the two cities. Site details for each city are shown in Table 3.1 and Table 3.2 for Ann Arbor and Grand Rapids, respectively. Figure 3.1 and Figure 3.2 show the spatial distribution of the sites in Ann Arbor and Grand Rapids, respectively.

Table 3.1: Selected sites in the city of Ann Arbor

No	Segment Name	Land use	Bike facility	Latitude	Longitude
1	5th Street	Commercial	Bike Lane	42.278684	-83.746129
2	Murfin Road	Institutional	Shared Lane	42.296415	-83.719707
3	Huron Pkwy	Residential	None	42.281959	-83.765323
4	Division Ave	Institutional	Bike Lane	42.275278	-83.744247
5	Platt Rd	Residential	Bike Lane	42.243612	-83.700060
6	Nixon Rd	Residential	Bike Lane	42.317163	-83.707602
7	Plymouth Rd	Institutional	Bike Lane	42.302561	-83.705764
8	Miller Rd	Residential	Shared Lane	42.283312	-83.750237
9	State@Liberty	Commercial	Shared Lane	42.279812	-83.740804
10	State@Packard	Commercial	None	42.270327	-83.740595

Table 3.2: Selected sites in city of Grand Rapids

No.	Road Name	Land use	Facility	Latitude	Longitude
1	Cherry St SE	Commercial	Bike lane	42.9594	-85.6586
2	Grandville Ave SW	Commercial	Shared lane	42.9603	-85.6735
3	Lake Dr SE	Residential	None	42.9544	-85.6299
4	Monroe Ave NE	Residential	Bike lane	43.0103	-85.6665
5	N Park St NE	Institutional	Bike lane	43.0224	-85.6602
6	Walker Ave NW	Institutional	Bike lane	42.9832	-85.7021
7	White Pine Trail	Recreational	Trail	43.0026	-85.6708
8	Oxford Street Trail	Recreational	Trail	42.9559	-85.6861
9	Kent Trail	Recreational	Trail	42.9506	-85.7095

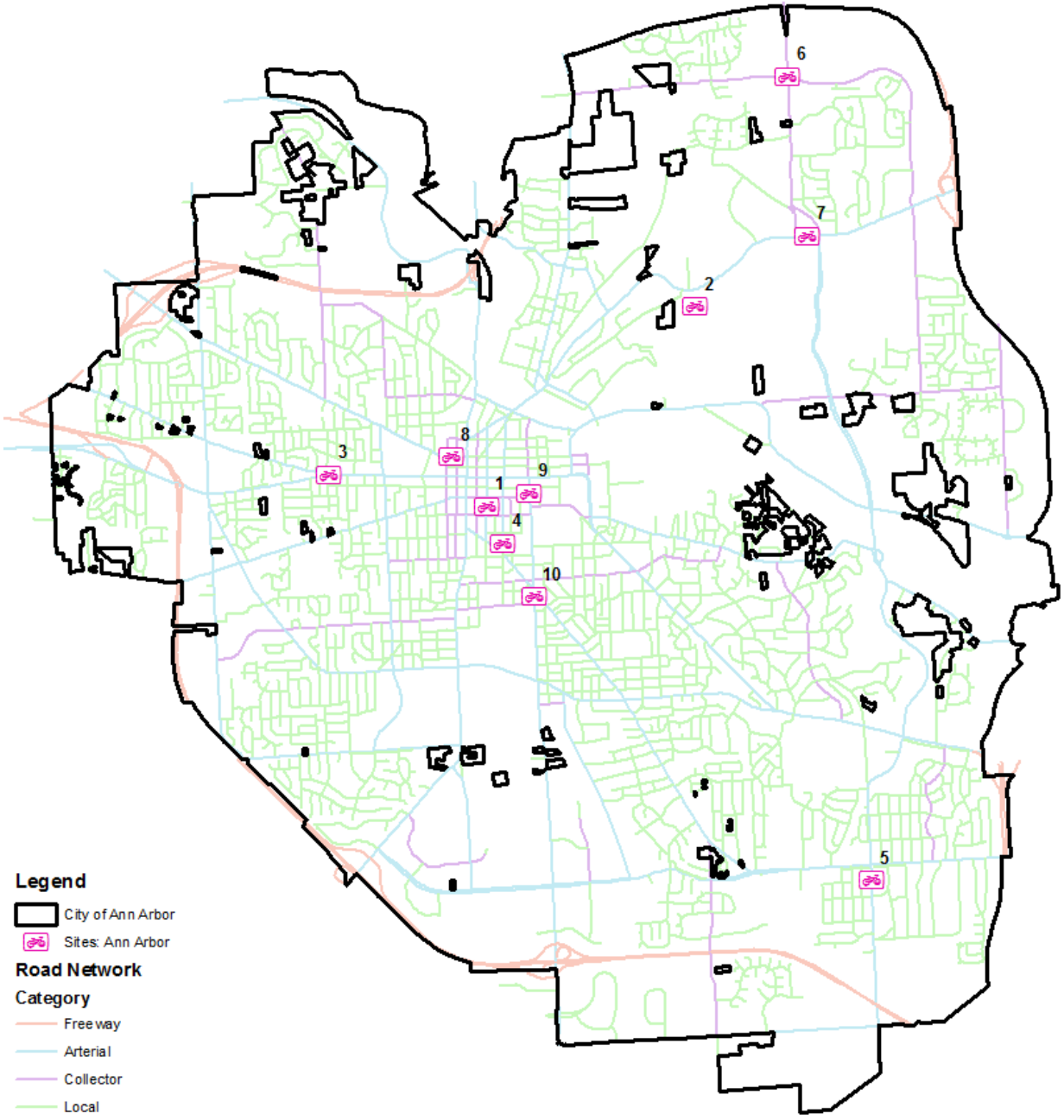


Figure 3.1: Spatial distribution of selected sites in Ann Arbor

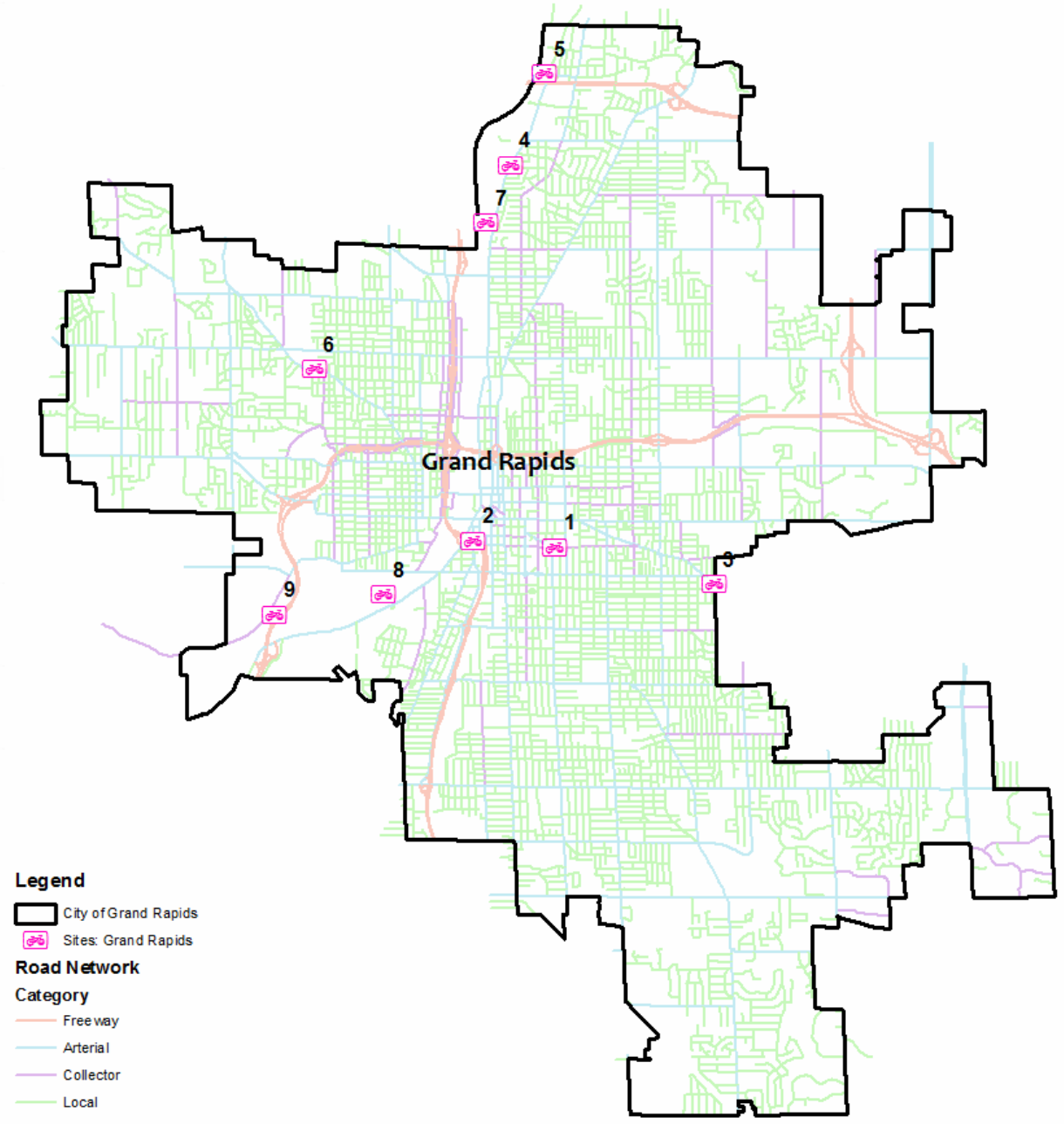


Figure 3.2: Spatial distribution of selected sites in Grand Rapids

3.2 Data Collection

Different types of data were collected to assist in the estimation of bicycle exposure. The main data types that were collected in this project were;

- Total cyclists count for the selected roadway segment,
- Cyclists' activities from Strava Metro data for each roadway segment,
- Weather data at hourly level,
- Bicycle facility data,
- Census data at block level,
- Landuse data, and
- Survey data.

3.3 Total cyclists count

The total cyclists count data were collected using video recordings at 16 sites and pneumatic tube counters at three sites. The video data were collected for one week continuously at each site. About 1,520 hours of video data were recorded at the 16 sites. The video camera had a feature which provided date and time stamp for each recording. This information was essential for data preprocessing as it enabled temporal match with other data types such as Strava data. A good location was selected for mounting the camera to facilitate the recording of cyclists in both directions of the road. The camera mounting height had to be greater than 6 feet for compliance with city regulations. In most cases the research team utilized the utility poles with the permission from the respective owner (i.e., city or utility company). The camera was firmly secured to the pole using fasteners and lockers to prevent it from vibration caused by wind and potential vandalism. The demonstration of how the camera were installed at the sites is shown in the Figure 3.3 below. Figure 3.4 and Figure 3.5 provide examples of video recordings at the site located along Monroe Avenue in Grand Rapids and along Nixon Road in Ann Arbor, respectively. Details about data collection are available in Appendix 8.1 – 8.4.

Figure 3.3: Installation of cameras on sites with bike lanes or shared lane markings

Figure 3.4: Position of the camera relative to the flow of cyclists in Monroe Ave, Grand Rapids

Figure 3.5: Position of the camera relative to the flow of cyclist in Nixon Road, Ann Arbor

The bicycle tube counters were installed across the White Pine trails, Oxford Trails and Kent trails in Grand Rapids. The bicycle tube counters are suitable for such locations as there were no vehicles in pedestrians and cyclists mix. The majority of people who were using the trails were walkers, runners and cyclists. The tube counters were installed for a week on each site to capture hourly variation within a day and daily variation within a week. Figure 3.6 and Figure 3.7 illustrate the position of bicycle tube counters relative to the movement of cyclists on trails.

Figure 3.6: Bicycle tube counters on White Pine Trails, Grand Rapids

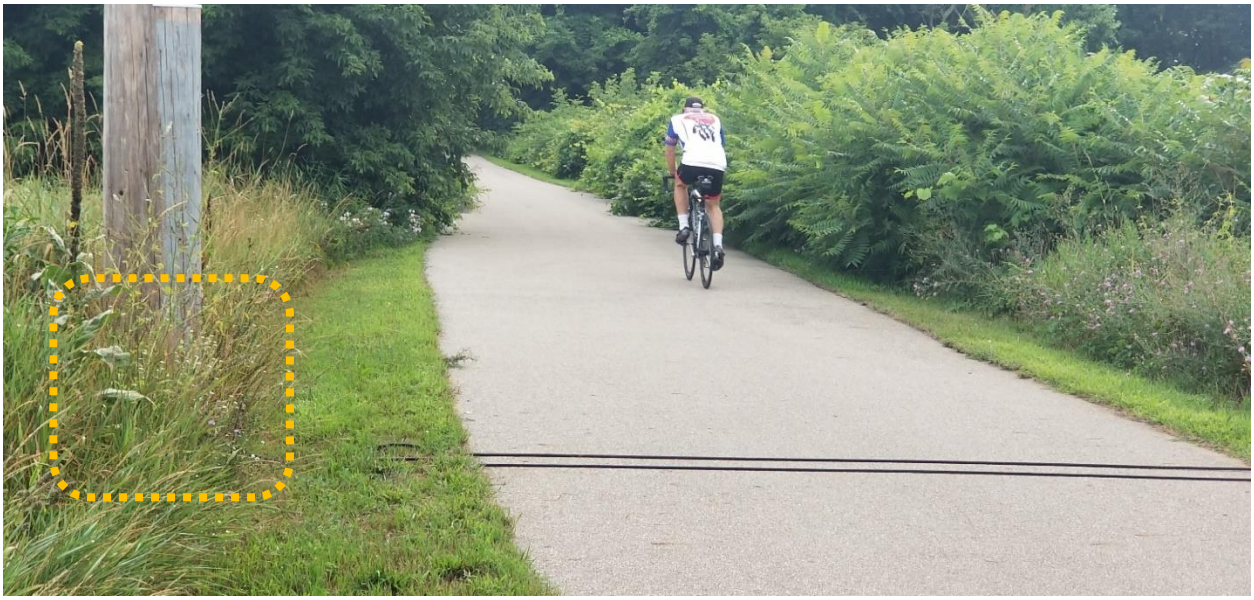


Figure 3.7: Bicycle tube counters on Oxford Trails, Grand Rapids

3.4 Video data processing

The COUNTPro Software was used to semi-automate the counting process. The software has an interface (see Figure 3.8) that allows a person who is counting to adjust the playback speed as desired, depending on the level of cyclists' activities. The data are automatically recorded by pressing the count pad once a cyclist is spotted in the video

recording. The use of CountPro software reduced the preprocessing time by almost 80 percent.

Figure 3.8: An interface of COUNTPro software and the COUNTpad used for counting cyclists

3.5 Strava Data

The Strava data was purchased covering a period of one year (February 2018 to January 2019). Strava offers a data package called Strava Metro which comprises of hourly or rollup cyclists' activities for each node (intersection) and edges (segment) of the roadway network. It further subdivides the cyclists' trips into commute and non-commute trips. The

time for each activity for a given hour or rollup (aggregated by time of the day, a week, a month, or a season) is also available. Other information in the Strava Metro package includes aggregated demographic information of Strava users by age and gender and origin-destination table aggregated within 350m hexagonal polygonal geometry. A comma delimited file of hourly Strava activities at all segments in the city of Ann Arbor and Grand Rapids was joined with the road edge shapefile using ArcGIS software. This enabled spatial merging with other data such as roadway bicycle facilities information, total cyclists count from video recordings, landuse data, weather data and census data.

Figure 3.9 and Figure 3.10 provide the distribution of Strava activities (monthly average) by different land use types for the city of Ann Arbor and Grand Rapids respectively. In Ann Arbor city, significant portion of Strava activities were found in the northern part. Strava users mostly used scenic routes such as West Huron River Drive going along the Huron River and trails passing through residential, forest, and rangeland areas such as the Pontiac trails. Roadway that were relatively close to the scenic routes and trails also had high number of Strava users. These nearby roadways acted as entry or exist to those areas or routes that attracted most of the Strava users. The Gallup Park Pathway border to border trail that starts from northwestern part to eastern part of Ann Arbor also attracted a significant number of cyclists who were using Strava app.

Similar pattern was observed in the city of Grand Rapids. Strava users were mostly attracted to trails that were passing through forest/wetland/rangeland areas or surrounded by water bodies such as Kent Trail and Oxford Trails in the western part area and White Pine Trail in Riverside Park in the Northern part area. Also, a relatively high number of Strava activities was observed in the central part of Grand Rapids which is characterized by commercial activities. The Monroe Avenue site which has a separate two-way bike lane (side path) also attracted significant number of cyclists who were using Strava app. The southern area of Grand Rapid had relatively low number of cyclists using Strava app with exception of Plaster Creek Trail which pass through Ken-O-Sha Park.

Overall, we can discern that majority of Strava users were found to ride in recreational and residential areas having trails or scenic routes close to water bodies. Roadway that were proximal to these locations also experienced a high number of cyclists using Strava app. These roadways mostly provided access to cyclists who were entering

or exiting from these trails and scenic routes. Furthermore, the data showed that roadways with dedicated bicycle facility for example, Monroe Avenue and Walker Avenue in city of Grand Rapids, attracted a significant number of Strava cyclists.

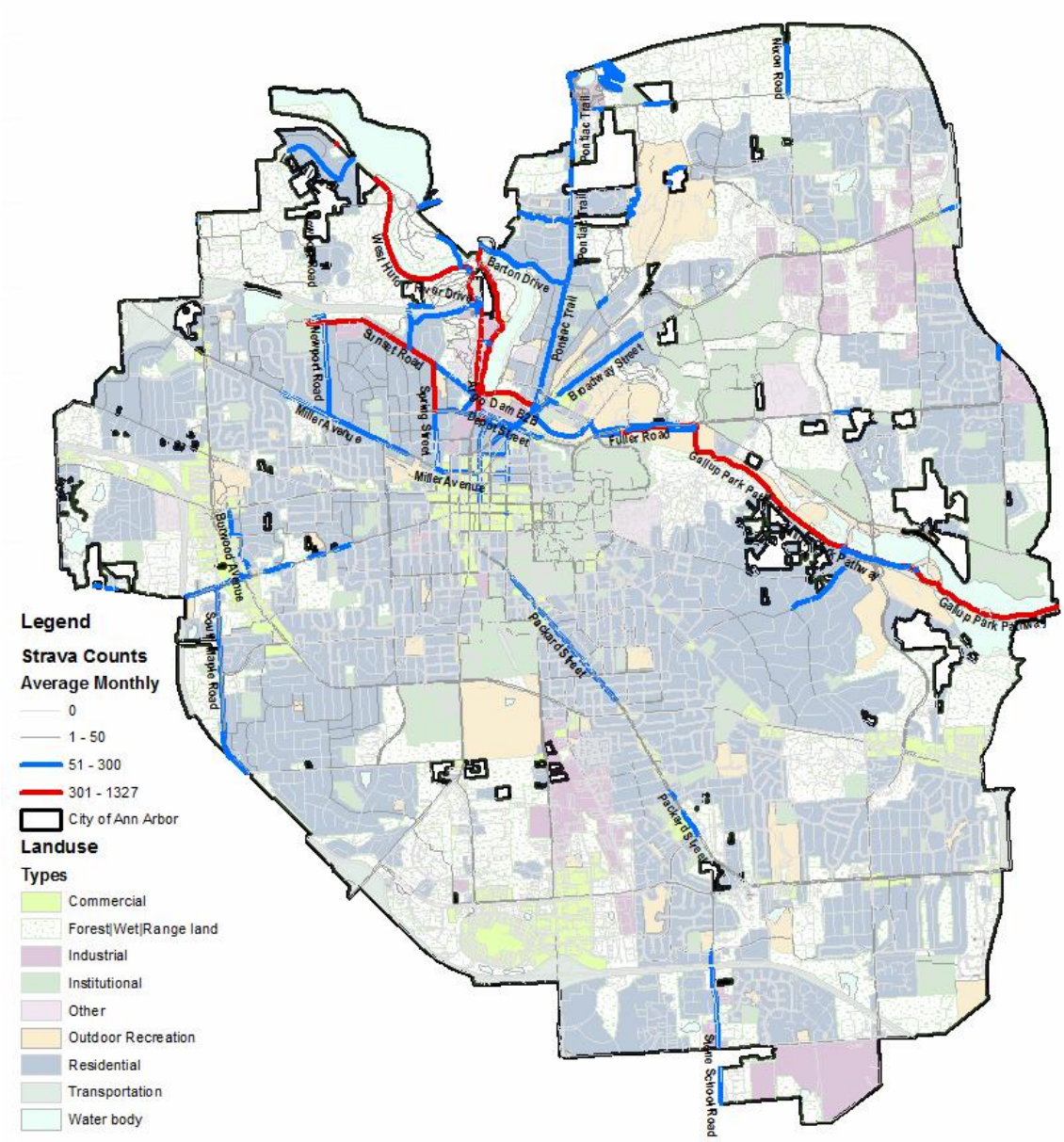


Figure 3.9: Distribution of Strava activities relative roadway and land use type in Ann Arbor

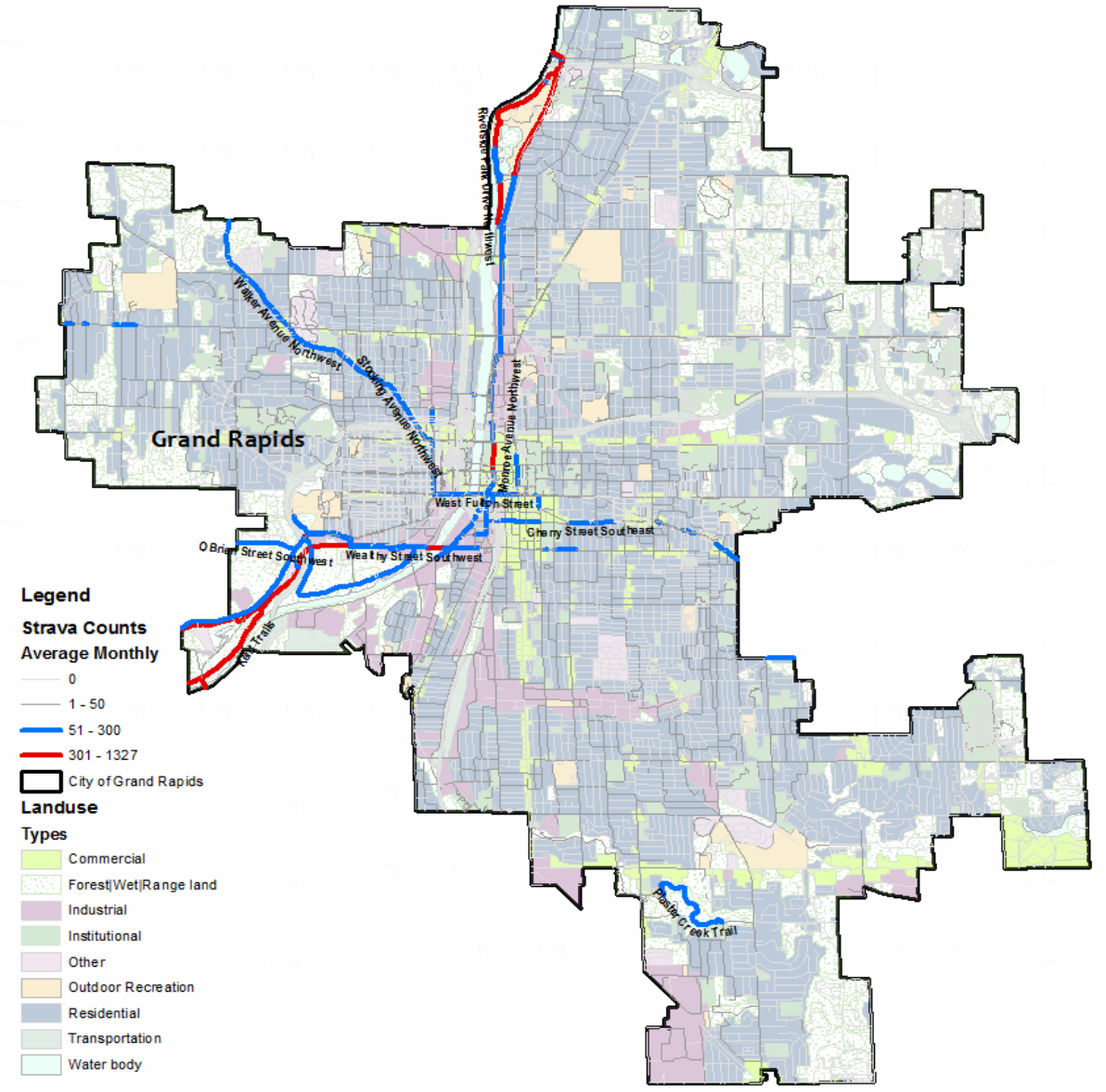


Figure 3.10: Distribution of Strava activities relative roadway and land use type in Grand Rapids

3.6 Weather Data

The historical hourly weather data for the city of Ann Arbor and Grand Rapids were obtained from an online weather repository, www.wunderground.com. The snapshot of the online weather information is shown in Figure 3.11. Each weather station provides an hourly weather information such as precipitation, temperature, relative humidity, and wind speed. For each site, the research team selected weather stations that was closest to the

video data collection site and the time were matched to ensure an exact match is obtained between the cyclists' activities and the weather data.

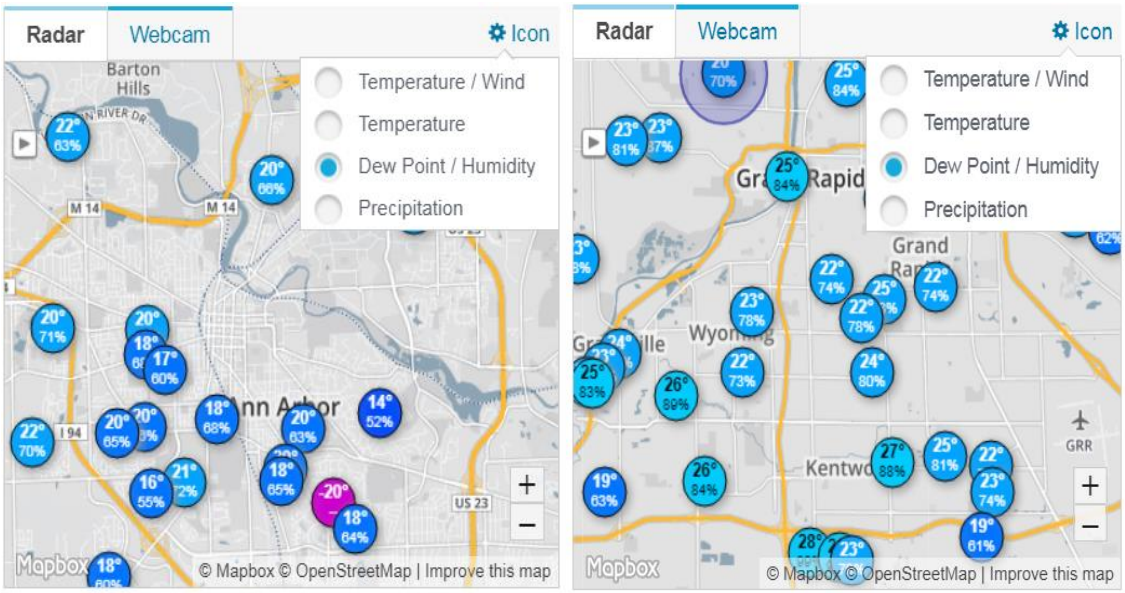


Figure 3.11: Weather Stations in Ann Arbor and Grand Rapids at hourly Level

3.7 Survey data

The survey of cyclists was conducted concurrently with video data collection of cyclist activities. It was designed to acquire the information which were difficult to discern using the video data. Such information includes (1) demographics and cycling behavior of the cyclists (2) characteristics of the trip(s) made by the cyclists, and (3) the proportion of cyclists using fitness and health apps to track their cycling activities. The survey had a total number of 10 questions which took a maximum of 2 minutes to complete. The survey was conducted at nearby bicycle racks and trail rest areas. Figure 3.12 shows some of the locations were the survey was administered including Kent Trail and Riverside trails in Grand Rapids. Figure 3.13 shows the number of cyclists that were surveyed for each location in Ann Arbor and Grand Rapids. The survey was conducted in the period between May to August 2018 – the same time when video data were also collected in the field. The survey questionnaire is available in Appendix 8.5.

Figure 3.12: Survey of cyclists on the bicycle racks or trail rest areas

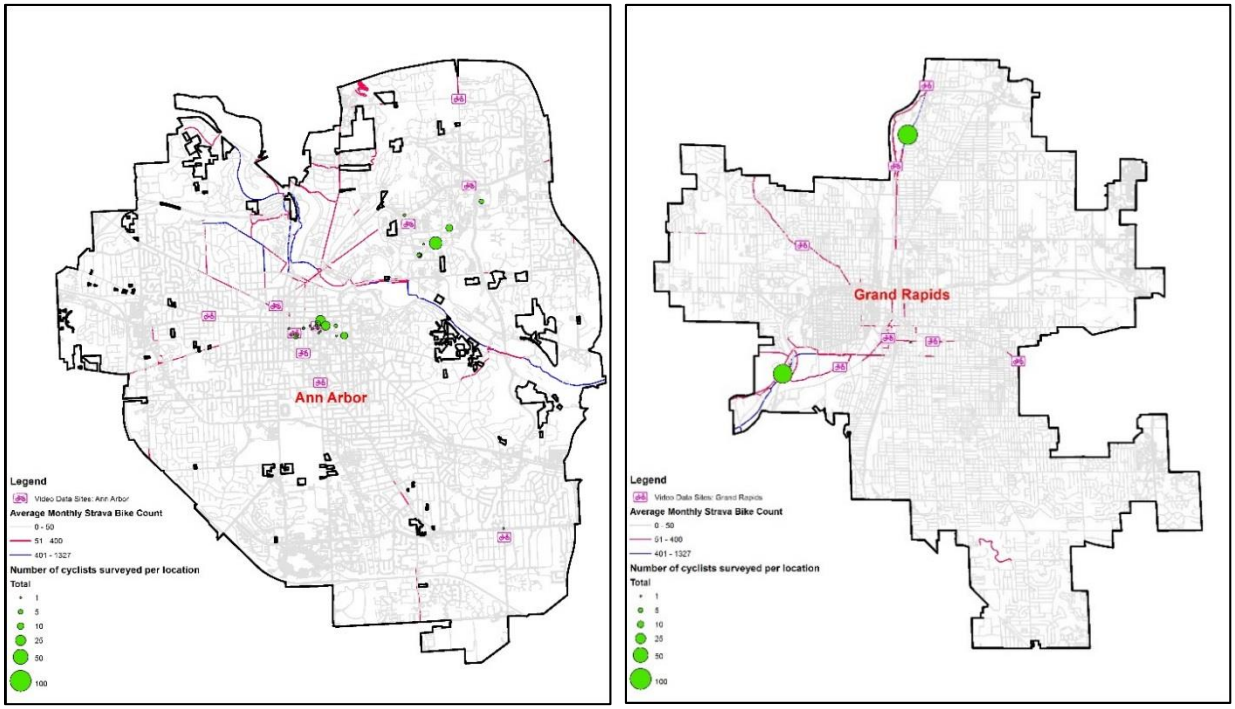


Figure 3.13: Survey locations with respect to video data collection of cyclists' activities

3.8 Descriptive statistics of the data

The analysis of descriptive statistics was carried out to understand basic relationships between different data types. Correlation plots and coefficient of determination were used to assess the association between total cyclists counts and Strava cyclists counts based on landuse, time of the day, roadway facility and weather. Also, Strava penetration rates which is the ratio of Strava count to total cyclists count were calculated based on landuse, bicycle facility and time of the day.

3.8.1 Comparison of Strava counts and total counts

Figure 3.14 shows the hourly distribution of total cyclist counts and Strava cyclist counts aggregated for all sites. The trend of total cyclist counts had a sharp increase starting from 6:00 am up to noon with no significant change afterwards until around 7:00 pm. The peak flow of total number of cyclists was observed between 6-7 pm. Thereafter, there was a continuous decrease in the number of cyclists with almost no cyclist recorded after midnight. Same pattern was observed for the segment of cyclists that were using Strava app with a significant positive correlation coefficient (R= 0.7244, p=0.0007) between 5:00 am to 11:00 pm. In particular, the correlation was observed to be stronger between 5:00 am to noon time (R= 0.9485, p=0.0003) and from 6:00 pm to 11:00pm (R= 0.9148, p= 0.0295).

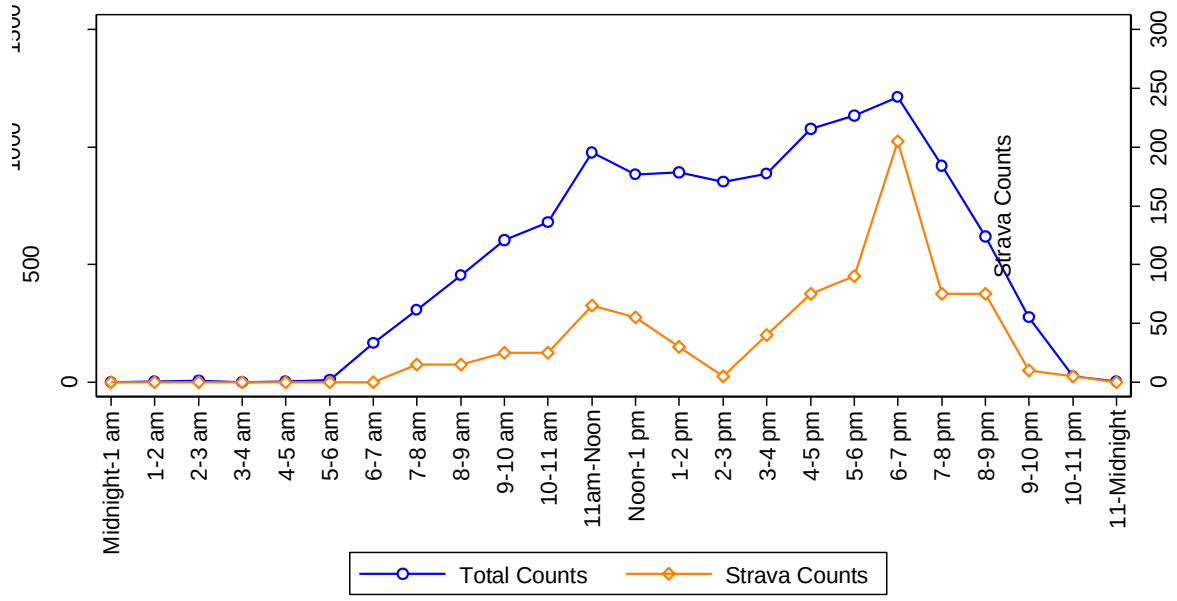


Figure 3.14: Hourly trend of Cyclist Activity: Total versus Strava Counts

3.8.2 Correlation of cyclist counts with weather data.

The hourly distributions of cyclists' counts were compared with hourly distribution of weather information particularly relative humidity expressed in percentage. The counts were aggregated in the relative humidity bins having a class interval of 5%. The trend of total cyclist counts and Strava counts were compared for each relative humidity bin. Figure 3.15 shows the frequency of hourly counts for each relative humidity bin represented by a class mark on an x-axis. The total number of cyclists aggregated over each bin is also plotted together with the frequency which is the number of hours spent in collecting cyclist counts for each bin of relative humidity. It can be observed that the number of hours per each bin didn't have a considerable variation at relative humidity between 50-100%. However, the total cyclists per each bin kept on decreasing, with maximum number of total cyclists observed at a relative humidity of 50%. The same trend was observed for cyclists who were using Strava app. The cyclist counts were normalized by frequency (number of hours) at each bin as shown in Figure 3.16. The spikes above average were observed for both total number of cyclists and Strava userss on a relative humidity range of 10-25% and 45-65%. Figure 3.17 shows the distribution of counts for each relative humidity bin separated by bicycle facility type. The high counts spikes at lower relative humidity (10-25%) were mostly on trails while the remaining observed spikes on moderate relative humidity (45-65%) were on bike lane and shared lane facilities. In summary, the results show that cyclists tend to avoid riding when the relative humidity is too high.

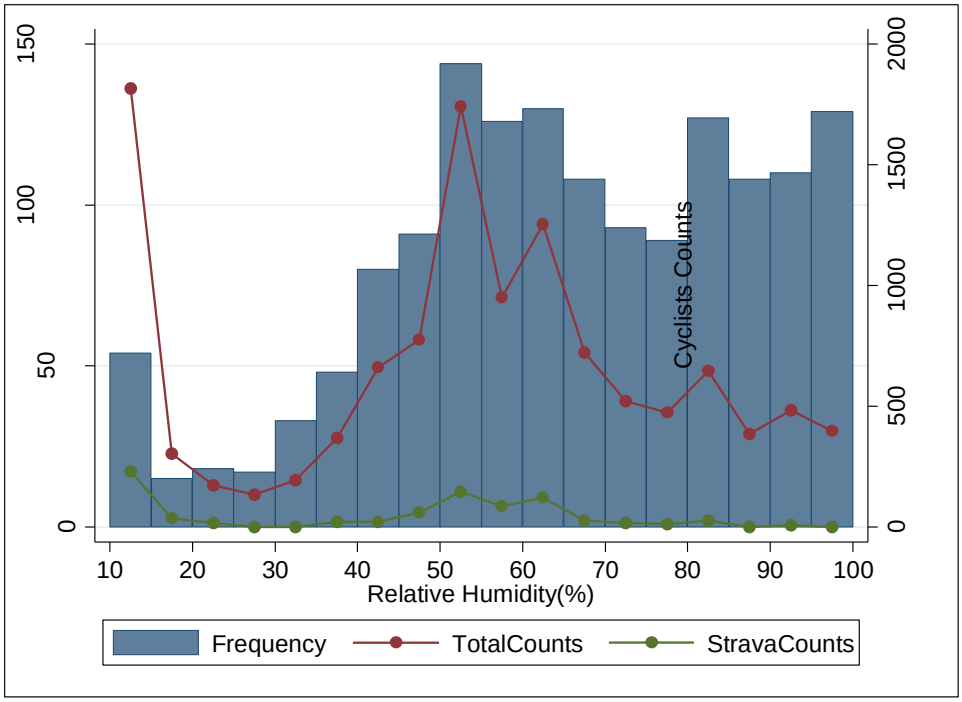


Figure 3.15: Distribution of counts with respect to relative humidity variations across sites

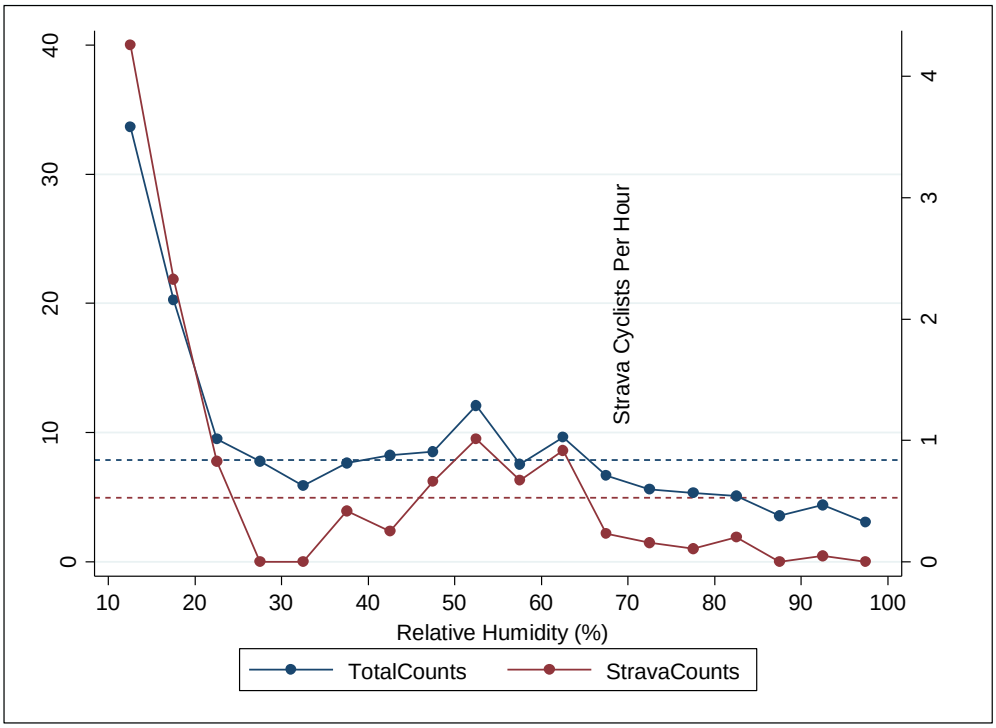


Figure 3.16: Normalized distribution of counts with respect to relative humidity variations across sites

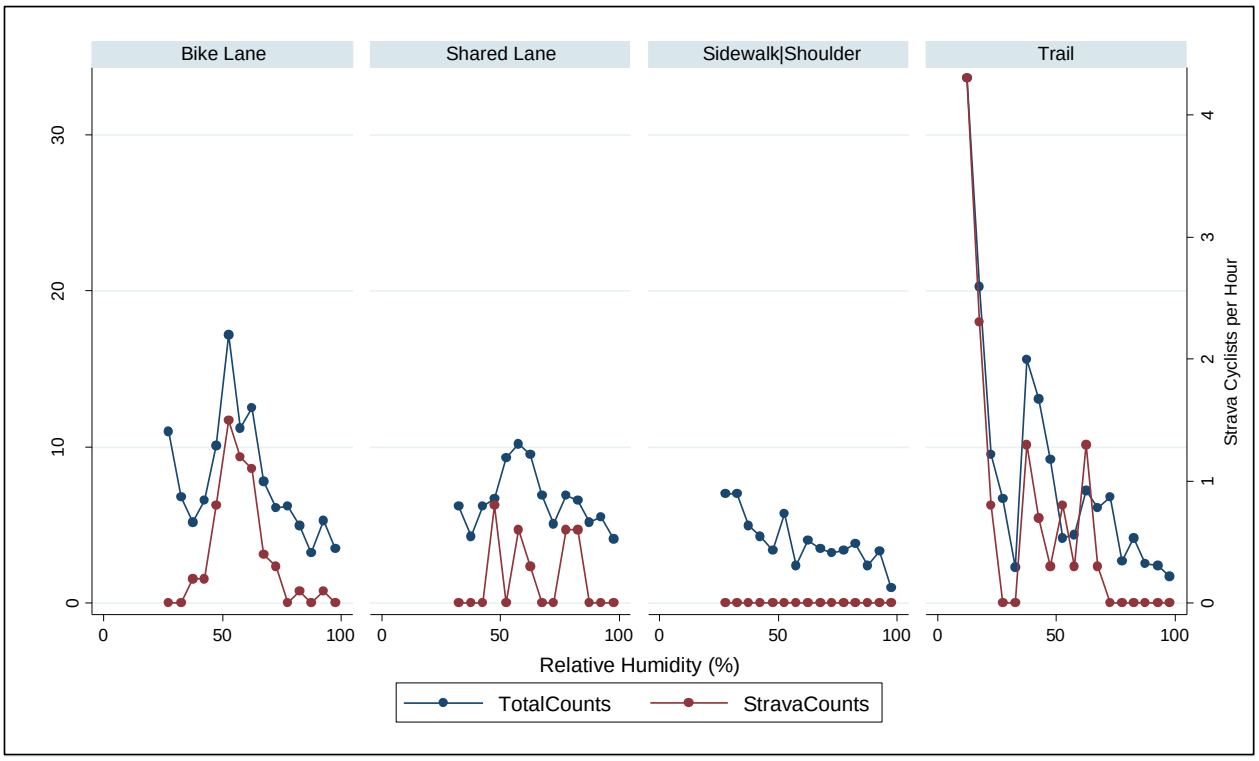


Figure 3.17: Distribution of cyclists counts with relative humidity by bicycle facility

3.8.3 Strava penetration rates

Strava penetration rates were computed as the percent of cyclists who were using Strava app in the total cycling population. The penetration rates were calculated for different categories as shown in Table 3.3. The Pearson correlation of coefficient (R) was computed for each attribute of a given category using hourly total cyclists' counts and hourly Strava counts.

The distribution of Strava penetration rates by hour of the day had a range of 0 to 10 percent. The minimum penetration rate was observed in Early AM hrs: 12am-5:59am while the maximum penetration rate was observed at PM hrs: 3pm-7:59pm. The coefficient of correlation by hour of the day ranged from 0.419-0.621. The Pearson correlation coefficient indicated a significant correlation between hourly total cyclist counts and Strava cyclist counts for each hour of the day.

Site location where there was no dedicated bicycle facility and cyclists had to use either sidewalk or shoulder didn't have any cyclists that were using the Strava app. The Strava penetration rates for shared lane was 4 percent followed by bike lane (6 percent)

and trails which had a maximum penetration rate of 9 percent. The correlation coefficient ranged from 0.209 (p=0.001) on shared lanes bicycle facility to 0.646 (p=0.000) on trails.

With respect to landuse, roadway segment passing through the commercial area had the lowest Strava penetration rate (3 percent) followed by residential areas (5 percent), institutional areas (6 percent) and recreational areas (9 percent). The Pearson correlation coefficient range was 0.114 to 0.646.

The distributions of Strava penetration rates by bicycle facility and land use type concurs with the spatial analysis which was conducted in the previous section. Significant number of Strava users were observed to use trails and scenic routes which had dedicated bicycle facilities that passed through residential and institutional areas.

Table 3.3: Strava Penetration Rates

Category	Attributes	Total count	Strava Count	Pen (%)	Pearson's correlation	p-value
Hourly adjustment	Early AM hrs: 12am-5:59am	34	0	0%	-	-
	AM hrs: 6am-9:59am	1756	55	3%	0.419	0.000
	Mid-Day hrs: 10am-2:59pm	4932	200	4%	0.643	0.000
	PM hrs: 3pm-7:59pm	5584	535	10%	0.621	0.000
	Evening hrs: 8pm-11:59pm	1028	95	9%	0.463	0.000
Bike Facility	Sidewalk|Shoulder	551	0	0%	-	-
	Shared Lane	1835	65	4%	0.209	0.001
	Bike Lane	6415	390	6%	0.600	0.000
	Trail	4533	430	9%	0.646	0.000
Land use Type	Commercial	1557	45	3%	0.114	0.053
	Residential	3027	165	5%	0.435	0.000
	Institutional	4217	245	6%	0.653	0.000
	Recreational	4533	430	9%	0.646	0.000
Grand Total		13334	885	7%	0.596	0.000

4 Survey of Cyclists in Ann Arbor and Grand Rapids

The survey was conducted to understand the similarities and differences between cyclists who were using fitness apps to monitor their cycling activities and cyclists who were not using any fitness tracking apps. The similarities and differences between these two cohorts of cyclists were analyzed based on cyclists' demographic characteristics, cycling experiences, cyclists' trip characteristics and cycling behavior. The statistical test of homogeneity was used to discern if the observed differences or similarities were significant. Furthermore, the logistic regression was used to understand how cyclists' demographic characteristics, trip characteristic, and cycling behaviors impact the likelihood of using a fitness tracking app(s).

4.1 The use of fitness trackers among cyclists

Cyclists were asked to report if they use Strava app or any other fitness app to track their cycling activities. Figure 4.1 displays the distribution of cyclists by fitness tracking app usage. A total of 321 cyclists were surveyed in the city of Ann Arbor and the city of Grand Rapids. The majority of cyclists who participated in the survey reported that they do not use any fitness tracking apps (66 percent). About 16 percent of cyclists reported to use Strava app to monitor their cycling activities and 18 percent reported to use other fitness tracker app(s) that are currently on the market.

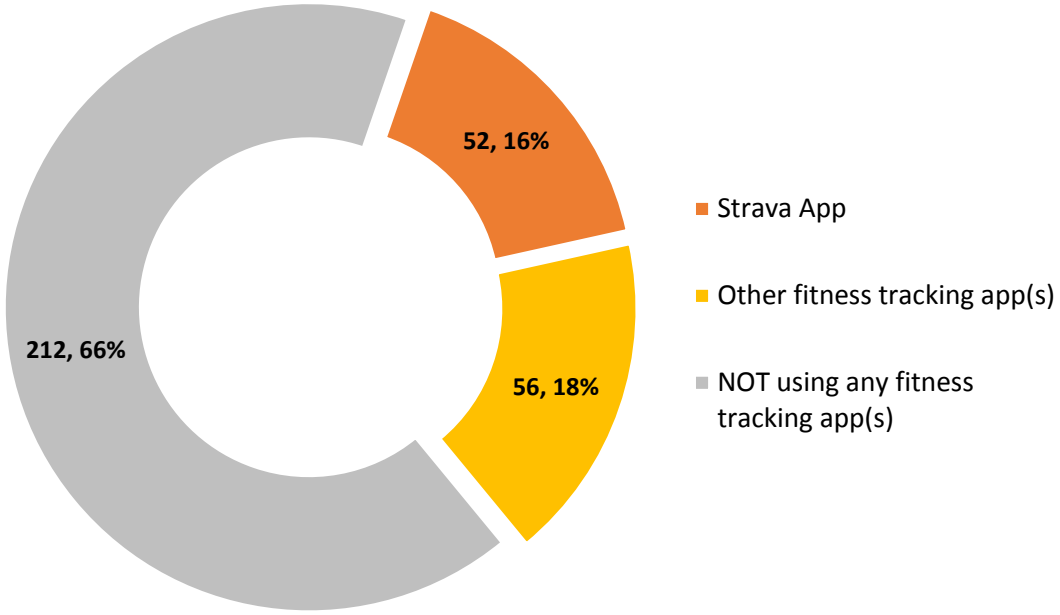


Figure 4.1: Reported tracking app(s) usage by cyclists

4.2 Demographics and cycling experience of cyclists

Figure 4.2 shows the distribution of cyclists by fitness tracker app usage for each cyclists' age group. The fitness tracker app usage was more among cyclists aged 35 years to 44 years and older cyclists aged 65 years and above. The high proportion that was observed among older cyclists using fitness tracker apps may suggest that older cyclists are keen on assessing potential benefits of cycling and not just riding for pleasure.

Age of a cyclist can be a factor in usage of fitness tracking apps. In this survey, cyclists who reported to use Strava app had higher percentage compared to those who are using other fitness app among the age groups of 25-34 years and 35-44 years. Other fitness apps were likely to be used more by older cyclists i.e., age greater than 55. Overall, the observed difference in fitness tracker app usage across age groups was significant at 95 percent confidence level (χ^2=34.22, p=0.00).

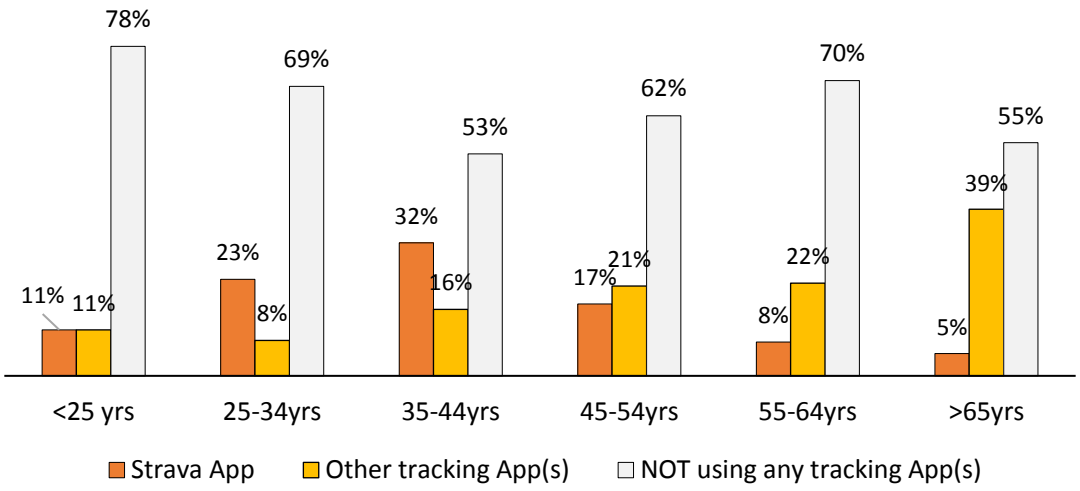


Figure 4.2: Cyclists' app usage across age groups

The distribution of fitness tracker app usage was slightly higher among female cyclists (37%) compared to male cyclists (24%) as shown in Figure 4.3. The observed differences in app usage between male and female was significant at 90% confidence level (χ^2=34.22, p=0.06). The proportion of Strava users was slightly higher than other tracking app(s) users among male cyclists while female cyclists reported to use other fitness tracker app(s) more than Strava App. This suggests that Strava can capture male cycling population, data from other fitness tracker app(s) can be used along with the Strava data to capture the female cycling population.

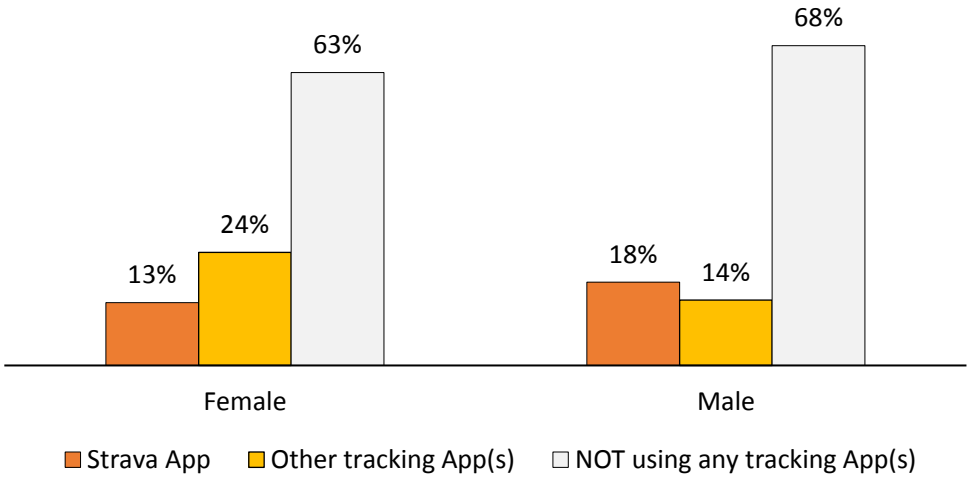


Figure 4.3: Tracking app utilization among males and females

Figure 4.4 indicates the increase in proportion of cyclists using tracking and fitness apps for experienced cyclists (intermediate skills and expert). Fitness tracking apps were more likely to be used by intermediate and expert cyclists compared to beginners. The percentage of other tracking apps utilization among beginners was slightly higher (17%) compared to Strava utilization (11 percent), but no difference in Strava app and other tracking apps utilization was observed for intermediate and expert cyclists. The observed difference in distribution between cyclists who were using fitness tracker apps and cyclists who were not using any fitness tracker app was significant at 95 percent confidence level (χ^2=11.15, p=0.03).

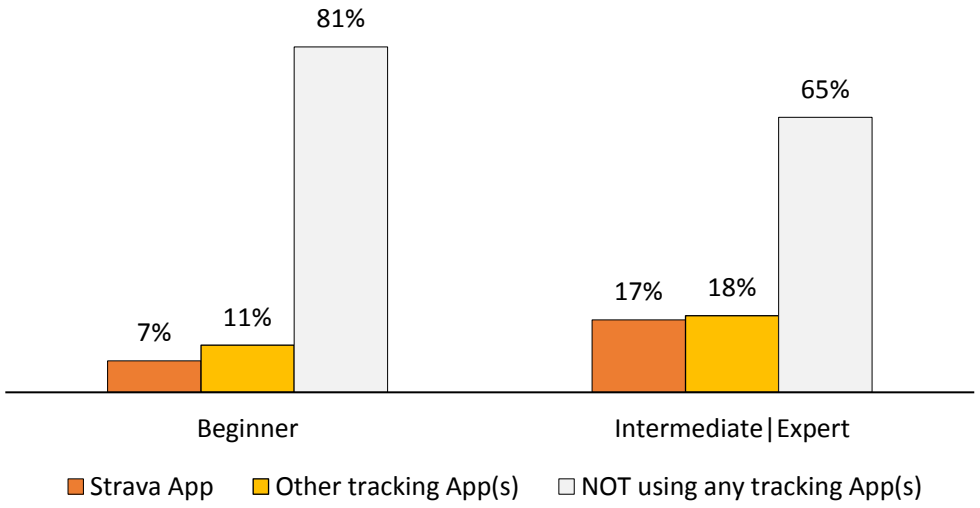


Figure 4.4: Tracking app utilization by cycling experience

4.3 Cycling characteristics by trip purpose

The trip purpose is one of the essential elements that is used in planning and designing of bicycle facilities. In most cases, it signifies the pattern and distribution of cycling activities on a given roadway network. The trip purpose was explored based on fitness track app utilization (Figure 4.5), cycling frequency (Figure 4.6) and bicycle facility usage (Figure 4.7).

Cyclists were more likely to use fitness tracker apps when making recreational trips (45%) compared to when they were making commute trips (22%). Strava app was likely to be utilized more among commuters (13%) compared to other tracking apps (9%). For

recreational trips, 25% of cyclists reported to use other fitness tracker apps compared to 20% of cyclists who were using Strava app. The observed difference of app utilization by trip type was significant at 95 percent confidence level (χ^2=20.05, p=0.00).

Moreover, cyclists who reported to be on recreational trips were more likely to use off-road facilities such as trails (89%) while commuters were more likely to use other bicycle facilities such as dedicated bike lanes and shared lanes (87%). For planning purposes, the results suggest that, in the absence of trip purpose information, the bicycle facility type information can be a potential surrogate for trip purpose.

The frequency of biking was strongly associated with the trip purpose (χ2=20.05, p=0.00). Cyclists who were commuting were likely ride always (87 percent). Cyclists who ride for recreation were likely ride several times in a month (86 percent) to few times a week (67 percent). The results offer practical significance in planning of bicycle facilities. For planners who resort to using crowdsourced data from fitness tracking app, necessary adjustment factors are needed in case the data from fitness tracking app is determined to be biased toward recreational trips. The adjustment factor can be found using a survey to get the ratios of frequency of use of a given bicycle facility by trip purpose. Alternatively, other modeling procedures, such as one proposed in this study (Chapter 5) can be used to offset the bias by incorporating other confounding factors.

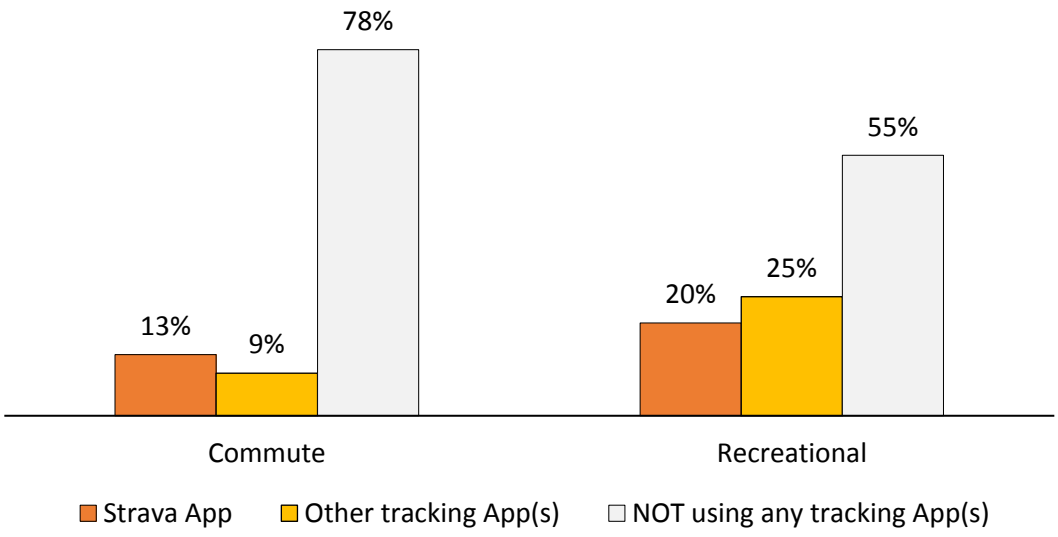


Figure 4.5: Distribution of tracking fitness app usage by trip purpose

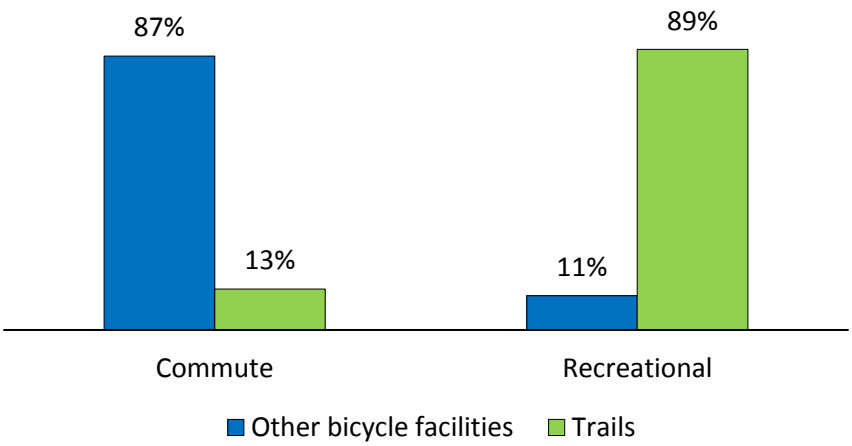


Figure 4.6: Bicycle facility usage by trip purpose

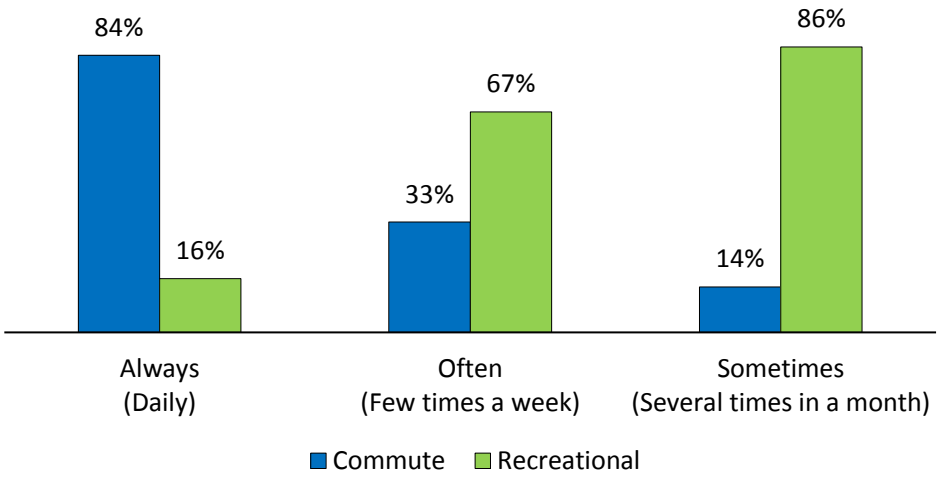


Figure 4.7: Frequency of biking by trip purpose

4.4 Factors influencing the choice of fitness tracking app utilization among cyclists

The logistic regression model was estimated to understand how each of the demographic characteristics, trip characteristics and cycling behavior affect the choice of using a fitness tracker app(s). Table 4.1 displays the summary of the dependent variables and the covariates that were used in the model. Age and gender represented the demographic characteristics of the cyclists of the participants while biking frequency, trip purpose and sidewalk use represent the riding behavior of cyclists.

Table 4.1: Variable descriptions

Variable	Category	Observation	Percent Freq.
Fitness Tracking app(s) Usability	Yes	56	34%
	No	212	66%
Strava Usability	Yes	52	16%
	No	268	84%
Age	<16	3	1%
	16-24 years	61	19%
	25-34 years	83	26%
	35-44 years	38	12%
	45-54 years	47	15%
	55-64 years	50	16%
	>65 years	38	12%
Gender	Female	112	35%
	Male	208	65%
Biking Frequency	Always	117	37%
	Often	144	45%
	Sometimes	59	18%
Biking Experience	Beginner	27	8%
	Intermediate	219	68%
	Expert	74	23%
Trip Purpose	Commute	154	48%
	Recreational	166	52%
Riding Behavior (Sidewalk Use)	Always	65	20%
	Only on Busy Roads	186	58%
	Never	54	17%
	No Preference	15	5%

A logistic regression was estimated for two scenarios. The first scenario dealt with cyclists who reported to use Strava app for tracking their fitness while the second scenario dealt with cyclists who reported to use other fitness tracking app(s) that available in the market. The analysis aimed at understanding the influence of each covariate for each of the scenario. The logistic regression allows us to measure the influence that a certain covariate on cyclists' choice of either to use or not using fitness tracking app while controlling for other confounding covariates. The odds ratios (OD) which is the exponentiated coefficients of the covariates has been used widely in the literature for assessing the impacts of covariates on a binary outcome. Table 4.2 displays the logistic

model results for the two scenarios. The covariates that were retained in the model were significant at least at 90% confidence level in either of the scenario.

Cyclist age information was obtained from the survey in a form of age group. The age cohorts that were significant in the model for the case of Strava app utilization were age groups 25-34 years, 35- 44 years and greater than 65 years. The reference group for age variable was cyclists whose age were below 25 years. The odds of using a Strava app were likely to increase by 6.7% for cyclists' age 25-34 years and 52% for cyclists age 35-44 years. The cyclist age group 35-44 years may likely represent a cyclist segment that is mostly captured in Strava cycling data. The older cyclists age 65 years had lower odds of using Strava app compared to the reference group of cyclists (age less than 25 years). Conversely, older cyclists had higher odds of utilizing other fitness app(s) for monitoring their fitness (OR=1.94). Further, cyclists whose ages were between 25 to 34 years had significant lower odds of using other fitness tracker apps compared to young cyclists (age less than 25). Conclusively, Strava app was found to significantly capture middle age group (25-44yrs) while other fitness apps were likely to be utilized by young cyclists age less than 25 and older cyclist age 65 years and above.

The odds of using Strava app for male cyclists increased by a factor of 2.204 compared to females. Conversely, the female cyclists had higher odds (OR=1.813) compared to males for cyclists who reported to use other fitness app(s). The results suggest that Strava app is likely to be utilized more by male cyclists while other tracking fitness apps are likely to be utilized more by female cyclists after controlling for other confounding factors.

Trip purpose was found to significantly affect the use of fitness tracking apps among cyclists. This covariate was significant among Strava users and other fitness tacker apps users. In both cases, cyclists who were making recreational trips had higher odds of using fitness tracking apps compared to cyclists who were making commute trips. This emphasizes the need to incorporate other data sources to account for underrepresentation of commute trips when estimating the total number of cyclists' trips as data from fitness tracking apps are likely to be biased towards recreational trips.

Cycling behavior in the context of sidewalk usage was a significant factor in determining whether a cyclist will use a Strava app. The cyclists were asked whether they

will use a sidewalk given the bicycle facilities such as bike lane or shared lane were available. The riding behavior of cyclists was a significant determinant of Strava app usability but not for other fitness tracking app usability. Cyclists who reported to always use the sidewalk regardless of presence of bicycle facilities were treated as a reference group when calibrating the logistic regression model. The odds of cyclists using a Strava app increased by 3.5% for cyclists who reported to use sidewalk only when riding on a high-speed or busy road. Cyclists who were keen in utilizing the dedicated bicycle facility if available regardless of any road condition were the most likely group of cyclists to utilize Strava app (IR=1.854).

Other significant factors were cycling frequency and cycling experience which were significant determinants only for Strava app utilization. Cyclists who were riding always, possibly commuters, had lower odds of using Strava app compared to occasional riders. Based on cycling experience, the odds of using Strava apps were higher for cyclists who had intermediate to advanced cycling skills.

The survey results suggest that cycling activities data from Strava app and other fitness apps can be used jointly in estimation of bicycle exposure since they represent different cycling population by age and gender. Male cyclists are likely to have high odds of using Strava app while female cyclists have higher odds of using other fitness apps. Further, Strava is more likely to be used by middle-age group of cyclists while other fitness apps are likely to be utilized by young and older cyclists. Therefore, combining different crowdsourced data may enable a more accurate estimation of total cycling activities covering a greater demographic diversity of cyclists.

Table 4.2: Results of Logistic Regression

Covariates	Strava App				Other fitness tracking app(s)			
	Coef.	OR	z	P>z	Coef.	OR	z	P>z
Cyclist's age (ref: <25 years)								
25-34 years	1.067	2.907	2.670	0.008*	-0.820	0.440	-1.750	0.080⁺
35-44 years	1.520	4.570	3.270	0.001*	-0.211	0.810	-0.410	0.683
>65 years	-1.569	0.208	-1.950	0.051⁺	0.662	1.940	1.530	0.126
Gender								
Male (ref: female)	0.790	2.204	2.070	0.038*				
Female (ref: Male)					0.595	1.813	1.820	0.069⁺
Trip purpose (ref: Commute)								
Recreational	0.873	2.395	2.390	0.017*	0.691	1.996	1.840	0.066⁺
Sidewalk use behavior (ref: Always)								
Never using a side walk	1.854	6.385	3.140	0.002*	-0.116	0.890	-0.240	0.813
Only on high-speed road	1.035	2.815	1.940	0.053⁺	-0.539	0.583	-1.490	0.137
Cycling frequency								
Often	0.591	1.805	1.690	0.092⁺	0.309	1.362	0.930	0.352
Cycling experience								
Intermediate	0.984	2.675	1.230	0.220	0.673	1.960	1.020	0.308
Expert	1.407	4.085	1.660	0.097⁺	-0.389	0.678	-0.500	0.620
Constant	-5.472	0.004	-5.210	0.000*	-2.421	0.089	-3.350	0.001*

*Note *significant at 95% CL, ⁺significant at 90% CL*

5 Integrating Crowdsourced Data in Estimation of Bicycle Exposure

Crowdsourced data from fitness tracker apps have been increasingly used in urban planning of non-motorized facilities as it has shown to be relatively fast, convenient and inexpensive means of acquiring public inputs. The market for fitness tracking apps is expected to continue growing in the coming years as more people are increasingly becoming interested in monitoring their health when engaged in physical activities such as walking and bicycling. This provides opportunities for leveraging such data to understand the travel patterns of non-motorized traffic at various levels of spatial granularity. This chapter address ways in which crowdsourced data can be integrated with other available data sources to improve the estimation of bicycle exposure. The chapter expounds on the specific modeling approaches that were used to estimate bicycle exposure at hourly level using crowdsourced cyclists' activities from Strava Metro as one of the exogenous variables. The inherent setback of Strava data being biased toward a specific segment of cyclists is circumvented by incorporating other covariates in the model such as census data, bicycle facility, hourly adjustment and land use data. Specific analyses covered in this section include using probabilistic and machine learning-based approaches to estimate hourly bicycle volume and a simulation study of Strava penetration rates. Model performance based on base and simulated Strava penetration rates and the practical implications of the results are discussed.

5.1 Distribution of total bicycle counts

Figure 5.1 displays the hourly distribution of total cyclist counts for all the sites that were used in the analysis. A total of 1,520 hours of cyclist counts were collected from all the 19 sites. About 13 percent of the total hours had zero counts. The percentage of zero cyclists count was not overrepresented in our sample to warrant the use of zero-inflated count models. Majority of hourly cyclists count were between 0 to 10 cyclists as shown in the cumulative frequency graph (Figure 5.2).

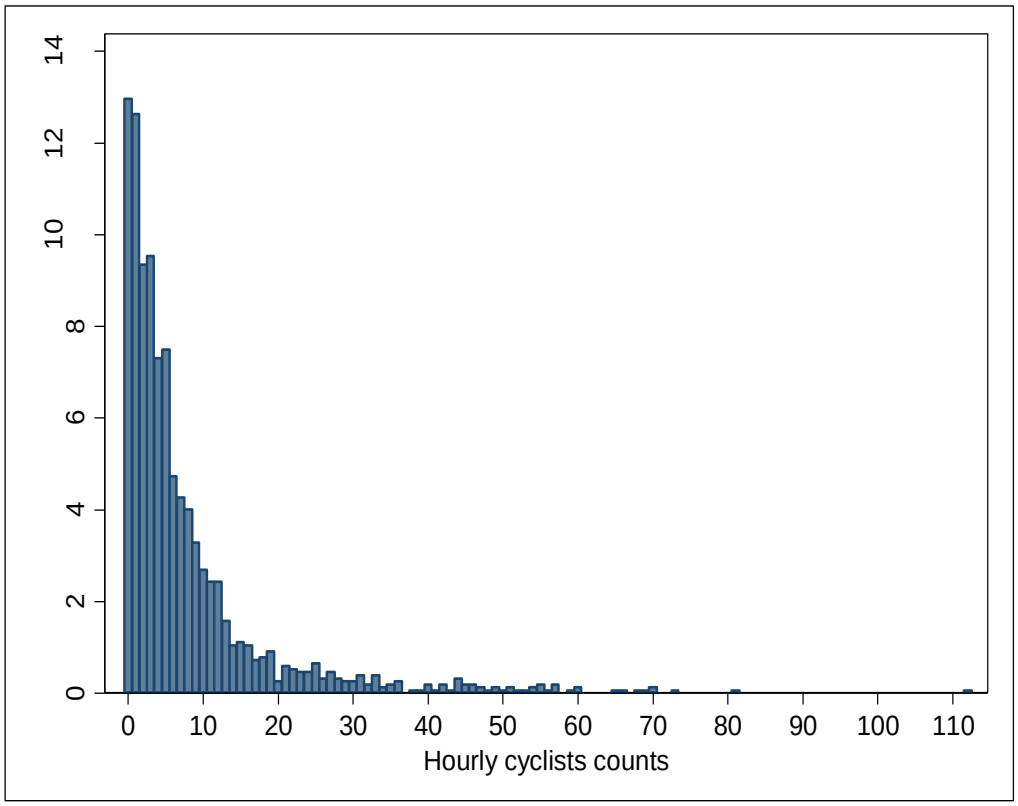


Figure 5.1: Histogram of hourly distribution of cyclist counts

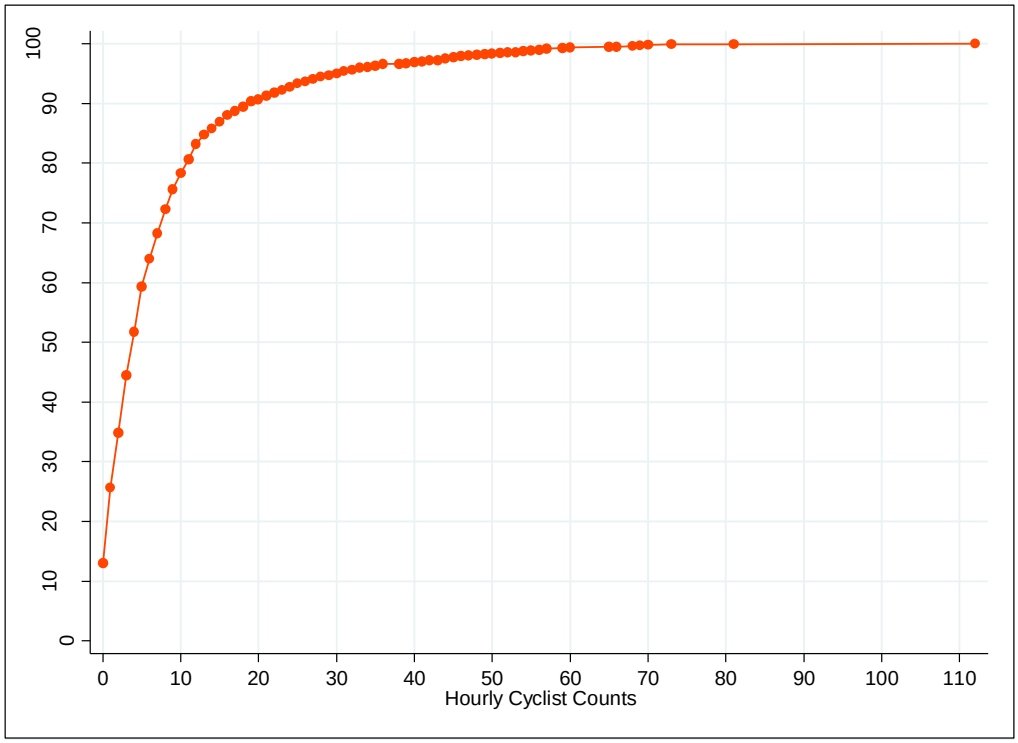


Figure 5.2: Cumulative hourly distribution of cyclist count

5.2 Correlation of covariates

Before going into the model calibration, it was important to understand how the variables correlate to one another. Variables that have a good correlation with the dependent variable tend to fit the model well during calibration. Figure 5.3 shows the correlation plots for all possible pairs of variables that were used in model calibration. The value in each box is the Pearson correlation coefficient for a given pair of variables. A column of interest is the one that compares the total cyclists count (dependent variable) with the covariates. Strava counts had the highest correlation coefficient of 0.6 followed by humidity and proportion of males in a given census block where the roadway segment is located.

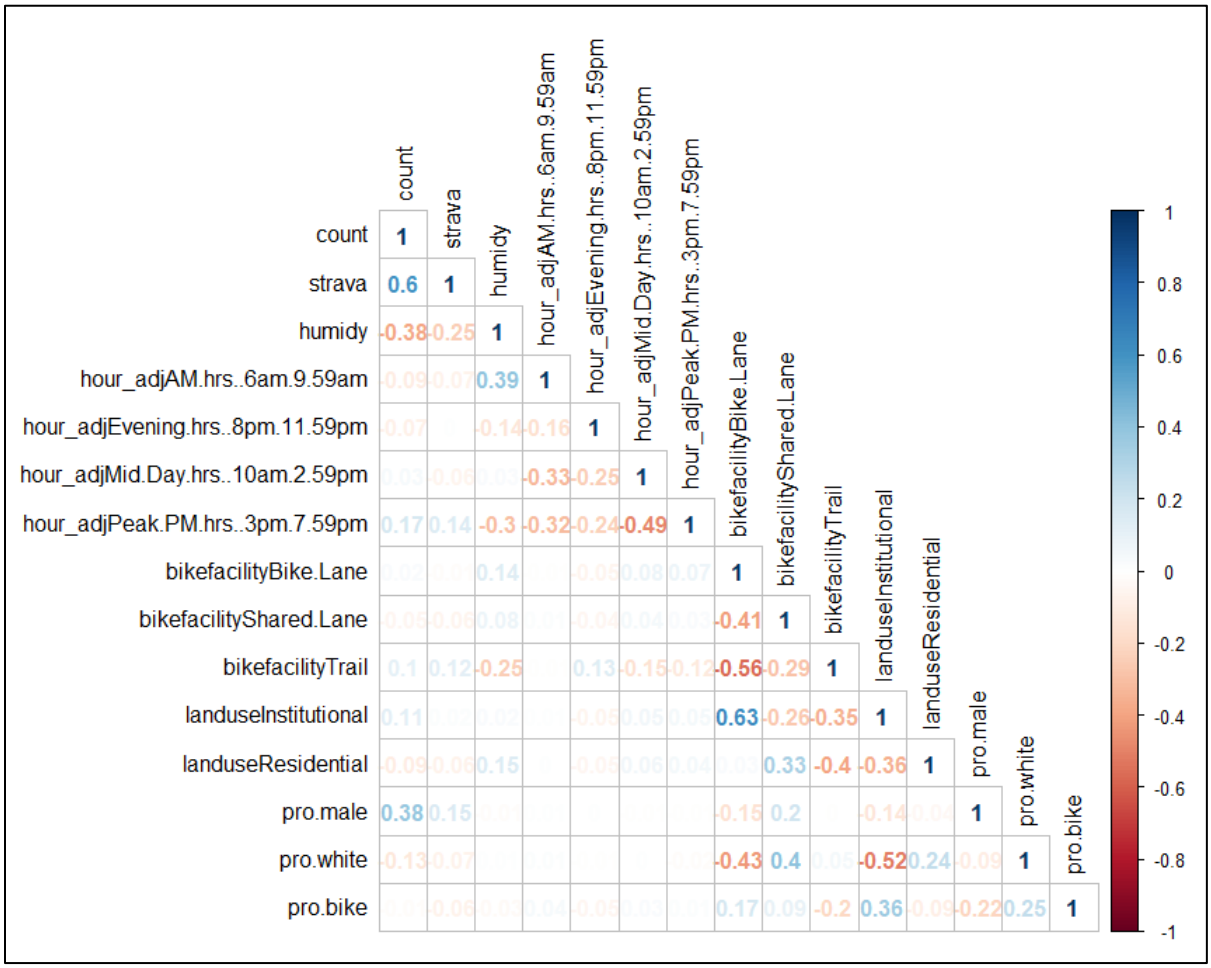


Figure 5.3: Correlation plots of variables used in the model

5.3 Modeling approaches

In calibrating the model, different modeling approaches were used in order to determine the model with the best fit. The use of different model approaches allowed us to investigate the consistency of independent variables, particularly Strava counts, in predicting total cyclist counts after controlling for other confounding factors. In this study we deployed a probabilistic model namely Negative Binomial regression. Furthermore, five machine-based learning models were used namely; Random Forest, Bagged regression tree, K-Nearest Neighbor, Support Vector Machines and Neural Network. A brief description of each machine learning technique is provided hereafter.

5.3.1 Negative Binomial

Negative binomial regression which handle cases where mean and variance of the count data are not equal can be derived from the Poisson model. The probability of segment i having n number of cyclists a given time period can be written as (Hilbe, 2011):

$$P(y_i) = \frac{EXP(-\lambda) \cdot \lambda_i^{y_i}}{y_i!} \tag{5.1}$$

λ_i is the Poisson parameter for segment i. In this study it can be defined as number of cyclists in a given hour. Generalizing Poisson model by introducing unobserved effect ε_i such that the expected Poisson parameter becomes

$$\lambda_i = EXP(\beta X_i + \varepsilon_i) \tag{5.2}$$

With $\lambda_i = EXP(\varepsilon_i)$ is known as gamma distributed error term with mean of one and variance of α^2, β = the vector of coefficient of predictor variables and X_i = the predictor variable i. Upon modifying mean-variance relationship for expected variance-mean relationship of hourly cyclist y_i becomes:

$$Var[y_i] = E(y_i) \cdot [1 + \alpha E(y_i)] = E[y_i] + \alpha E(y_i)^2 \tag{5.3}$$

If α is significantly different from zero then the cyclists counts in a given hour are said to be overdispersed for positive α values and underdispersed for negative α values. For overdispersion case, the resulting Negative binomial probability distribution can be written as:

$$P(y_i) = \frac{\Gamma\left(\left(1/\alpha\right) + y_i\right)}{\Gamma\left(1/\alpha\right) y_i!} \left(\frac{1/\alpha}{\left(1/\alpha\right) + \lambda_i}\right)^{1/\alpha} \left(\frac{\lambda_i}{\left(1/\alpha\right) + \lambda_i}\right)^{y_i} \quad (5.4)$$

Whereby, $\Gamma(x)$ = A value of the gamma function, α = Overdispersion parameter and y_i = Number of hourly cyclist counts in a given road segment *i*

5.3.2 Random Forest

The RF-model is one of the most popular and accurate predictive machine learning techniques that have been used across various disciplines for classification and regression purposes (Hastie et al., 2009). The RF is an ensemble classifier that builds the decision tree from the bootstrapped training sample using a subset of predictors selected from the total predictors (Pal, 2017). Unlike other decision tree methods, a random sampling of predictor variables de-correlate the bootstrapped tree, thus increasing the overall performance of the RF model upon aggregation of all the trees. The ability of RF model to simultaneously de-correlate the trees and apply variable selection during the calibration process reduces the chances of model overfitting, and less variances when used in a different dataset (Gareth et al., 2013). The optimization problem of random forest involved finding the optimal selection tuple at each node by applying the impurity measure that is proportional to the heterogeneity of the node. A common node impurity measure is Gini impurity Index. If there are p classes the category set can be defined as (Bonaccorso, 2018):

$$Y = \{y_1, y_2, \ldots\ldots\ldots y_M\} \; where \; y_i \in [1, p] \quad (5.5)$$

The Gini Impurity Index which has to be minimized is given as

$$I_{Gini}\ (X_k) = \sum_{j=1}^{p} p(j|k)(1 - p(j|k)) \quad (5.6)$$

Whereby, *p(j|k)* represents the proportion of training observations in the jth region that are from the kth class. Gini index takes on a small value if all of the p(j|k) are close to zero or one

5.3.3 Artificial Neural Network

Artificial Neural Network (ANN) or sometimes referred as the connectionist systems is the framework that allows different machine learning algorithm to work together in solving complex tasks. It has recently become one of the most powerful computing algorithms for solving complex tasks in various disciplines such as object recognition and speed processing due to its ability of producing accurate results (Chien, 2019). The ANN is the collection of nodes commonly referred to as artificial neutrons. The input structure has the number of nodes equals to the dimensions of the input while the output has two nodes if it is a classification problem or one node if it is a regression problem. The ANN can have a layer of nodes that are neither input nor output that are commonly known as hidden layers. The complexity of the network increases with the increase in number of hidden layers. Figure 5.4 shows an example of ANN structure that was tested using our dataset. Each link that connects node *i* and node *j* is assigned the weights W_{ij} based on the chosen learning rule A neuron with label *j* will receive an input from a predecessor neutron *i* consisting of the threshold/bias *(b),* and activation energy, $Netj$ and activation functions, f_a. The cost function is used to iteratively change the initial tuning values of weight, bias and activation energy while minimizing the mismatch between the target output and the ANN output. The net activation energy can be computed as(Duda et al., 2001)

$$Netj = \sum_{i=1}^{d} x_i w_{ji} + w_{jo} \quad (5.7)$$

Whereby, w_{ji} is the vector weight and w_{jo} is the bias.

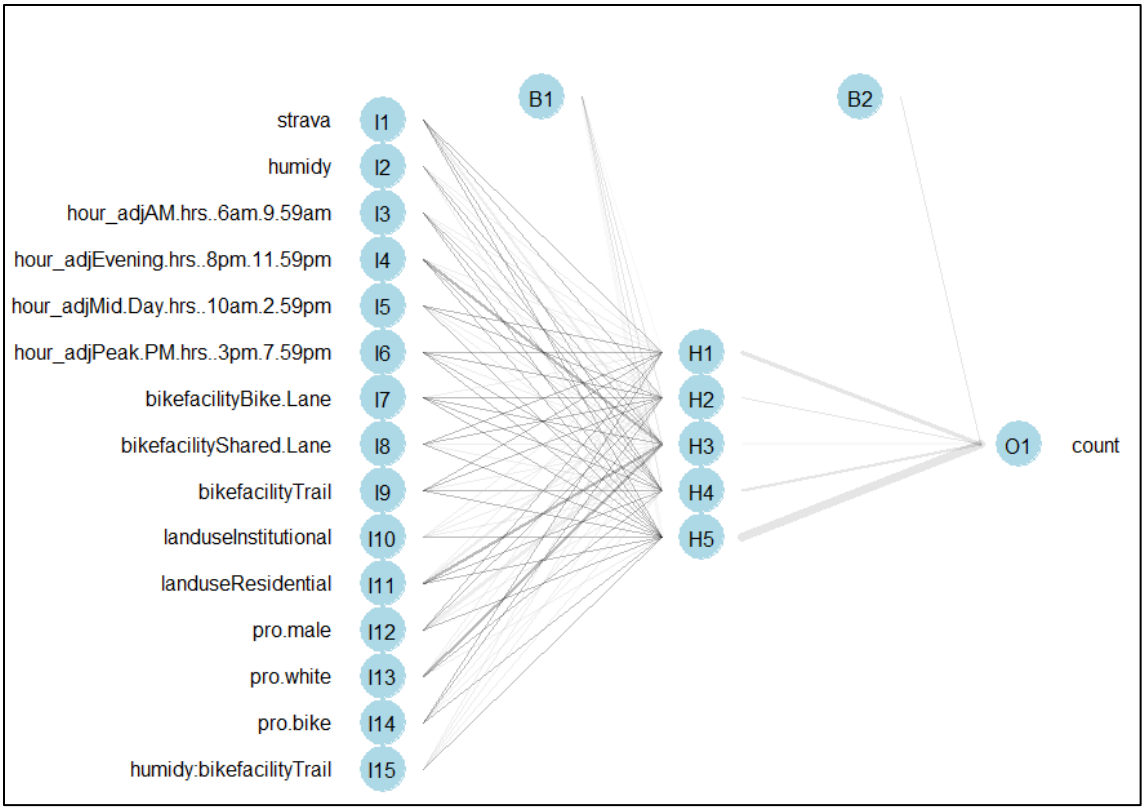


Figure 5.4: Example of Artificial Neural Network used in this research

5.3.4 Classification tree

The decision tree algorithm recursively assigns the observations into the most common occurring class using orthogonal splits. The splitting is performed using various unbiased splitting criteria such as Gini Index and Entropy measure of homogeneity (Strobl et al., 2007). Classification trees have a greater advantage over other machine training tools as they can be visualized graphically for high dimensional cases. They are very easy to explain interpret by non-experts. However, simple classification trees are non-robust and may likely to suffer from high variance when tested in a different dataset (Gareth et al., 2013). Bagging is one of the ensemble methods that was used in our analysis to improve decision tree predictive ability by lowering the variance of the estimates. For bagged classification tree, multiple training datasets are created using bootstrap method. The trees are sequentially built on the training data by learning from the previous trees. The first tree is built by using the response variable, y while subsequent trees are built based on the residuals, r, of their previous trees. This way, boosting creates a splitting classifier

that put an emphasis on the misclassified samples, thus minimizing overall error upon aggregation of all trees (Sutton, 2005).

The tuning parameters for boosting include number of trees, B, shrinkage parameter, λ, which controls the boosting learning rate and the number of trees or interaction depth, d, which controls the complexity of boosting assemble. Usually, cross-validation approach is used to find the optimal tuning parameters that provide a balance of predictive power, efficiency, and flexibility of the model. A general simplified routine for boosting is as follows (Gareth et al., 2013);

- Set f(x)=0, and $r_i = y_i$ for all i in the training set
- For b=1,2,3……B, repeat
    - Fit a tree $\hat{f}^b$ with d splits to the training data (X, r).
    - Update $\hat{f}$ by adding a shrunken new tree: $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^{\,b}(x)$
    - Update the residuals; $r_i \leftarrow r_i + \lambda \hat{f}^{\,b}(x)$
- Output the boosted model; $\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^{\,b}(x)$.
-

5.3.5 Support vector machines

Support vector machines is another popular machine learning technique used in classification problems. It evolves from the idea of the maximum marginal classifier. The maximum marginal classifier works well for separable cases where it is optimized to separate class labels by maximizing the minimal distance from the training observations to the separating hyperplane known as the margin, M. The support vector classifier is an extension of the maximum marginal classifier for inseparable cases. It tolerates a given level of misclassification, ϵ_i by specifying a budget, C that the margin can be violated by training observations, x_{ij} (Gareth et al., 2013). The support vector classifier tends to work well for both linear and non-linear inseparable cases.

The optimization problem for non-linear cases becomes (Gareth et al., 2013);

$$maximize \ M$$

$$\text{Subject to } \ y_i\left(\beta_o + \sum_{j=1}^{p}\beta_{j1}x_{ij} + \sum_{j=1}^{p}\beta_{j2}x_{ij}^2\right) \geq M(1-\epsilon_i)$$

$$\sum_{i=1}^{n}\epsilon_i \leq C, \ \epsilon_i \geq 0, \ \sum_{j=1}^{p}\sum_{k=1}^{2}\beta_{jk}^2 = 1$$

Whereby ϵ_i is the misclassification error rate, M is the width of the margin, C is the tuning parameter (Cost) budget, β_{ij} is the coefficient for the training observation, x_{ij}.

5.3.6 K-Nearest Neighbors

K-Nearest Neighbors is one of the simplest machine learning algorithms that can be used for both classification and regression problems. Like many other machine learning algorithms, it is a non-parametric technique as it doesn't make any assumption about the distribution of the data. In many cases, the kNN will be outperformed by complex machine learning algorithms. However, it is useful to include it in the analysis to establish the baseline of the expected performance before embarking on using more complex algorithm which will utilize more resources. For a regression problem, the output of the kNN are the weighted values of k-nearest neighbors. The number of k-nearest neighbors for a given problem can be found using cross-validation approach. The number of neighbors that yield the output with the smallest root mean square error (RMSE) is usually selected as the optimal number of k-nearest neighbors. The function $f(x_o)$ is then estimated using the average of all the training responses inN_o. Mathematically it can be presented as (Gareth et al., 2013)

$$f(x_o) = \frac{1}{K} \sum_{x_i \in N_o} y_i$$

Whereby, $f(x_o)$ is the average of K closest point.

5.4 Model calibration

In training the model, the sample was randomly divided into training and testing dataset. The training dataset had 80 percent of the total sample and the rest was used for testing the calibrated models. Each model was calibrated using *k*-fold cross-validation resampling strategy in which the dataset is divided into *k* different parts. The model is then fitted on the remaining *k-1* training parts. The left-out part is used to test the model performance based on the calibrated parameters. The procedure is repeated k times, each time leaving out a different part of the dataset for testing the model performance.

The root mean square error (RMSE), which is the mean square difference between the observed and predicted outcome, is used as the criteria for selecting the best model.

The description of variables used in the analysis is provided in Table 5.1 for continuous variables and Table 5.2 for categorical variables. The hourly total cyclist counts ranged from 0 to 112 cyclists with standard deviation of 11 cyclists per hour. For Strava data, the deviation of count across sites was 2.2 cyclist per hour with the hourly count ranging from 0 to 20 cyclists. The average relative humidity had a standard deviation of 22.3% ranging from 10% to 99%. As for the categorical variables, the majority of the segments in our sample had bike lanes followed by trails and shared lanes with markings. The segments passed mostly through the residential (28.8%) and recreational areas (28%) followed by institutional and commercial areas.

Table 5.1: Descriptive summary of the continuous variable used in the analysis

Variable	N	Mean	SD	Min	Max
Hourly Total Cyclist Counts	1520	7.9	11.2	0.0	112.0
Hourly Strava Counts	1520	0.5	2.2	0.0	20.0
Average Hourly Relative humidity (%)	1520	64.7	22.3	10.0	99.0
% of males in the population in the census block	1520	48.7	3.9	39.6	59.0
% of white in the census block population	1520	88.2	10.3	62.0	100.0
% of bikers (workers) in the census block population	1520	0.3	0.9	0.0	3.9

Table 5.2: Descriptive summary of the discrete variables used in the analysis

Variable	Level	Freq.	Percent
Bike facility	Bike Lane	680	44.74
	None	149	9.8
	Shared Lane	266	17.5
	Trail	425	27.96
	Total	1520	100
Hourly Adjustment	Early AM hrs: 12am-5:59am	90	5.92
	AM hrs: 6am-9:59am	270	17.76
	Mid-Day hrs: 10am-2:59pm	509	33.49
	PM hrs: 3pm-7:59pm	487	32.04
	Evening hrs: 8pm-11:59pm	164	10.79
	Total	1520	100
Landuse	Commercial	289	19.01
	Institutional	368	24.21
	Recreational	425	27.96
	Residential	438	28.82
	Total	1520	100

Figure 5.5 shows the cross-validation results for the different machine learning tuning parameters. The tuning parameters are specific for each machine learning model. The optimal tuning parameter was the one that minimizes the overall RMSE value averaged across all the cross-validation samples. Random forest also optimizes its calibration process by doing feature selection. The use of 8 randomly features yielded a minimum RMSE value. For kNN and support vector machine, 7-nearest neighbors and the cost factor of 1 yielded optimal results. The Artificial Neural Network with 5 hidden layers and a weight decay of 0.5 had the lowest RMSE values.

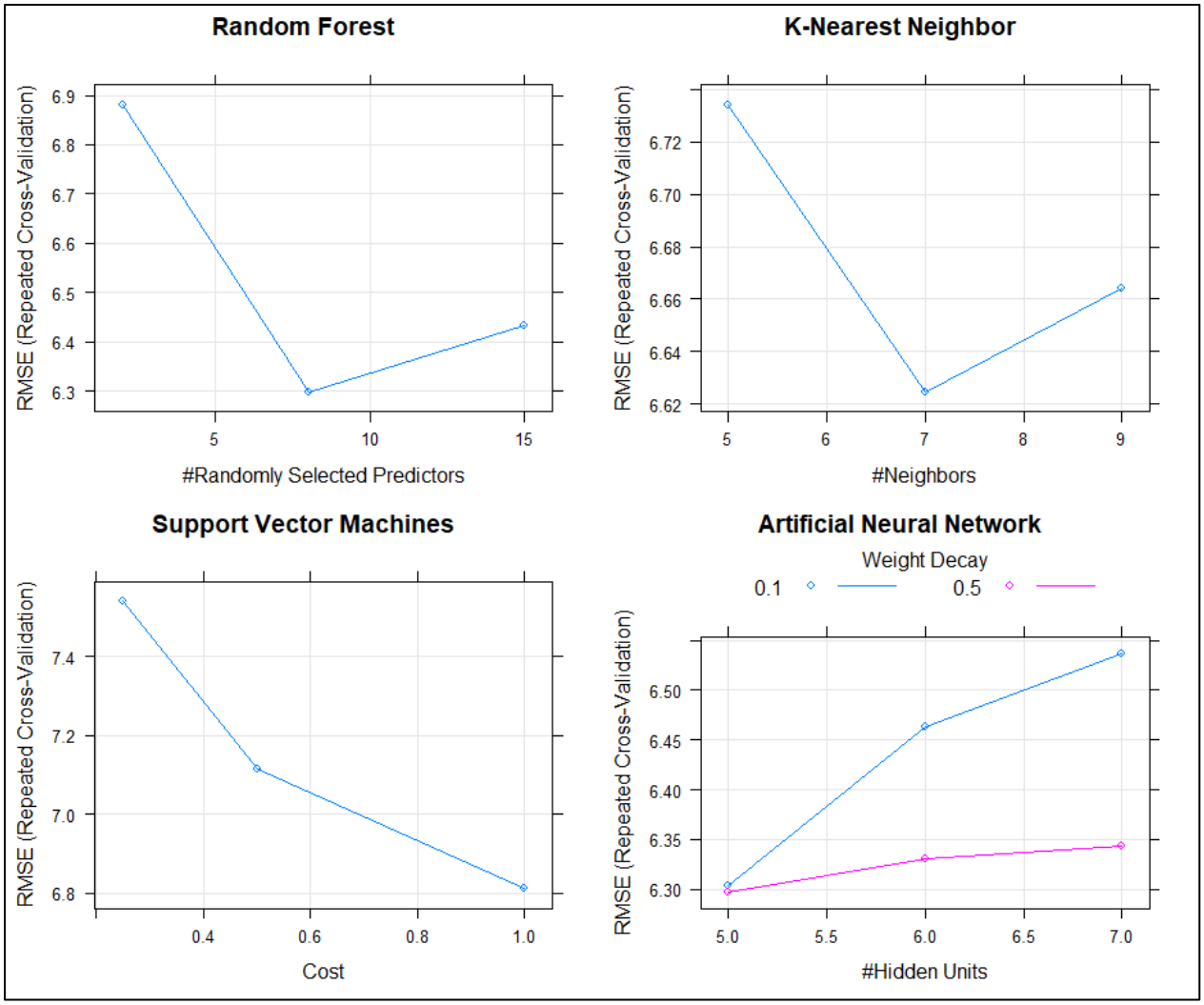


Figure 5.5: Model tuning parameters in the training dataset

5.5 Influence of significant covariance in predicting bicycle count

The Negative binomial model was used for making inferences on how each of the predictor variables impacts total bicycle counts, as shown in Table 5.3. Variables that were significant in the model includes hourly Strava counts, relative humidity(%), hourly adjustment, bicycle facility, land use types, proportion of males in the census block population, proportion of white people in the census block population and proportion of bikers (workers) in the census block population. The exponentiated coefficients or incident rate ratios (IRR) are provided to facilitate the interpretation. An hourly adjustment factor was also added to the model to account for hourly variation of cyclist counts for a

given day. The adjustment factors were divided into five groups with Early AM hrs: 12am-5:59am as a reference hourly adjustment factor.

The Strava count which is one of the main predictor variables had a positive coefficient. By controlling for other confounding variables, one unit increase in hourly Strava cyclist counts in given segment increases the expected number of total hourly cyclist counts by a factor of 1.09 (9 percent).

The effect of weather was also investigated using relative humidity as one of the predictor variables. The change in relative humidity on total cyclists' counts was measured by controlling for trip purpose. It was hypothesized that weather was likely have more impact on recreational trips which are optional compared to commute trips. Previous studies have also shown that the influence of weather condition on outdoor activities such as walking, running and biking is likely be mediated or confounded by trip purpose (Vanky et al., 2017). In our analysis, bicycle facility was used as a surrogate for trip purpose as we observed high correlation between trip purpose and facility type from the survey data. Majority of cyclists (89 percent) who reported to make a recreational trip were found using the trails while cyclists who were commuting were found mostly on other facilities (87 percent) .The interactive variable combining relative humidity and facility type (Trail vs No trail) was created during model calibration to discern the effect of relative humidity on total cyclist counts after controlling for trip type. A unit increase in percent of average hourly relative humidity decreases the total hourly cyclists count by 1.1 percent (IRR=0.989). The total hourly cyclist counts were further reduced by 1.4 percent (IRR=0.986) when considering the unit decrease in relative humidity on recreational trips i.e. trips on trails.

The presence or absence of the dedicated bicycle facility in a given segment had a significant impact on the expected number of cyclists. The influence of each bicycle facility was measured by keeping segments which had no dedicated bicycle facility i.e., shoulder or sidewalk as the reference. Trail was found to attract the highest number of cyclists compared to the rest of the facilities (IRR=5.011) followed by bike lane (IRR=1.672) and a shared lane marking markings/ sharrows (IRR=1.319). The IRR values that were obtained for each of the bicycle facility type were seemingly correct as cyclists are likely to be attracted more on bicycle facilities that offers exclusive space for riding.

Shared lane markings which had the lowest IRR value compared to trail and bike lanes is likely to be used by a segment of intermediate to expert cyclists who are confident to share the road space with the vehicles.

Segments located in residential areas were found to have more cyclists (IRR=1.286) compared to Institutional and commercial areas. The residential areas are the major cyclist trip production areas, thus accounting for most of cyclists that were counted.

Census data at block level were also incorporated in the model. Variables that were tested in the model include age, gender, poverty level, education level, race and means of transportation to work. Only three variables from census data (gender, race and means of transportation to work) were found to significantly affect the total number of cyclists in a given hour as shown in Table 5.3. A unit increase in percent of males in a given census block increases the expected number of cyclists by 14.9 percent after controlling for other covariates in the model. Further, a unit increase in the percent of workers who were using bicycle to commute to work increases the expected number of cyclists in a given hour by 21.1 percent. The percentage of white population also had a significant impact (at 90% confidence level) with IRR value of 1.006.

Table 5.3: Model results for Negative Binomial regression model

Bike counts	IRR	Std. Err.	z	P>z
Strava Count	1.090	0.010	9.710	0.000
Relative Humidity (%)	0.989	0.002	-7.040	0.000
Relative Humidity (%) & Trails	0.986	0.002	-6.930	0.000
Hourly adjustment (Ref: Early AM hrs: 12am-5:59am)				
AM hrs: 6am-9:59am	22.617	5.204	13.550	0.000
Mid-Day hrs: 10am-2:59pm	21.635	4.935	13.480	0.000
Peak PM hrs: 3pm-7:59pm	22.086	5.056	13.520	0.000
Evening hrs: 8pm-11:59pm	13.117	3.065	11.020	0.000
Bike Facility (Ref: Sidewalk/Shoulder)				
Shared Lane	1.319	0.119	3.080	0.002
Bike Lane	1.672	0.143	6.020	0.000
Trail	5.011	0.740	10.910	0.000
Landuse (Ref: Commercial)				
Institutional	1.192	0.108	1.940	0.052
Residential	1.286	0.081	4.020	0.000
Census data				
% of males in the population	1.149	0.007	23.270	0.000
% of white in the population	1.006	0.004	1.690	0.092
% of bikers(workers) in the population	1.211	0.036	6.430	0.000
Constant	0.000	0.000	-15.400	0.000
Alpha	0.418	0.024		

5.6 Assessing the influence of Strava counts on model performance

Sensitivity analysis was conducted to assess the influence of Strava cyclist counts in predicting the total cyclist counts in a given hour. The analysis involved two scenarios. The first scenario measured each models' predictive performance using Strava cyclist count data as one of the predictor variables and the second scenario measured model performance without the Strava cyclist counts. The performance measures that were used were the RMSE value and the percentage of total cyclist count variance explained by predictor variables commonly referred as R-Squared value. The two performance measures complemented each other in explaining the model predictive performances. The model performance with and without Strava cyclist counts is shown in Figure 5.6. For each model, the RMSE and R-Squared values were estimated 100 times using 10 times 10-fold cross-validation technique. This offered a way to make inferences based on distribution of the performance measures as shown in Figure 5.6. The addition of Strava count was found to increase the model performances consistently across all the models that were used in the training dataset. The increase in model performance was indicated by reduction in RMSE value and consequently the increase in R-Squared value. The paired one tail t-test was also conducted for the two scenarios as shown in Table 5.4. The null hypothesis that we wish to reject varied by the given performance measures. The null hypothesis for RMSE states that the average error value before adding Strava counts was less than RMSE value after adding Strava Counts. For R-Squared values, the null hypothesis stated that the average R-Squared values before the addition of Strava count is greater than R-Squared value after the addition of Strava counts. For all the models, the null hypotheses were rejected, concluding that the observed improvement of model performances was indeed significant.

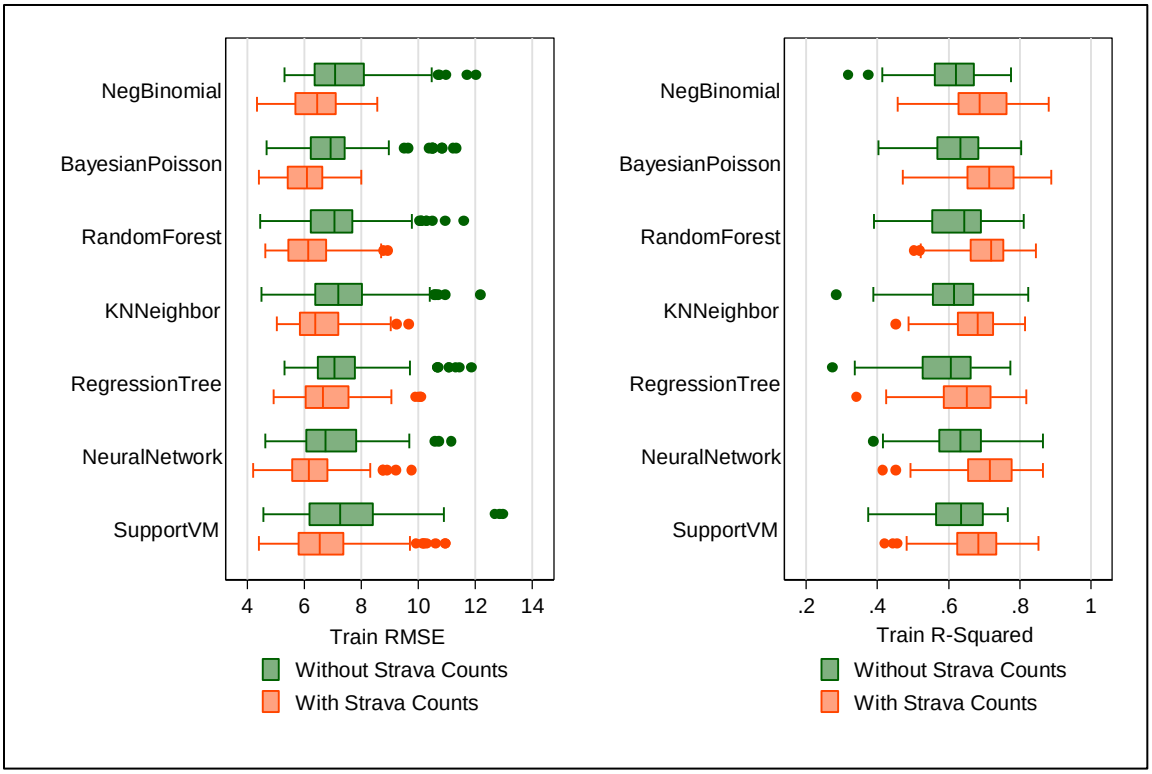


Figure 5.6: Model predictive performance on the training dataset with and without Strava count

Table 5.4: Paired t-test on predict on models' predictive performance on the training dataset with and without Strava Count Feature

Method	RMSE		R-Squared	
	d_{RMSE}= ($RMSE_{str}$-$RMSE_{nostr}$)	$p(d_{RMSE}>=0)$	d_{R2} ($R2_{str}$-$R2_{nostr}$)	$p(d_{R2}<=0)$
NegBinomial	-0.968	0.000	0.082	0.000
RandomForest	-0.867	0.000	0.075	0.000
KNNeighbor	-0.727	0.000	0.072	0.000
RegressionTree	-0.533	0.003	0.055	0.000
NeuralNetwork	-0.787	0.001	0.072	0.000
SupportVM	-0.659	0.002	0.049	0.000

5.7 The best model for predicting cyclists counts

The training RMSE value for each model were compared to select the best model that will be recommended for predicting hourly bicycle counts. Figure 5.7 shows the models' performances based on the RMSE values obtained from the training dataset. The best model that yielded the lowest RMSE value was selected as our best model for predicting bicycle counts.

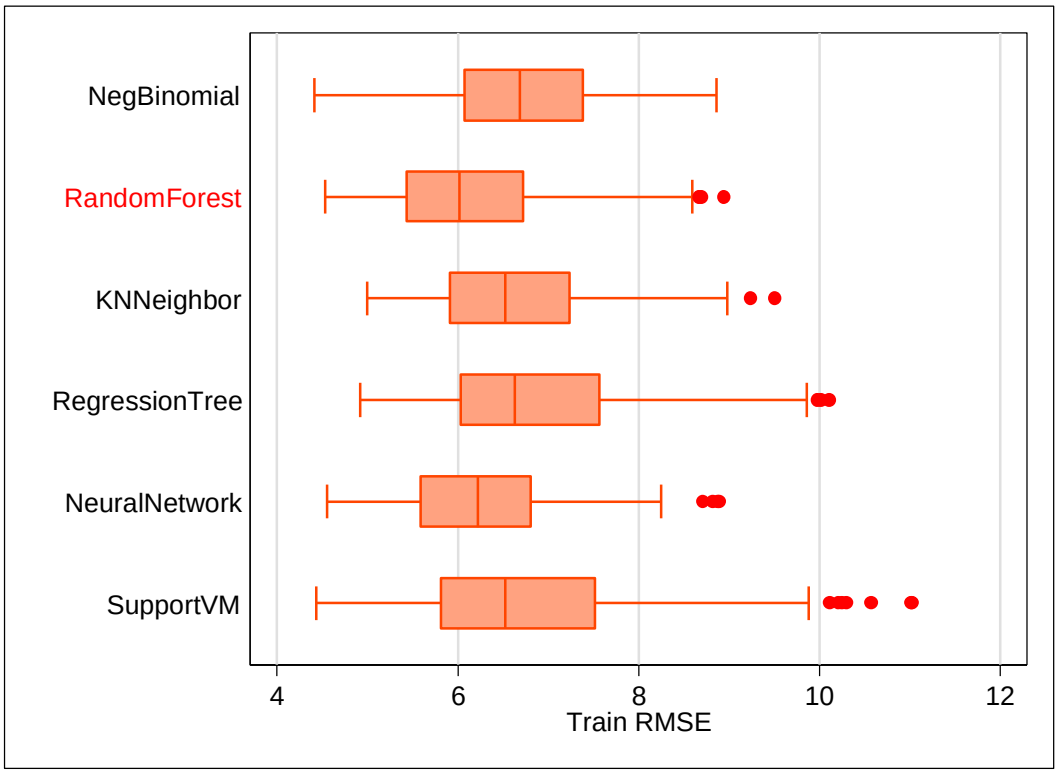


Figure 5.7: Final model section for predicting the bicycle exposure

The performance of the calibrated random forest model was tested using the remaining 30 percent of the data that was not used in training the model. The out-of-sample prediction helps to discern the extent to which the model is generalizable. The model was able to explain about 71 percent of variation in hourly cyclists counts (R-Squared=0.71) when tested on a different dataset. The final RF model was used to create an online tool available at https://trclc.shinyapps.io/BikeExposure/. This tool utilizes

crowdsourced data in addition to other data such as bicycle facility type, landuse, census data, relative humidity level, and time-of-the day to estimate hourly bicycle volume. Details are available in Appendix 8.6.

5.8 Simulation study of Strava penetration rates

The previous section assesses the significance of Strava count in predicting the total counts based on the current observed penetration rate of 7 percent. However, the number of Strava users keeps on growing each year. Definitely, the increase in Strava penetration rates will have an impact on model predictive performances in estimating total cyclist counts. Therefore, it is imperative to be able to envisage analytically how the model performance will change in the coming years as number of cyclists using Strava app keep on increasing. To achieve this, we conducted a simulation study by incrementing Strava counts from the base condition without distorting the observed hourly variation of Strava counts across our sites. Figure 5.8 shows the hourly distribution of different simulated Strava penetration rates aggregated for all the sites. The models were then calibrated by altering simulated Strava counts while keeping other predictor variables unchanged.

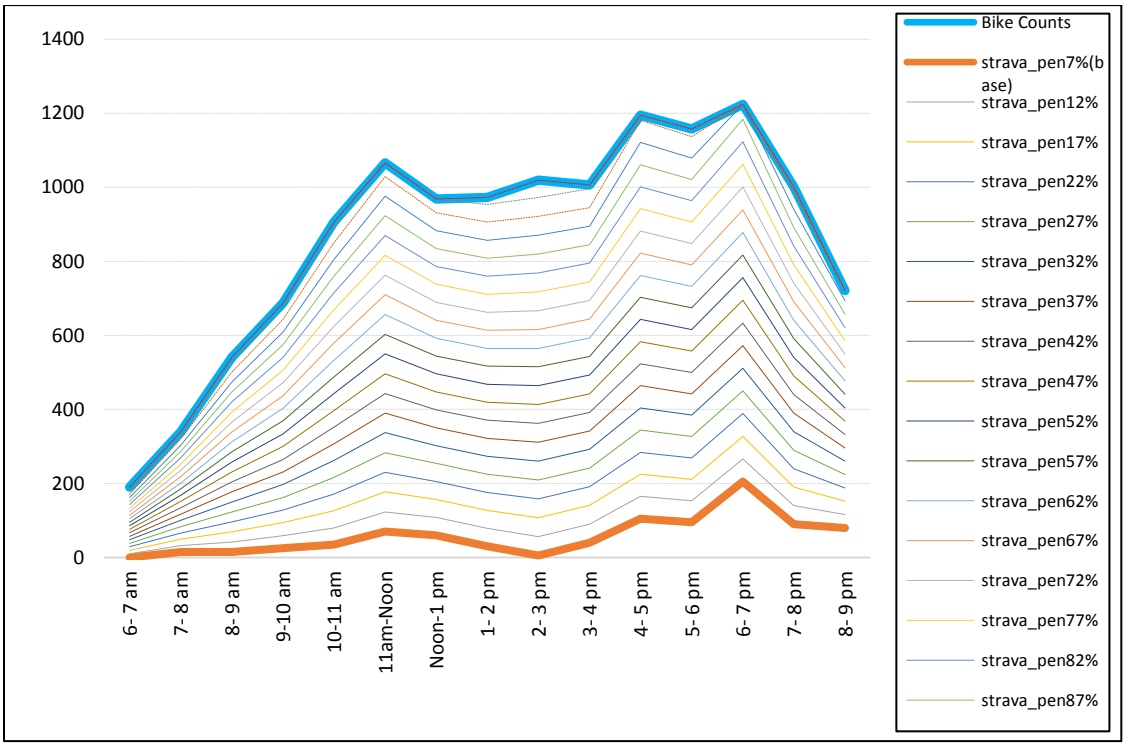


Figure 5.8: Illustration of different hourly Strava penetration rates for all sites

A sensitivity analysis of the predictor variables was conducted for the five most important predictor variables which were Strava counts, relative humidity, the proportion of male population in a census block group, presence of a bike lane and presence of a trail. Figure 5.9 and Figure 5.10 show the variable importance plot of predictor variables across different model approaches. The importance of the variable *i* was based on how the model performances changes due to the presence of variable *i* as opposed to its absence. The change in predictive performances of each model, which for our case was the change in RMSE was scaled from 0 to 100 to facilitate the comparison across different models. At the base condition i.e., Strava penetration rate of 7 percent, Strava count was the most important predictor in 2 out of 6 modeling approaches. A simulated Strava penetration rate of 10 percent consistently indicated Strava as the most important predictor for all the 7 model approaches. Apparently, a unit change in the percent of simulated Strava penetration rate has a very significant influence on the model's performances.

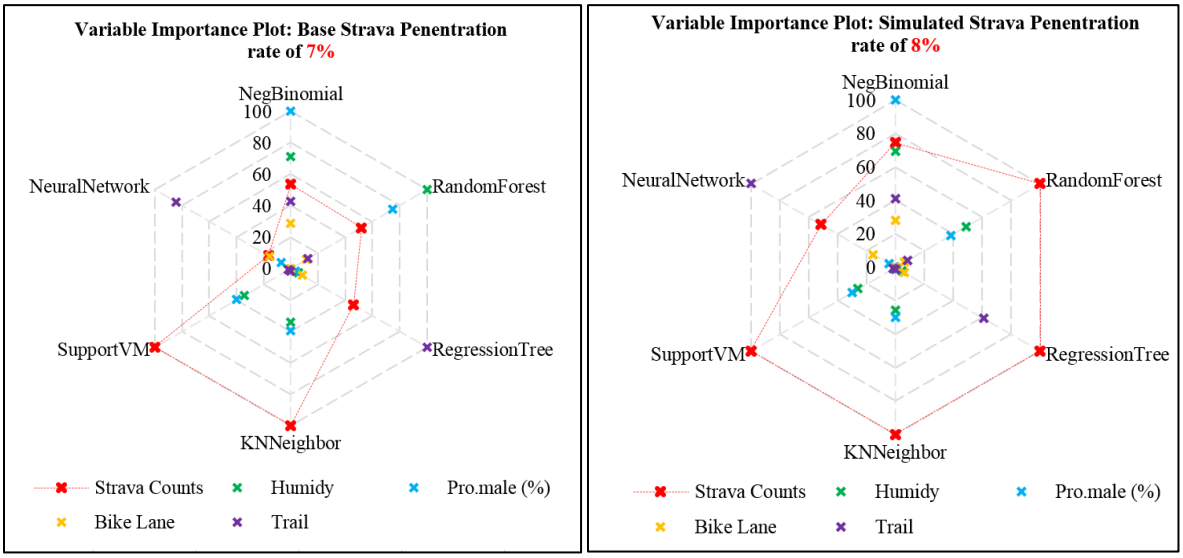


Figure 5.9: Variable importance plot for Strava penetration rate of 7 percent and 8 percent

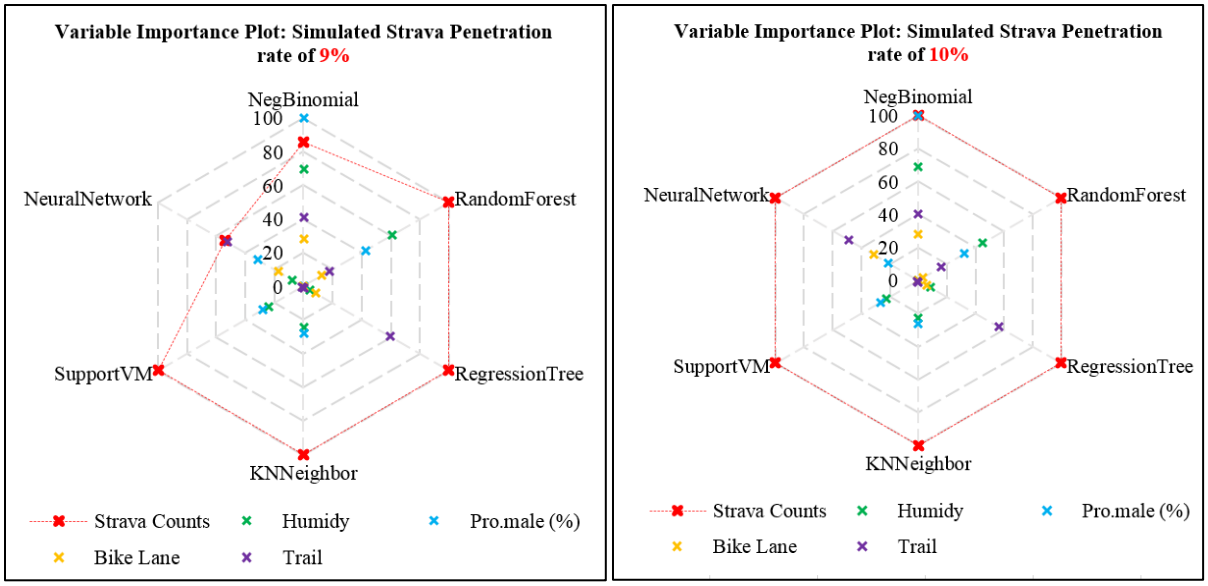


Figure 5.10: Variable importance plot for Strava penetration rate of 9 percent and 10 percent

Figure 5.11 shows the models' predictive performances on the test dataset for different simulated Strava penetration rates. Figure 5.12 shows the model performances only for Random Forest which outperformed other models. From Figure 5.12, the inclusion of Strava counts based on the current penetration rate increases the percent of cyclist count variance explained by the model from 65 percent to 71 percent. The graph shows a sharp increase of model performance from the base Strava penetration rate of 7% to 10% and gentle increase in model performance up to a simulated Strava penetration rate of 40%. Thereafter, the increase in simulated Strava penetration rate do not offer appreciable improvement of model's predictive performances.

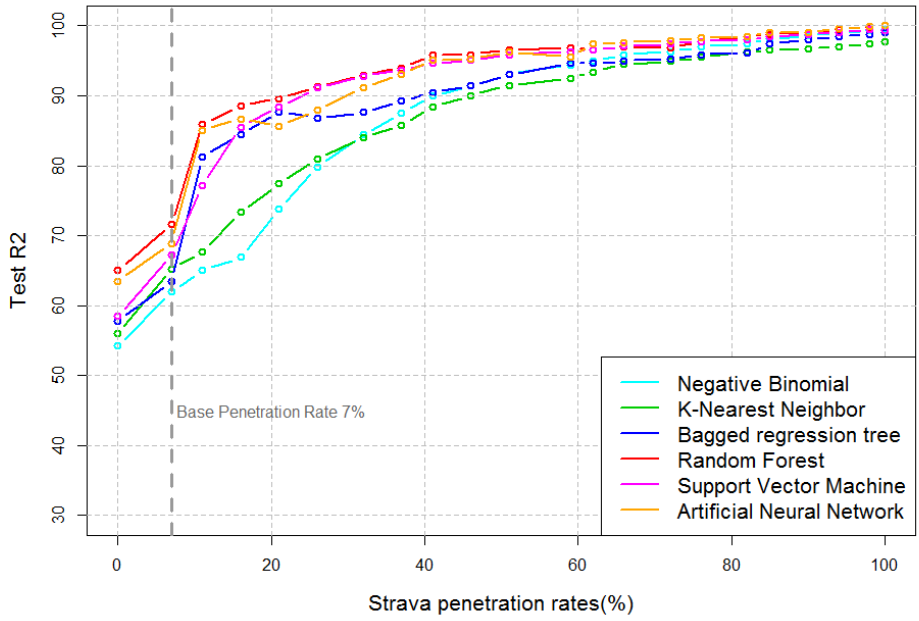


Figure 5.11: Models performance at various Strava penetration rates on the test dataset

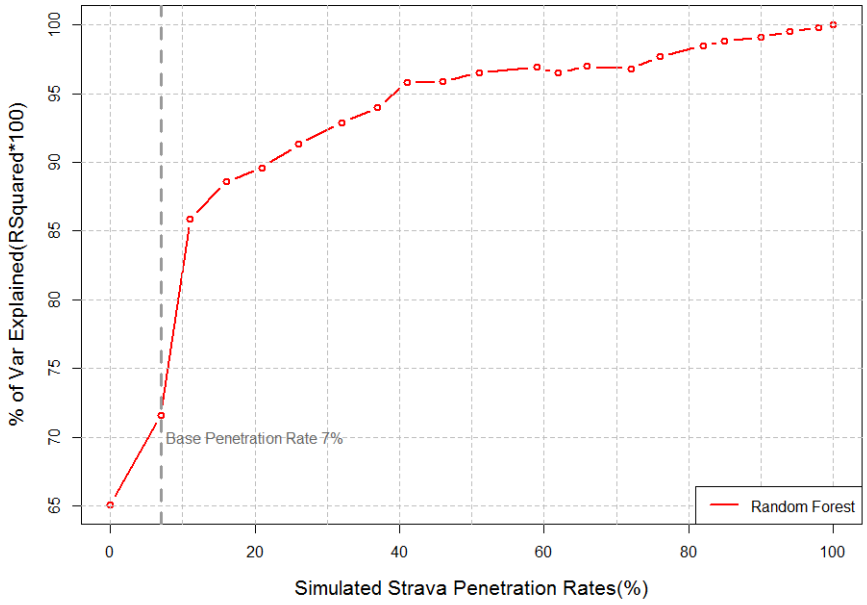


Figure 5.12: Best model performance at various Strava Penetration rates on the test dataset

6 Conclusions and Recommendations

As the technology continues to advance, crowdsourcing will continue to offer fast, plausible, convenient and cheaper platform of obtaining useful information from the crowd. In transportation, there is a growing interest from urban planners to use crowdsourced data of pedestrian and bicycle activities in network planning of non-motorized facilities. Traditional means that are currently used to collect bicycle information lack spatial and temporal details of cyclists' activities which are essential for effective planning of bicycle facilities at network level. This research investigated opportunities and limitations of integrating traditional cyclists count data with the crowdsourced cycling data. In this research, cyclists' activities from the Strava Metro data was used as the source of the crowdsourced cyclists' activities. Strava is one among commercial fitness apps that sells tracking data to public agencies to assist in making informed decisions and planning better facilities for pedestrians and bicyclists.

Ground truth cyclist activities on selected segments were recorded using video data and later processed to obtain cyclist counts in each given hour. At each site, the video recordings were observed for a week to capture hourly variation within a day and daily variations within a week. Site characteristics at each of the selected segments were collected and joined spatially with other data sources.

A field survey was conducted to identify the proportion of cyclists that were using fitness app to track their cycling activities. Further, the analysis investigated trip characteristics and demographics of cyclists who reported to use cycling apps in comparison to cyclists who reported not to use any of the fitness apps. The survey was conducted in the city of Ann Arbor and Grand Rapids. A total of 321 cyclists participated in the survey. Interesting findings were obtained that have practical significance to planners and engineers who are currently using crowdsourced data or considering in the future to incorporate the crowdsourced data for assessment and planning of bicycle facilities.

About 16 percent of cyclists that were surveyed reported to use Strava app while 18 percent reported to use other fitness apps. The rest of the cyclists (66 percent) reported not using any fitness app. From the survey data, the reported average proportion of

cyclists who were using Strava app was 16 percent. This percentage was higher than the actual average proportion (7 percent) that was obtained after comparing the Strava cyclist counts and total ground truth counts from the selected segments. This might indicate the needs to adjust for penetration rates of Strava and other fitness apps users obtained from survey data to reflect the actual number of cyclists who are using fitness app(s) in the total cycling population.

Furthermore, the analysis investigated if the cycling activities data coming from different fitness apps represent different cycling population demography. There has been a growing concern of using crowdsourcing data due to possible inequalities that may arise in distribution of resources and services. The resulting decisions may favor a segment of population that have digital access and therefore hinder opportunities for the remaining disadvantaged group (Griffin and Jiao, 2019). Based on the results of this study, the odds of using a Strava fitness app were likely to increase by 6.7 percent for cyclists' age 25-34 years and 52 percent for cyclists age 35-44 years. Conversely, young cyclists (less than 25 years) and older cyclists (55 years and greater) had higher odds of utilizing other fitness apps. The dispersion of apps usage was also observed when analyzing app usage by gender. Male cyclists were more likely use Strava app (OR=2.204) while the female cyclists had higher odds (OR=1.813) of using other fitness app(s). Therefore, fusing cycling activities from different fitness apps can help to address the issue of digital inequality inherent in crowdsourcing platforms. However, there is a need to address the setbacks of data integration first before thinking about fusing dataset from different fitness apps. These setbacks include variation in data accuracy, data quality and possible redundancy in representation of certain segment of cycling population.

Another important variable that was explored in our survey data is cyclist trip purpose. The cyclists were asked to report the purpose of their trips – either recreational or commuting. By controlling for cyclists age, gender and level of experience, it was found that cyclists who were making recreational trips had higher odds of using fitness tracking apps (Strava or/and any other fitness apps) compared to cyclists who were taking commute trips. Descriptive statistics of Strava cyclist activities from February 2018 to January 2019 was consistency with the survey results. The commute trip constituted only 6 percent of the total cyclist activities. Therefore, it is likely for the fusion of different

crowdsourced cycling data to adjust for inequality in demographic representation, but not trip characteristic. Other data types such as roadway, landuse and census data can be used to account for underrepresentation of commute trips when estimating the total number of cyclists' trips.

The estimation of bicycle exposure was one of the main components of this research. A predictive approach was developed to estimate the hourly cyclist volume using Strava data as one of the independent variables. Other variables that were included in the model include average hourly relative humidity (%), census data such as proportion of males, white and bikers in a given census block where the roadway segment is located, bike facility information, and land use types. Hourly adjustment factors were also applied in the model. Different probabilistic and machine learning models were tested using dataset, namely, the Negative Binomial model, Random Forest, Support Vector Machines, Artificial Neural Network and k-Nearest Neighbors. In terms of prediction, the Random Forest was found to have a better performance, explaining about 71% of the variations in total bicycle volume(R-Squared=0.71). The final product was a method to estimate hourly bicycle volume using the random forest model by inputting crowdsourced data in addition to other data such as bicycle facility type, landuse, census data, relative humidity level, and time-of-the day. This method can be implemented through an online tool available at https://trclc.shinyapps.io/BikeExposure/.

The contribution of Strava data in overall model predictive performance was assessed by conducting a sensitivity analysis of Strava data by comparing model performances with and without Strava cycling counts. For each model that was used in the training dataset, consistent results were obtained showing an improvement in model performance after adding Strava data. The significance of the improvement was tested by a paired t-test of the models' performance indicators (RMSE, R-Squared) before and after adding Strava data in the model estimation. Again, the improvement in model performance was significant at 95% confidence level for all the models that were tested in the data. In conclusion, Strava data showed a significant contribution to overall model predictive performance.

The research team also explored what might be the change in model performance in the coming years. Number of Strava users is growing and its growth will have an impact

on future model predictive performances. The research team conducted a simulation study to assess the change in model performance based on different simulated Strava penetration rates. The base Strava penetration rate was 7% and the penetration rates were incrementally increased without distorting the observed variance of bicycle counts across the sites. The output of this analysis was a graph showing how the model performances (R-Squared) changes with simulated Strava penetration rate. The graph can be useful to planners and engineers who are currently or in the future considering using Strava data for estimating bicycle exposure. Example on how the graph can assist planner to make informed decisions is discussed in Chapter 5.

In conclusion, this research demonstrated how crowdsourced cycling data can be integrated with the traditional counts to improve the estimation of bicycle exposure-an essential component in planning of non-motorized facilities at a given roadway network. With some additional data, the procedure developed in this research can be extended to estimate pedestrian exposure, which is also a fundamental component of urban planning.

7 References

Afzalan, N., Sanchez, T., 2017. Testing the Use of Crowdsourced Information: Case Study of Bike-Share Infrastructure Planning in Cincinnati, Ohio. Urban Plan. 2 3 , 33. doi:10.17645/up.v2i3.1013

Assemi, B., Schlagwein, D., Safi, H., Mesbah, M., 2015. Crowdsourcing as a method for the collection of revealed preference data. Proc. - 9th IEEE Int. Symp. Serv. Syst. Eng. IEEE SOSE 2015 30, 378–382. doi:10.1109/SOSE.2015.52

Barbier, G., Zafarani, R., Gao, H., Fung, G., Liu, H., 2012. Maximizing benefits from crowdsourced data. Comput. Math. Organ. Theory 18 3 , 257–279. doi:10.1007/s10588-012-9121-2

Blanc, B., Figliozzi, M., 2016. Modeling the Impacts of Facility Type, Trip Characteristics, and Trip Stressors on Cyclists' Comfort Levels Utilizing Crowdsourced Data. Transp. Res. Rec. J. Transp. Res. Board 2587 2587 , 100–108. doi:10.3141/2587-12

Blanc, B., Figliozzi, M., Clifton, K., 2016. How Representative of Bicycling Populations Are Smartphone Application Surveys of Travel Behavior? Transp. Res. Rec. J. Transp. Res. Board 2587, 78–89. doi:10.3141/2587-10

Boss, D., Nelson, T., Winters, M., Ferster, C.J., 2018. Using crowdsourced data to monitor change in spatial patterns of bicycle ridership. J. Transp. Heal. 9 March , 226–233. doi:10.1016/j.jth.2018.02.008

Bonaccorso, G., 2018. Mastering Machine Learning Algorithms. Packt Publishing.

Chien, J.-T., 2019. Deep Neural Network. Source Sep. Mach. Learn. 259–320. doi:10.1016/B978-0-12-804566-4.00019-X

Boyd, D., Crawford, K., 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. Inf. Commun. Soc. 15 5 , 662–679. doi:10.1080/1369118X.2012.678878

Brandway, M., Bettenardi, A., Smith, P., Lyons, S., Weidner, T., Juul, B., Sperley, M., 2014. Purpose and Need For the Strava Bicycle Data Project April , 1–8.

Branion-Calles, M., Nelson, T., Winters, M., 2016. Comparing Crowdsourced Near Miss and Collision Cycling Data and Official Bike Safety Reporting. Transp. Res. Rec. 480 , 778–782. doi:10.3141/2662-01

Buckland, L., Jones, M., 2008. Estimating Bicycle and Pedestrian Demand in San Diego, TRB 2008 Annual Meeting.

Chien, J.-T., 2019. Deep Neural Network. Source Sep. Mach. Learn. 259–320. doi:10.1016/B978-0-12-804566-4.00019-X

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

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

Cupples, J., Ridley, E., 2008. Towards a heterogeneous environmental responsibility: Sustainability and cycling fundamentalism. Area. doi:10.1111/j.1475-4762.2008.00810.x

Dhakal, N., Cherry, C.R., Ling, Z., Azad, M., 2018. Using CyclePhilly data to assess

wrong-way riding of cyclists in Philadelphia. J. Safety Res. 67, 145–153. doi:10.1016/j.jsr.2018.10.004
Estellés-Arolas, E., González-Ladrón-De-Guevara, F., 2012. Towards an integrated crowdsourcing definition. J. Inf. Sci. 38 2 , 189–200. doi:10.1177/0165551512437638
Ferster, C.J., Nelson, T., Robertson, C., Feick, R., 2018. Current Themes in Volunteered Geographic Information, in: Comprehensive Geographic Information Systems. Elsevier, pp. 26–41. doi:10.1016/B978-0-12-409548-9.09620-2
Figliozzi, M.A., Blanc, B.P., 2015. Evaluating the Use of Crowdsourcing as a Data Collection Method for Bicycle Performance Measures and Identification of Facility Improvement Needs 123.
Fröhlich, S., Springer, T., Dinter, S., Pape, S., Schill, A., Krimmling, J., 2016. BikeNow: A Pervasive Application for Crowdsourcing Bicycle Traffic Data. Proc. 2016 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Adjun. - UbiComp '16 1408–1417. doi:10.1145/2968219.2968419
Gareth, J., Daniela, W., Trevor, H., Rober, T., 2013. An Introduction to Statistical Learning with Applications in R. Springer Texts in Statistics. doi:10.1007/978-1-4614-7138-7
Griffin, G., Jiao, J., 2015a. Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. J. Transp. Heal. 2 2 , 238–247. doi:10.1016/j.jth.2014.12.001
Griffin, G., Jiao, J., 2015b. Crowdsourcing Bicycle Volumes : Exploring the role of volunteered geographic information and established monitoring methods. Compend. Transp. Res. Board Annu. Meet. 27 1 , 1–19.
Griffin, G., Nordback, K., Götschi, T., Stolz, E., Kothuri, S., 2014. Transportation Research Circular E-C183, Monitoring Bicyclist and Pedestrian Travel and Behavior, Current Research and Practice. Transp. Res. Board.
Griffin, G.P., Jiao, J., 2019. The Geography and Equity of Crowdsourced Public Participation for Active Transportation Planning. Transp. Res. Rec. doi:10.1177/0361198118823498
Griswold, J.B., Medury, A., Schneider, R.J., 2011. Pilot Models for Estimating Bicycle Intersection Volumes. Transp. Res. Rec. J. Transp. Res. Board. doi:10.3141/2247-01
Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, The Elements of Statistical Learning. doi:10.1007/978-0-387-98135-2
Haworth, J., 2016. Investigating the potential of activity tracking app data to estimate cycle flows in urban areas. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 41 July , 515–519. doi:10.5194/isprsarchives-XLI-B2-515-2016
Heesch, K.C., Langdon, M., 2016. The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour. Heal. Promot. J. Aust. 27 3 , 222–229. doi:10.1071/HE16032
Hilbe, J., 2011. Modeling count data. Cambridge University Press.
Hochmair, H., Bardin, E., Ahmounda, A., 2016. Estimating bicycle trip volume for Miami-Dade county frm Strava Tracking Data 1–17.
Hood, J., Sall, E., Charlton, B., 2011. A GPS-based bicycle route choice model for San Francisco, California. Transp. Lett. 3 1 , 63–75. doi:10.3328/TL.2011.03.01.63-75

Howe, J., 2006. The Rise of Crowdsourcing. Wired Mag. 14 06 , 1–5. doi:10.1086/599595
Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern classification. New York John Wiley, Sect.
Jestico, B., Nelson, T., Winters, M., 2016. Mapping ridership using crowdsourced cycling data. J. Transp. Geogr. 52, 90–97. doi:10.1016/j.jtrangeo.2016.03.006
Jestico, B., Nelson, T.A., Potter, J., Winters, M., 2017. Multiuse trail intersection safety analysis: A crowdsourced data perspective. Accid. Anal. Prev. 103 March , 65–71. doi:10.1016/j.aap.2017.03.024
Krykewycz, G.R., Pollard, C., Canzoneri, N., He, E., 2012. Web-Based “Crowdsourcing” Approach to Improve Areawide “Bikeability” Scoring. Transp. Res. Rec. J. Transp. Res. Board 2245, 1–7. doi:10.3141/2245-01
Leao, S., Izadpanahi, P., Leao, S.Z., Lieske, S.N., Pettit, C., 2017a. Factors motivating bicycling in Sydney : Analysing crowd-sourced data July .
Leao, S., Lieske, S., Conrow, L., Doig, J., Mann, V., Pettit, C., 2017b. Building a National-Longitudinal Geospatial Bicycling Data Collection from Crowdsourcing. Urban Sci. 1 3 , 23. doi:10.3390/urbansci1030023
Leao, S., Lieske, S., Pettit, C., 2017c. Validating crowdsourced bicycling mobility data for supporting city planning. Transp. Lett. 7867, 1–12. doi:10.1080/19427867.2017.1401198
Lee, K., Sener, I., 2019. Understanding Potential Exposure of Bicyclists on Roadways to Traffic-Related Air Pollution: Findings from El Paso, Texas, Using Strava Metro Data. Int. J. Environ. Res. Public Health 16 3 , 371. doi:10.3390/ijerph16030371
Lesiv, M., Moltchanova, E., Schepaschenko, D., See, L., Shvidenko, A., Comber, A., Fritz, S., 2016. Comparison of data fusion methods using crowdsourced data in creating a hybrid forest cover map. Remote Sens. 8 3 . doi:10.3390/rs8030261
McArthur, D.P., Hong, J., 2019. Visualising where commuting cyclists travel using crowdsourced data. J. Transp. Geogr. 74 December 2018 , 233–241. doi:10.1016/j.jtrangeo.2018.11.018
McKenzie, B., 2014. Modes Less Traveled: Bicycling and Walking to Work in the United States, 2008-2012. Am. Community Surv. Reports.
Misra, A., Gooze, A., Watkins, K., Asad, M., Le Dantec, C., 2014. Crowdsourcing and Its Application to Transportation Data Collection and Management. Transp. Res. Rec. J. Transp. Res. Board 2414, 1–8. doi:10.3141/2414-01
Molino, J.A., Kennedy, J.F., Johnson, P.L., Beuse, P.A., Emo, A.K., Do, A., 2009. Pedestrian and Bicyclist Exposure to Risk: Methodology for Estimation in an Urban Environment. Transp. Res. Rec. J. Transp. Res. Board. doi:10.3141/2140-16
Musakwa, W., Selala, K.M., 2016. Mapping cycling patterns and trends using Strava Metro data in the city of Johannesburg, South Africa. Data Br. 9, 898–905. doi:10.1016/j.dib.2016.11.002
Oh, J.-S., Kwigizile, V., Van Houten, R., McKean, J., Abasahl, F., Dolatsara, H., Wegner, B., Clark, M., 2013. Development of Performance Measures for Non-Motorized Dynamics 296p.
Pal, R., 2017. Chapter 7 - Predictive modeling based on random forests, in: Pal, R.B.T.-P.M. of D.S. (Ed.), . Academic Press, pp. 149–188. doi:https://doi.org/10.1016/B978-0-12-805274-7.00007-5

Piatkowski, D., Marshall, W., Afzalan, N., 2015. Does Crowdsourcing Community Input Lead to Equitable Transportation? The Application of Web-based Tools to Inform Bikeshare System Development. Statew. Agric. L. Use Baseline 2015 1, 1–11. doi:10.1017/CBO9781107415324.004
Proulx, F.R., Pozdnukhov, A., 2017. Bicycle Traffic Volume Estimation using Geographically Weighted Data Fusion. J. Transp. Geogr. 1–14.
Ryus, P., Ferguson, E., Laustsen, K.M., Schneider, R.J., Proulx, F.R., Hull, T., Miranda-Moreno, L., 2016. Guidebook on Pedestrian and Bicycle Volume Data Collection, Guidebook on Pedestrian and Bicycle Volume Data Collection. doi:10.17226/22223
Sanders, R.L., Frackelton, A., Gardner, S., Schneider, R., Hintze, M., 2017. Ballpark Method for Estimating Pedestrian and Bicyclist Exposure in Seattle, Washington. Transp. Res. Rec. J. Transp. Res. Board 2605, 32–44. doi:10.3141/2605-03
Selala, M.K., Musakwa, W., 2016. The potential of strava data to contribute in non-motorised transport (NMT) planning in johannesburg. Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci. - ISPRS Arch. 41 July , 587–594. doi:10.5194/isprsarchives-XLI-B2-587-2016
Smith, A., 2015. Crowdsourcing Pedestrian and Cyclist Activity Data. White Pap. Ser. January , 34.
Statistica, 2016. Number of cyclists/bike riders in the U.S. 2016 [WWW Document]. URL https://www.statista.com/statistics/227415/number-of-cyclists-and-bike-riders-usa/ (accessed 12.6.17).
Strauss, J., Miranda-Moreno, L.F., Morency, P., 2015. Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. Accid. Anal. Prev. 83, 132–142. doi:10.1016/j.aap.2015.07.014
Strobl, C., Boulesteix, A.L., Augustin, T., 2007. Unbiased split selection for classification trees based on the Gini Index. Comput. Stat. Data Anal. doi:10.1016/j.csda.2006.12.030
Sultan, J., Ben-haim, G., Haunert, J., Dalyot, S., 2015. Using Crowdsourced Volunteered Geographic Information for Analyzing Bicycle Road Networks. Int. Fed. Surv. Artic. Mon. – December 2015 December , 1–14.
Sutton, C., 2005. 11 - Classification and Regression Trees, Bagging, and Boosting, in: Handbook of Statistics. pp. 303–329. doi:https://doi.org/10.1016/S0169-7161(04)24011-1
Vanky, A.P., Verma, S.K., Courtney, T.K., Santi, P., Ratti, C., 2017. Effect of weather on pedestrian trip count and duration: City-scale evaluations using mobile phone application data. Prev. Med. Reports 8, 30–37. doi:10.1016/j.pmedr.2017.07.002
Watkins, K., Ammanamanchi, R., LaMondia, J., Le Dantec, C.A., 2016. Comparison of Smartphone-based Cyclist GPS Data Sources, in: Transportation Research Board 95th Annual Meeting.
Whitfield, G.P., Ussery, E.N., Riordan, B., Wendel, A.M., 2016. Association Between User-Generated Commuting Data and Population-Representative Active Commuting Surveillance Data — Four Cities, 2014–2015. MMWR. Morb. Mortal. Wkly. Rep. 65 36 , 959–962. doi:10.15585/mmwr.mm6536a4
Wu, M., Frias-Martinez, V., 2015. Crowdsourcing biking times. Proc. 2015 ACM Int. Jt. Conf. Pervasive Ubiquitous Comput. Proc. 2015 ACM Int. Symp. Wearable Comput. - UbiComp '15 1123–1131. doi:10.1145/2800835.2800976

8 APPENDICES

8.1 Field data collection: City of Ann Arbor

No	Road Name	Land use	Bike facility	Lat	Long
1	5th@Liberty	Commercial	Bike Lane	42.278684	-83.746129
2	Murfin@Plymouth	Institutional	Shared Lane	42.296415	-83.719707
3	Huron@Dexter	Residential	None	42.281959	-83.765323
4	Division@Packard	Institutional	Bike Lane	42.275278	-83.744247
5	Platt@Packard	Residential	Bike Lane	42.243612	-83.700060
6	Nixon@Green	Residential	Bike Lane	42.317163	-83.707602
7	Plymouth@Huron Pkwy	Institutional	Bike Lane	42.302561	-83.705764
8	Miller@1st	Residential	Shared Lane	42.283312	-83.750237
9	State@Liberty	Commercial	Shared Lane	42.279812	-83.740804
10	State@Packard	Commercial	None	42.270327	-83.740595

Physical addresses

No	Camera Location	Physical address
1	Light pole	325-301 S 5th Ave, Ann Arbor, MI 48104
2	Light pole	1799-1691 Murfin Ave, Ann Arbor, MI 48105
3	Electric pole	1499-1419 W Huron St, Ann Arbor, MI 48103
4	Electric pole	494-580 S Division St, Ann Arbor, MI 48104
5	Electric pole	3086-3166 Platt Rd, Ann Arbor, MI 48108
6	Light pole	3029-3009 Nixon Rd, Ann Arbor, MI 48105
7	Light pole	2777-2703 Plymouth Rd, Ann Arbor, MI 48105
8	Electric pole	200-212 Miller Ave, Ann Arbor, MI 48104
9	Light pole	233-201 S State St, Ann Arbor, MI 48104
10	Electric pole	917-901 S State St, Ann Arbor, MI 48104

Schedule

No	Name	Week (Date)				
		1	2	3	4	5
		(6/4/18-6/10/18)	(6/11/18-6/17/18)	(6/18/18-6/24/18)	(6/25/18-7/1/18)	(7/2/18-7/8/18)
1	5th@Liberty					
2	Murfin@Plymouth					
3	Huron@Dexter					
4	Division@Packard					

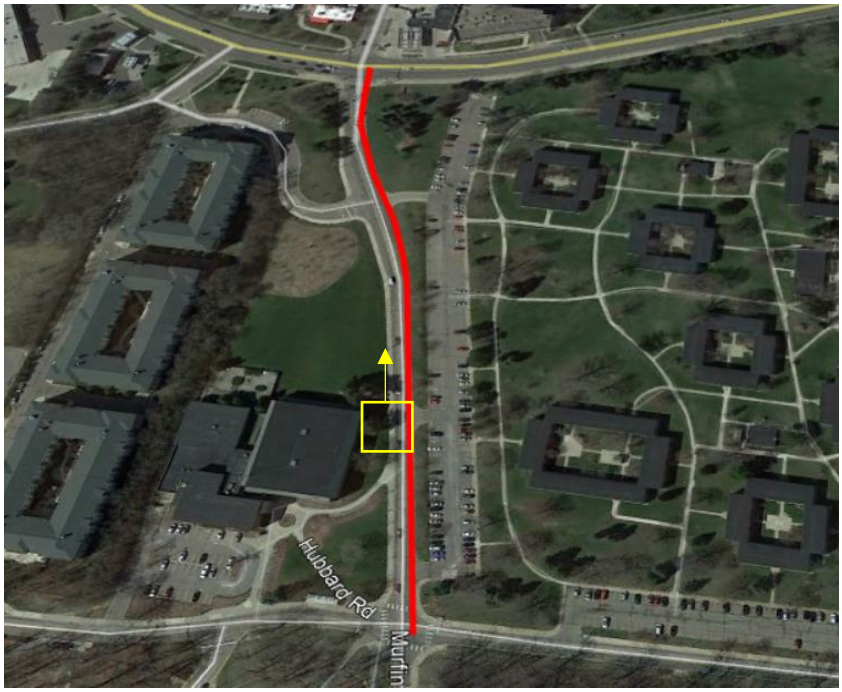
5	Platt@Packard					
6	Nixon@Green					
7	Plymouth@Huron					
8	Miller@1st					
9	State@Liberty					
10	State@Packard					

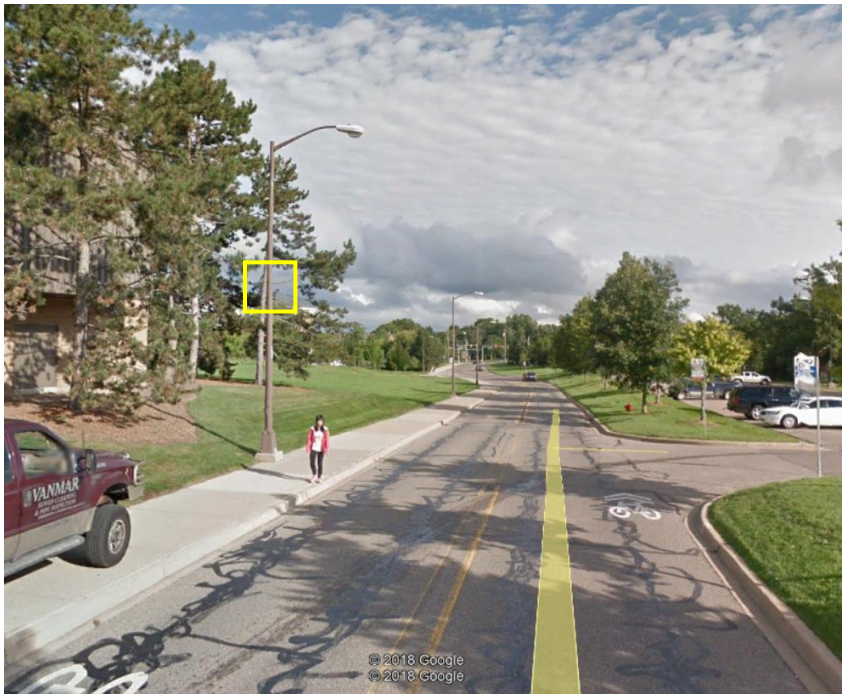
Details of each site

No	Segment Name	Road type	Bike facility	Land use	**Owner**	Lat	Long
1	5th@Liberty	Collector	Bike Lane	Commercial	**DTE**	42.278684	-83.746129

Video camera location: Light pole

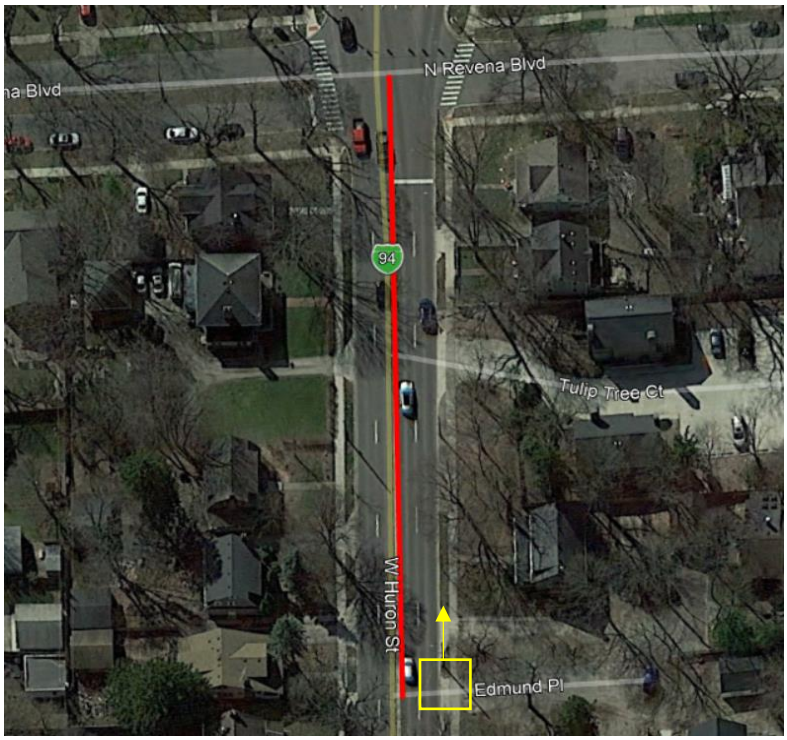
No	Segment Name	Road type	Bike facility	Land use	Owner	Lat	Long
2	Murfin@Plymouth	Local	Shared Lane Marking	Institutional	UoM	42.296415	-83.719707

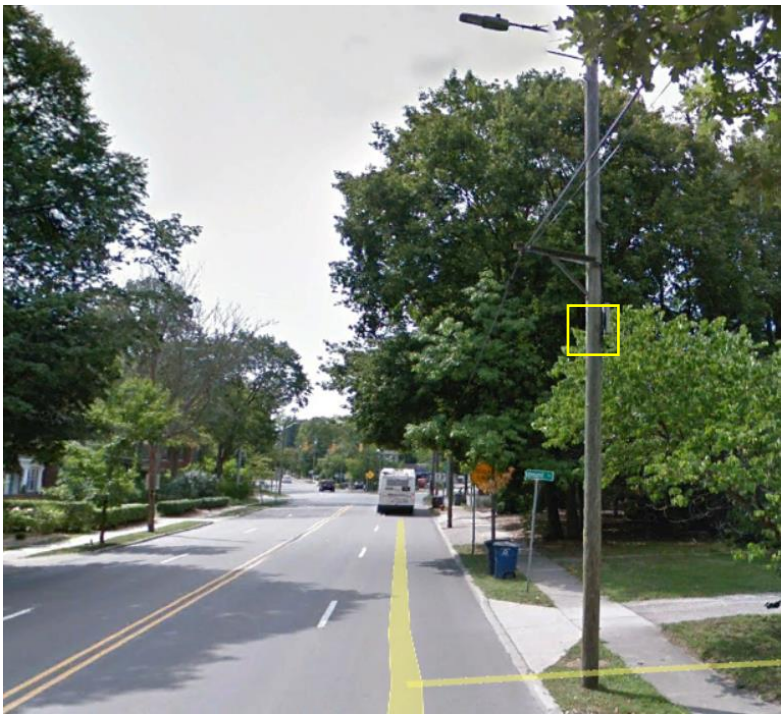




Video camera location: Light pole

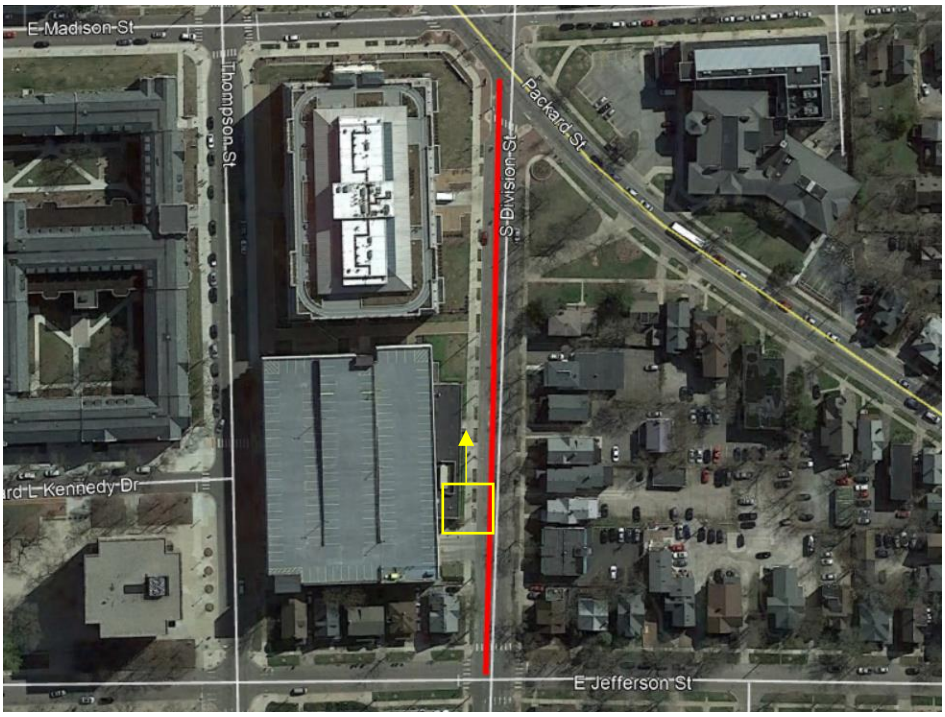
No	Segment Name	Road type	Bike facility	Land use	**Owner**	Lat	Long
3	Huron@Dexter	Arterial	None	Residential	**DTE**	42.281959	-83.765323

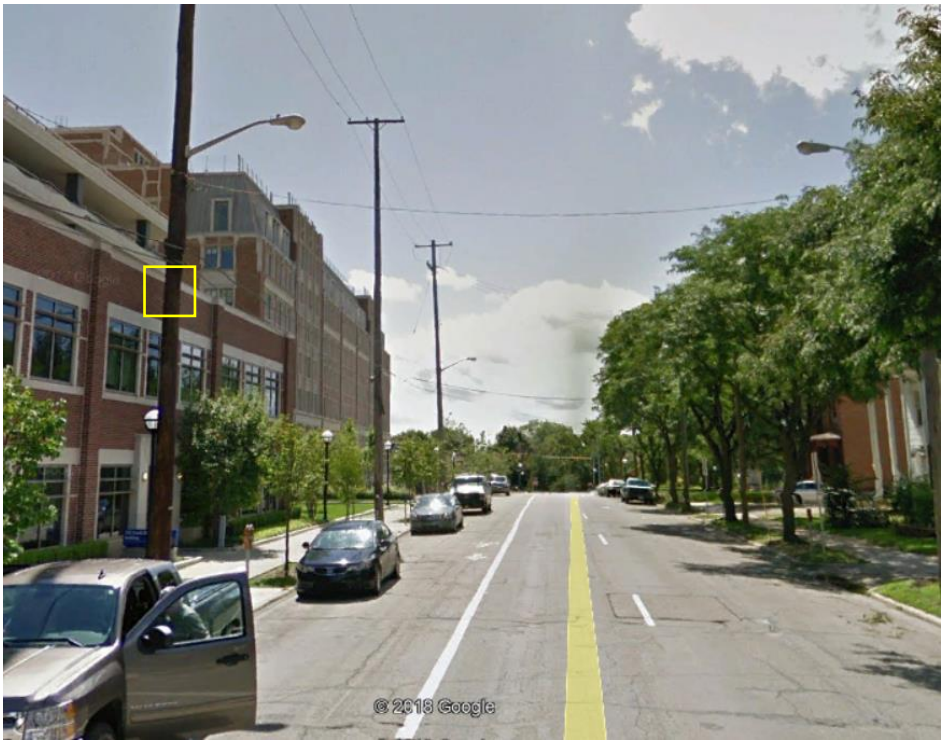




Video camera location: Electric pole

No	Segment Name	Road type	Bike facility	Land use	**Owner**	Lat	Long
4	Division@Packard	Arterial	Bike Lane	Institutional	**DTE**	42.275278	-83.744247

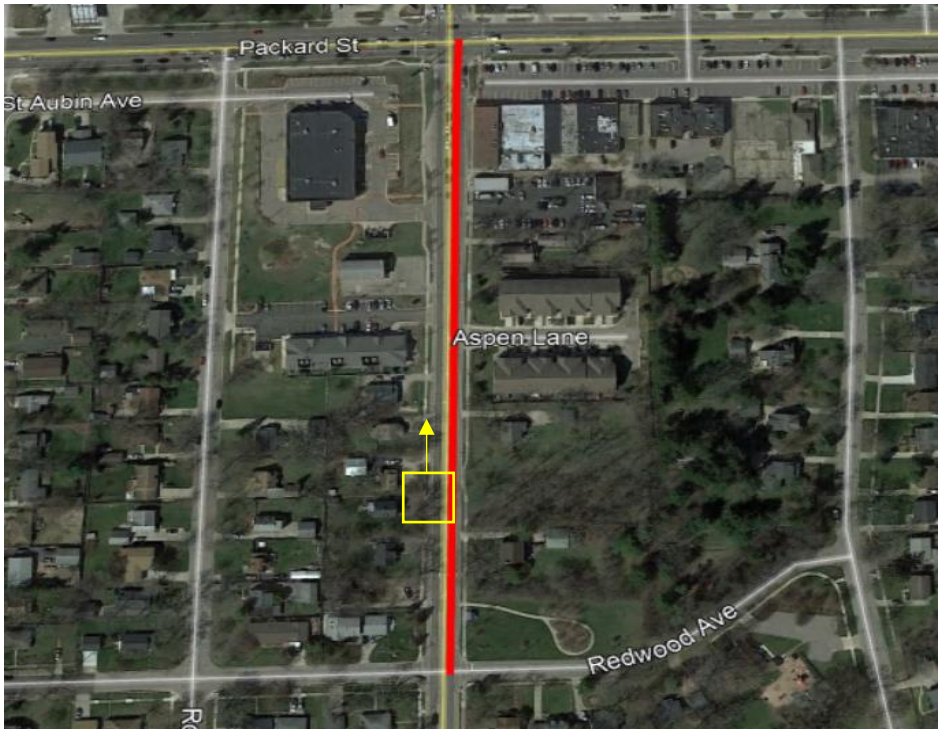




Video camera location: Electric pole

No	Segment Name	Road type	Bike facility	Land use	**Owner**	Lat	Long

5	Platt@Packard	Collector	Bike Lane	Residential	**DTE**	42.243612	-83.700060

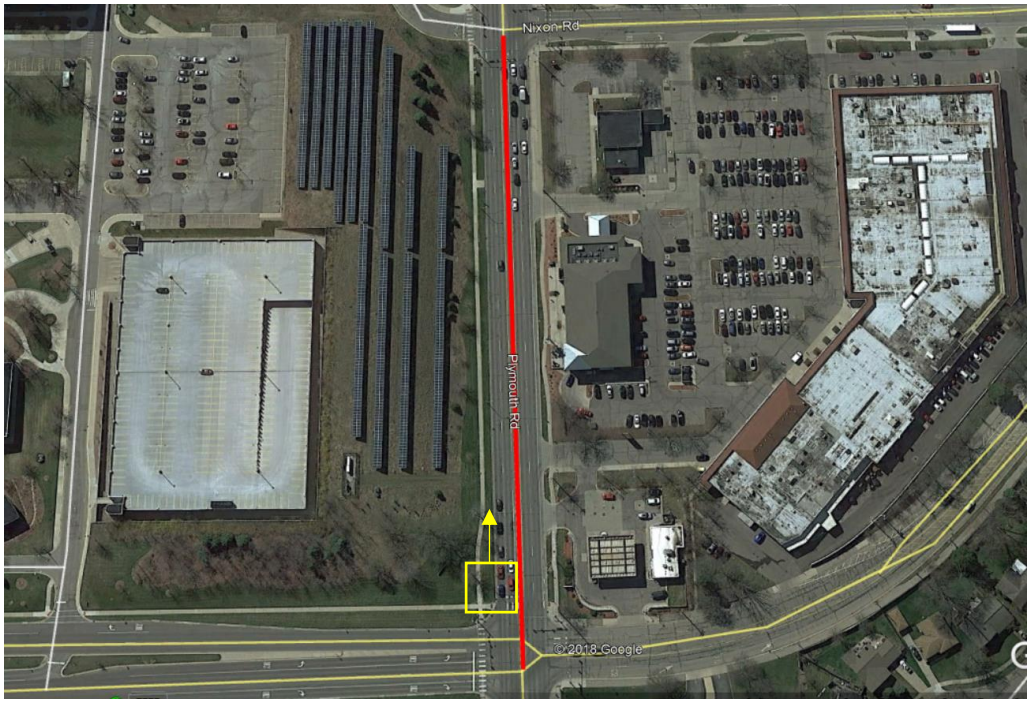


Video camera location: Electric pole

No	Segment Name	Road type	Bike facility	Land use	Owner	Lat	Long
6	Nixon@Green	Collector	Bike Lane	Residential	City	42.317163	-83.707602

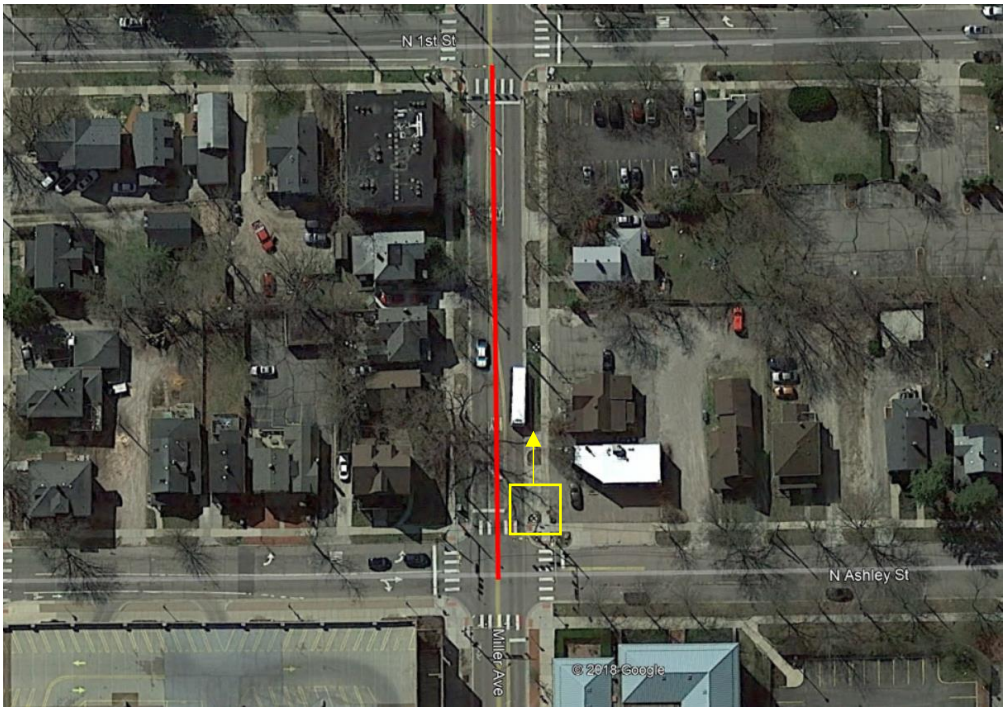
Video camera location: Light pole

No	Segment Name	Road type	Bike facility	Land use	**Owner**	Lat	Long
7	Plymouth@Huron Pkwy	Arterial	Bike Lane	Industrial	**DTE**	42.302561	-83.705764



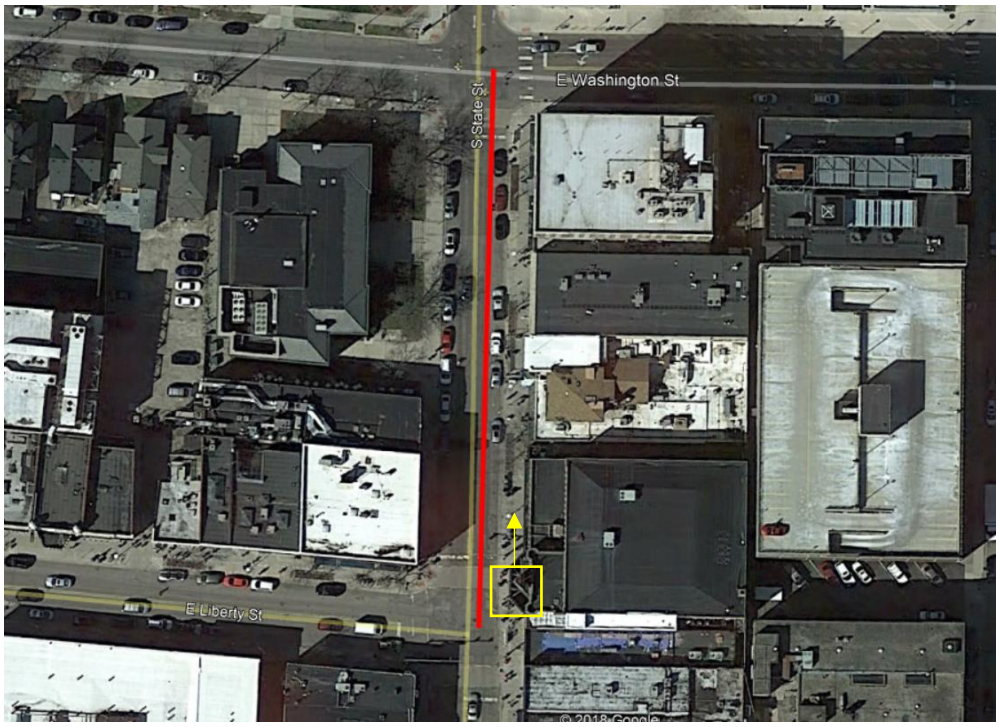
Video camera location: Light pole

No	Segment Name	Road type	Bike facility	Land use	**Owner**	Lat	Long
8	Miller@1st	Arterial	Shared Lane Marking	Residential	**DTE**	42.283312	-83.750237



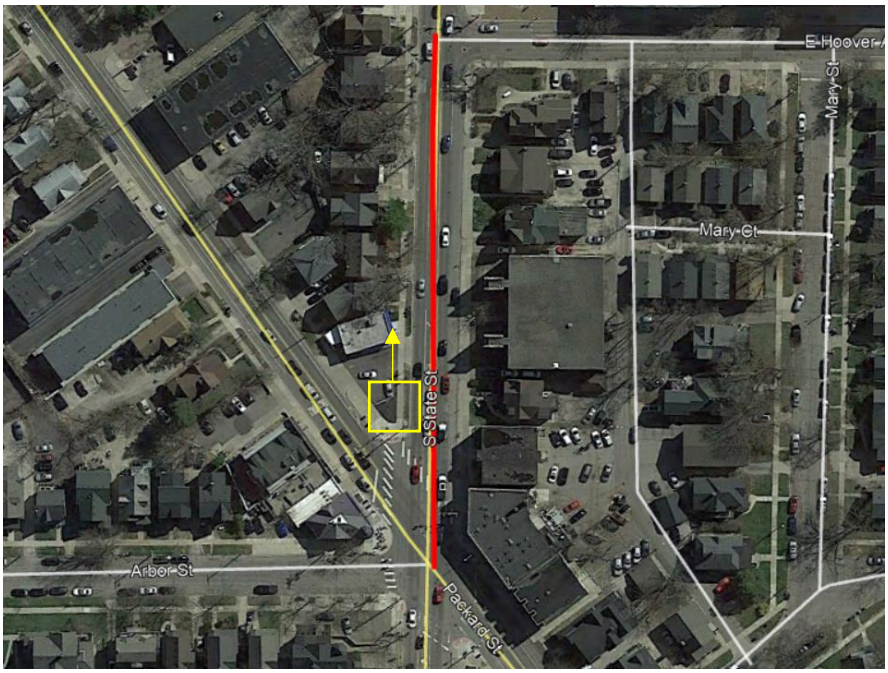
Video camera location: Electric pole

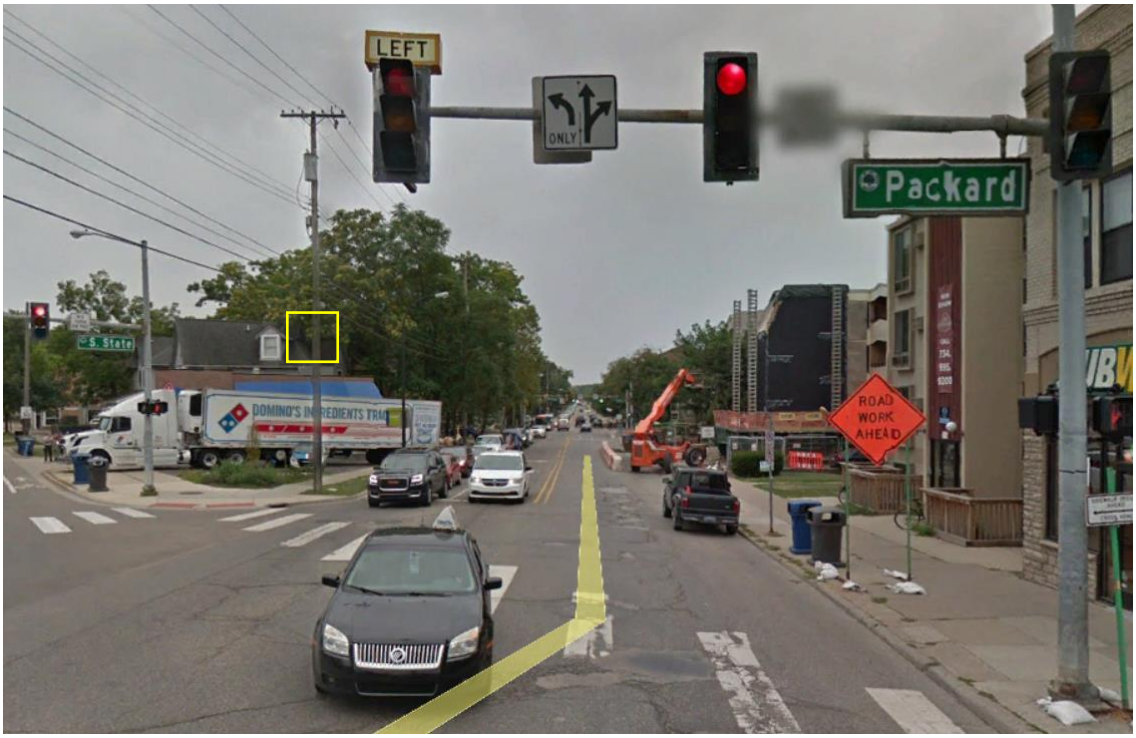
No	Segment Name	Road type	Bike facility	Land use	**Owner**	Lat	Long
9	State@Liberty	Arterial	Shared Lane Marking	Commercial	**City**	42.279812	-83.740804



Video camera location: Light pole

No	Segment Name	Road type	Bike facility	Land use	**Owner**	Lat	Long
10	State@Packard	Arterial	None	Commercial	**DTE**	42.270327	-83.740595





Video camera location: Electric pole

8.2 Field data collection: City of Grand Rapids

List of sites

No.	Road Name	Land use	Facility	Lat	Long
1	Cherry St SE	Commercial	Bike lane	42.9594	-85.6586
2	Grandville Ave SW	Commercial	Shared lane	42.9603	-85.6735
3	Lake Dr SE	Residential	None	42.9544	-85.6299
4	Monroe Ave NE	Residential	Bike lane	43.0103	-85.6665
5	N Park St NE	Institutional	Bike lane	43.0224	-85.6602
6	Walker Ave NW	Institutional	Bike lane	42.9832	-85.7021
7	White Pine Trail	Recreational	Trail	43.0026	-85.6708
8	Oxford Street Trail	Recreational	Trail	42.9559	-85.6861
8	Oxford Street Trail	Recreational	Trail	42.9531	-85.6896
9	Kent Trail	Recreational	Trail	42.9506	-85.7095

Physical addresses

Site. No	Camera Location	Address
1	Light pole	419-401 Cherry St SE, Grand Rapids, MI 49503
2	Light pole	135-165 Grandville Ave SW, Grand Rapids, MI 49503
3	Electric pole	1572-1634 Lake Dr. SE, East Grand Rapids, MI 49506
4	Light pole	2600-2698 Monroe Ave NE, Grand Rapids, MI 49505
5	Light pole	422-446 N Park St NE, Grand Rapids, MI 49525
6	Electric pole	1327-1399 Walker Ave NW, Grand Rapids, MI 49504
7	None(Tubes)	White Pine Trail, Grand Rapids, MI 49505
8	None(Tubes)	SWAN, Grand Rapids, MI 49504
9	None (Tubes)	Kent Trails, Grand Rapids, MI 49534

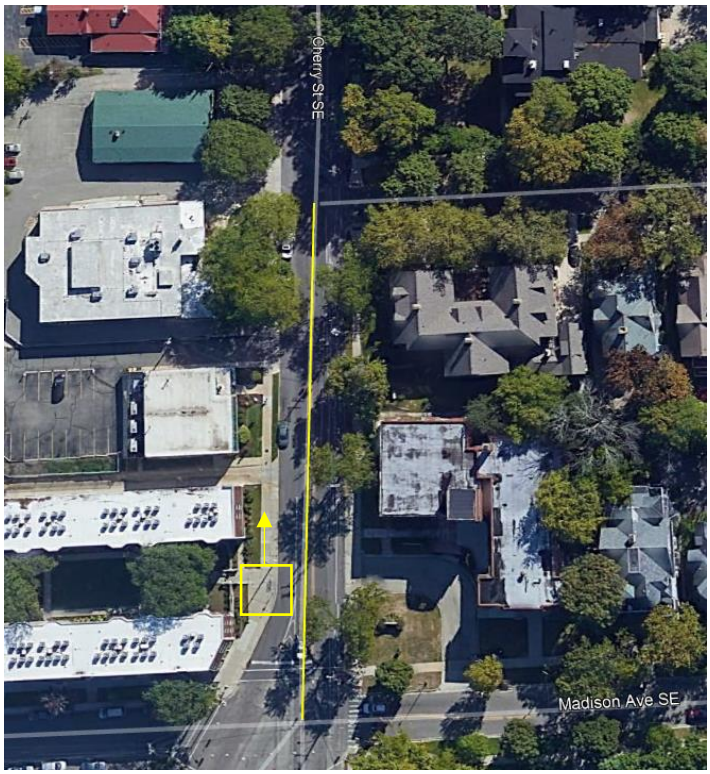
Schedule

No.	Road Name	Week 1	Week 2	Week 3
		7/16/18-7/22/18	7/23/18-7/29/18	7/30/18-8/5/18
1	Cherry St SE			
2	Grandville Ave SW			
3	Lake Dr SE			
4	Monroe Ave NE			
5	N Park St NE			
6	Walker Ave NW			
7	White Pine Trail			
8	Oxford Street Trail			

9	Kent Trail			

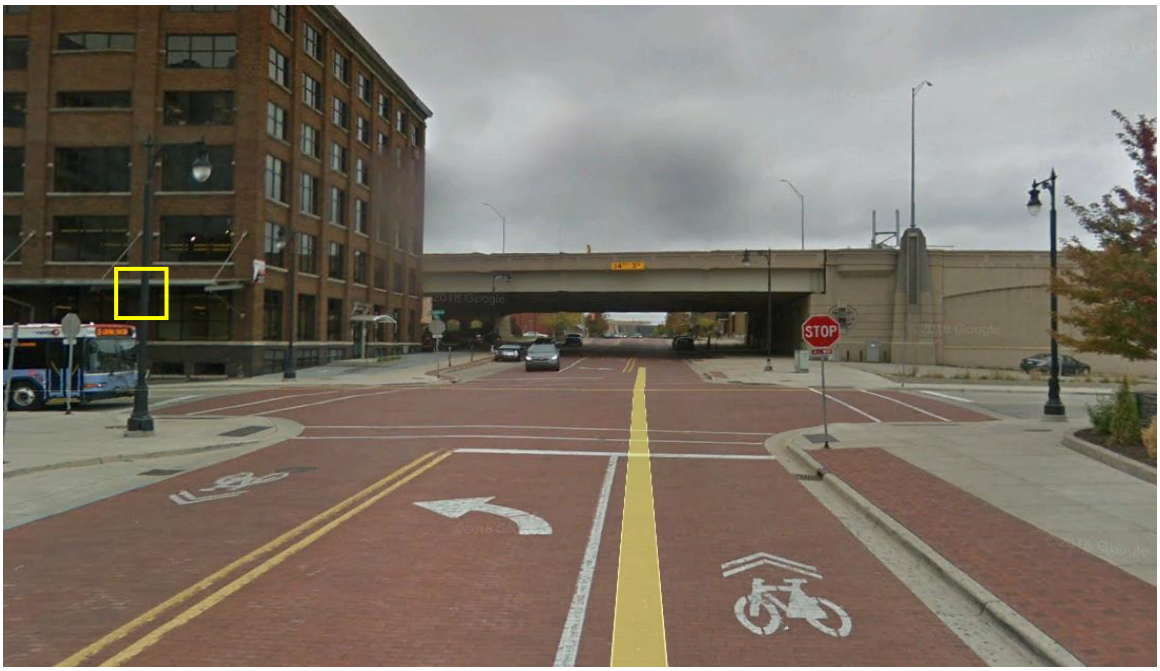
Details of each site

No.	Road Name	Land use	Facility	Lat	Long
1	Cherry St SE	Commercial	Bike lane	42.9594	-85.6586



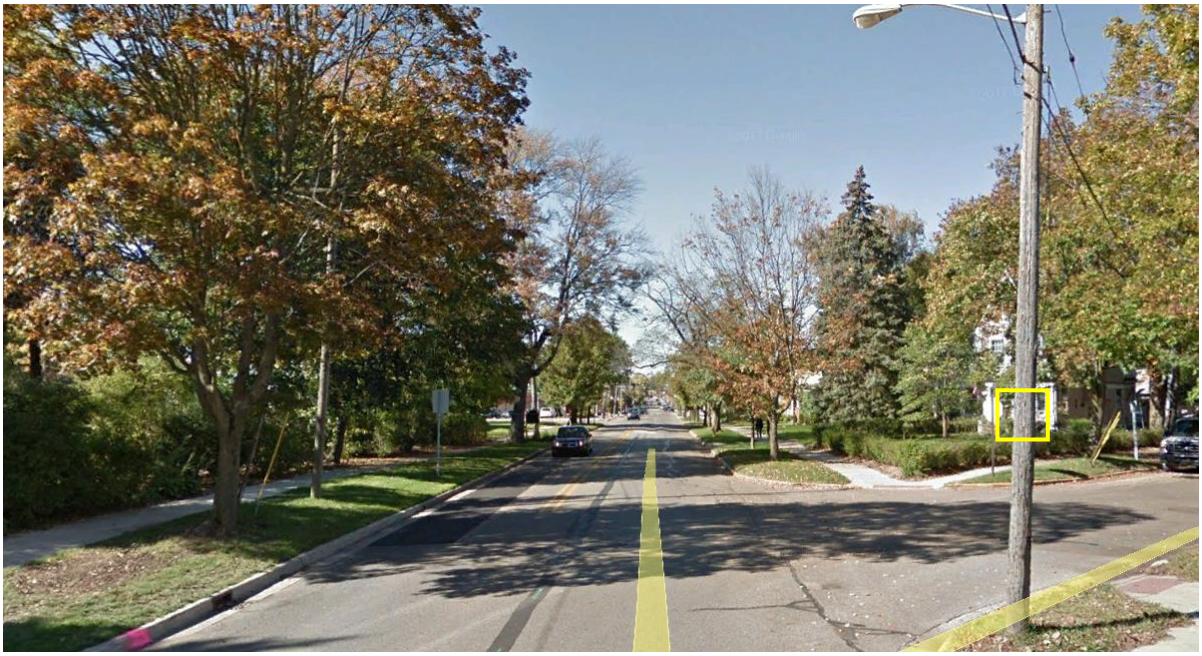
Video camera location: Light pole

No.	Road Name	Land use	Facility	Lat	Long
2	Grandville Ave SW	Commercial	Shared lane	42.9603	-85.6735



Video camera location: Light pole

No.	Road Name	Land use	Facility	Lat	Long
3	Lake Dr SE	Residential	None	42.9544	-85.6299



Video camera location: Electric pole

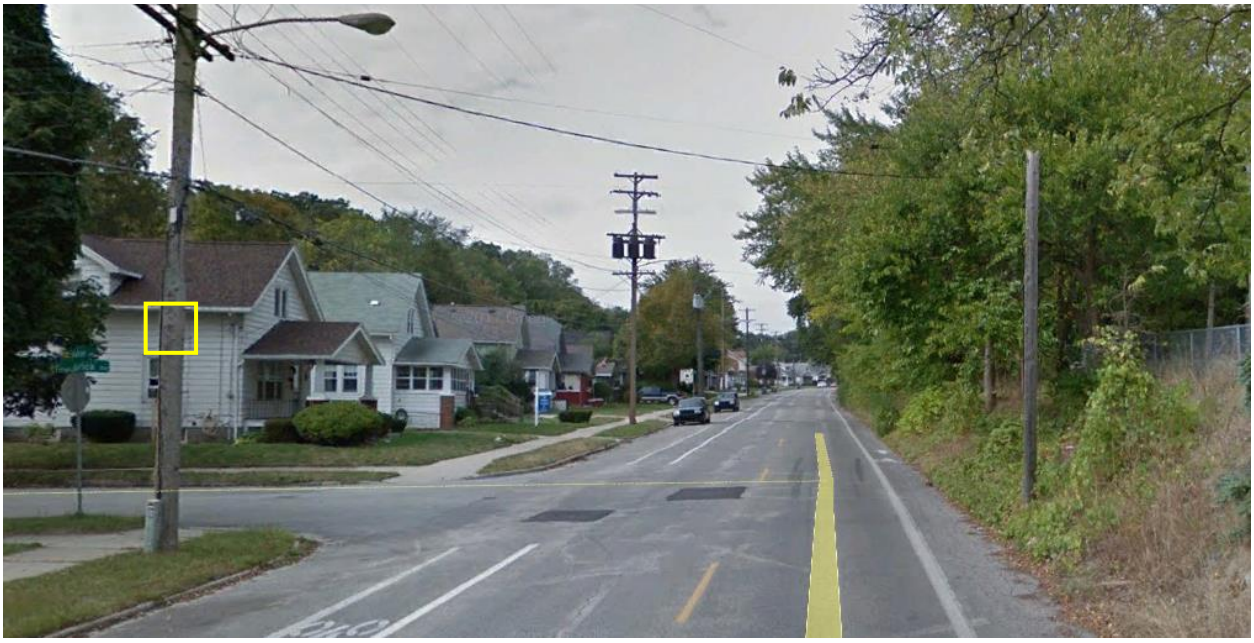
No.	Road Name	Land use	Facility	Lat	Long
4	Monroe Ave NE	Residential	Bike lane	43.0103	-85.6665

Video camera location: Light pole

No.	Road Name	Land use	Facility	Lat	Long
5	N Park St NE	Institutional	Bike lane	43.0224	-85.6602

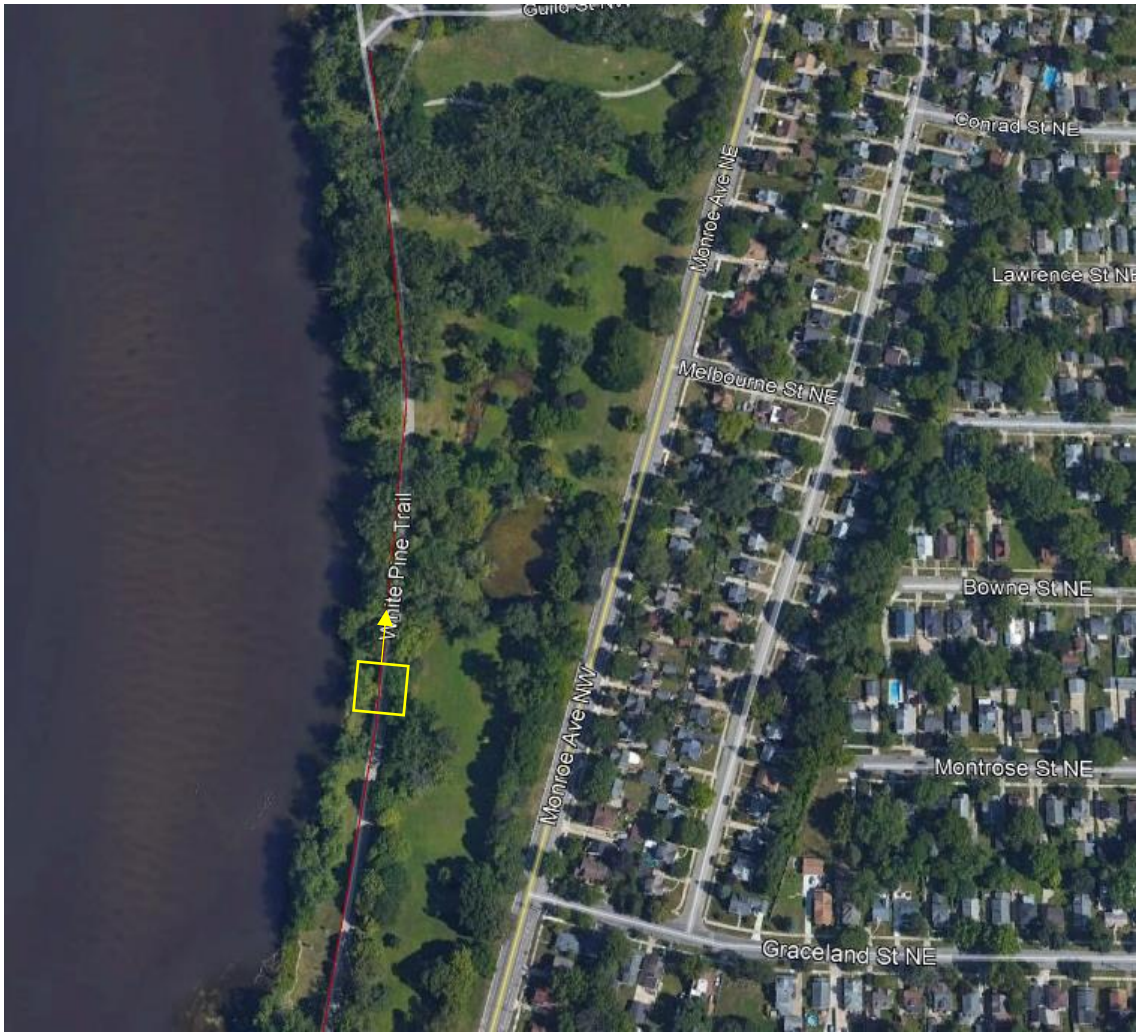
Video camera location: Light pole

No.	Road Name	Land use	Facility	Lat	Long
6	Walker Ave NW	Institutional	Bike lane	42.9832	-85.7021



Video camera location: Light pole

No.	Road Name	Land use	Facility	Lat	Long
7	White Pine Trail	Recreational	Trail	43.0026	-85.6708



Location: Bike pneumatic tube counters across the trail

No.	Road Name	Land use	Facility	Lat	Long
8	Oxford Street Trail	Recreational	Trail	42.9559	-85.6861

Location: Bike pneumatic tube counters across the trail

No.	Road Name	Land use	Facility	Lat	Long
9	Kent Trail	Recreational	Trail	42.9506	-85.7095

Location: Bike pneumatic tube counters across the trai

8.3 Hourly Bicycle Counts: City of Ann Arbor

Site	Weekday /Hour	5- 6 am	6- 7 am	7- 8 am	8- 9 am	9-10 am	10-11 am	11am-Noon	Noon-1 pm	1- 2 pm	2- 3 pm	3- 4 pm	4- 5 pm	5- 6 pm	6- 7 pm	7- 8 pm	8- 9 pm	9-10 pm	10-11 pm	Total
5th@Liberty	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Wed	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Thu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Fri	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sat	0	0	1	6	2	3	8	3	2	4	2	4	4	2	4	0	0	0	45
	Sun	0	1	0	1	3	2	2	6	5	1	4	7	6	3	0	0	0	0	41
Plymouth @ Murfins	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	2	5	0	1	2	5	9	5	6	5	7	9	1	0	57
	Wed	0	4	1	5	5	5	7	4	3	4	3	1	2	1	0	0	0	0	45
	Thu	0	0	0	0	7	1	0	2	2	1	1	5	5	2	4	5	1	0	36
	Fri	0	3	2	12	0	6	3	0	4	2	1	4	10	10	0	0	0	0	57
	Sat	0	0	0	0	0	9	6	3	6	5	2	10	8	5	6	1	2	1	64
	Sun	0	0	0	2	6	6	4	2	2	6	5	9	8	3	0	0	0	0	53
Huron@Dexter	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	5	3	6	3	2	1	5	0	1	1	4	3	1	0	35
	Wed	0	0	1	1	3	0	2	2	3	1	2	0	2	5	4	0	0	0	26
	Thu	0	0	0	0	1	1	1	0	2	1	3	1	3	1	4	0	0	0	18
	Fri	0	1	0	0	0	1	2	3	1	1	1	0	0	0	0	0	0	0	10
	Sat	0	0	0	0	0	1	2	2	5	4	1	2	1	2	1	5	0	0	26
	Sun	0	1	0	1	4	1	6	3	0	2	1	0	0	1	0	0	0	0	20

Site	Weekday /Hour	5- 6 am	6- 7 am	7- 8 am	8- 9 am	9-10 am	10-11 am	11am-Noon	Noon-1 pm	1- 2 pm	2- 3 pm	3- 4 pm	4- 5 pm	5- 6 pm	6- 7 pm	7- 8 pm	8- 9 pm	9-10 pm	10-11 pm	Total
Division@Packard	Mon	0	0	0	25	12	10	6	9	11	3	7	7	14	17	9	2	0	0	132
	Tue	0	5	12	16	20	3	8	10	4	4	3	10	14	7	5	0	0	0	121
	Wed	0	0	0	9	25	5	8	10	1	6	5	6	12	11	9	7	2	0	116
	Thu	0	3	8	24	10	14	13	12	12	4	11	12	14	7	0	0	0	0	144
	Fri	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sat	0	0	1	5	18	3	7	4	11	9	0	7	3	8	8	7	0	0	91
	Sun	0	4	4	13	3	2	4	1	7	3	1	5	3	0	0	0	0	0	50
Platt@Packard	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	0	0	3	5	2	3	2	5	3	6	1	4	1	0	35
	Wed	0	1	5	0	3	0	2	0	1	2	0	4	2	0	0	0	0	0	20
	Thu	0	0	0	0	1	0	0	0	2	2	0	4	0	4	0	3	0	0	16
	Fri	0	1	0	2	0	0	0	1	0	3	0	1	0	6	1	0	0	0	15
	Sat	0	0	0	0	0	0	4	0	4	1	0	0	3	0	6	1	1	0	20
	Sun	1	0	4	5	0	5	4	0	2	0	2	0	0	3	0	0	0	0	26
Nixon@Green	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	0	2	6	2	1	0	3	6	3	7	3	5	3	1	42
	Wed	0	0	0	2	1	0	0	0	0	0	0	0	0	0	4	0	0	0	7
	Thu	0	0	0	0	0	3	1	1	1	2	0	5	5	7	11	9	1	1	47
	Fri	0	0	0	3	2	3	0	3	0	2	1	1	0	2	3	0	0	0	20
	Sat	0	0	0	0	3	3	3	1	3	2	1	4	3	2	2	5	2	0	34
	Sun	0	3	7	5	2	6	0	4	2	3	1	5	5	2	0	0	0	0	45

Site	Weekday /Hour	5- 6 am	6- 7 am	7- 8 am	8- 9 am	9-10 am	10-11 am	11-12 pm	12-1 pm	1- 2 pm	2- 3 pm	3- 4 pm	4- 5 pm	5- 6 pm	6- 7 pm	7- 8 pm	8- 9 pm	9-10 pm	10-11 pm	Total
Plymouth@Huron Pkwy	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	1	0	2	0	0	0	0	1	6	1	1	2	0	0	14
	Wed	0	1	0	3	0	1	1	0	0	0	1	0	0	1	0	0	0	0	8
	Thu	0	0	0	0	6	2	2	1	2	5	1	1	1	5	0	3	0	0	29
	Fri	0	3	5	4	5	0	1	0	3	0	2	0	3	2	2	0	0	0	30
	Sat	0	0	0	0	0	0	0	1	1	1	0	0	1	1	0	3	1	0	9
	Sun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Miller@1st	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	27	15	2	10	7	21	13	15	12	15	11	15	6	0	169
	Wed	0	3	5	15	12	5	7	1	6	10	4	11	22	14	11	0	0	0	126
	Thu	0	0	4	10	28	9	10	11	15	10	10	17	29	19	19	16	12	0	219
	Fri	2	5	10	10	15	15	8	11	12	2	13	4	3	0	0	0	0	0	110
	Sat	0	0	0	0	9	16	20	13	16	3	13	6	2	5	6	8	4	0	121
	Sun	2	7	3	8	8	3	16	10	23	8	10	11	1	7	2	0	0	0	119
State@Liberty	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Wed	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Thu	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Fri	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sat	0	0	1	2	2	2	7	5	0	0	1	1	3	2	4	0	0	0	30
	Sun	0	0	0	3	4	1	3	6	5	11	2	8	3	1	0	0	0	0	47
State@Packard	Mon	0	0	0	0	8	5	9	7	14	12	3	10	6	7	9	6	4	0	100
	Tue	0	1	2	14	5	7	7	3	7	5	4	10	5	8	3	0	0	0	81
	Wed	0	0	0	0	1	4	10	9	5	2	2	1	7	11	5	8	0	0	65
	Thu	0	2	3	12	8	8	10	5	2	3	6	8	11	3	0	0	0	0	81
	Fri	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Sat	0	0	0	1	0	3	8	0	5	8	3	6	0	3	3	2	3	0	45
	Sun	0	0	1	4	8	3	1	3	6	5	9	1	1	2	0	0	0	0	44

8.4 Hourly Bicycle Counts: City Grand Rapids

Site	Weekday /Hour	5- 6 am	6- 7 am	7- 8 am	8- 9 am	9-10 am	10-11 am	11am-Noon	Noon-1 pm	1- 2 pm	2- 3 pm	3- 4 pm	4- 5 pm	5- 6 pm	6- 7 pm	7- 8 pm	8- 9 pm	9-10 pm	10-11 pm	Total
Cherry St SE	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	0	9	4	5	8	9	3	7	12	17	13	10	3	0	100
	Wed	0	10	5	6	3	2	6	8	11	5	3	15	12	15	15	3	0	0	119
	Thu	0	0	0	0	0	4	3	15	5	18	7	10	9	13	11	6	12	6	119
	Fri	0	2	4	1	1	4	3	7	4	10	6	13	8	2	0	0	0	0	65
	Sat	0	0	0	0	0	1	3	2	7	3	3	9	7	2	3	1	4	0	45
	Sun	0	1	1	3	0	6	4	4	5	5	7	13	3	7	9	4	2	0	74
Grandville Ave SW	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	0	2	1	5	4	3	3	7	3	4	5	9	6	2	54
	Wed	0	2	4	1	3	0	6	9	5	5	6	6	5	10	4	4	0	0	70
	Thu	0	0	0	0	1	7	0	3	6	5	3	3	1	11	4	2	1	0	47
	Fri	1	2	0	1	2	3	5	1	4	2	8	6	4	0	0	0	0	0	39
	Sat	0	0	0	0	0	1	1	4	7	8	8	7	4	2	0	1	0	0	43
	Sun	0	0	1	2	1	2	2	3	6	11	9	8	8	9	6	4	3	0	75
Lake Dr SE	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	6	3	0	4	8	4	7	7	9	21	11	10	10	0	0	100
	Wed	0	4	7	8	2	4	2	10	8	7	4	9	18	6	0	0	0	0	89
	Thu	0	0	0	0	4	5	4	6	12	6	6	8	5	11	5	5	4	0	81
	Fri	0	5	13	1	1	8	5	2	8	9	10	7	8	0	0	0	0	0	77
	Sat	0	0	0	0	1	5	8	6	12	9	7	9	6	8	8	9	5	0	93
	Sun	0	1	4	0	5	8	1	11	3	8	10	3	1	7	4	1	0	0	67

Site	Weekday /Hour	5- 6 am	6- 7 am	7- 8 am	8- 9 am	9-10 am	10-11 am	11am-Noon	Noon-1 pm	1- 2 pm	2- 3 pm	3- 4 pm	4- 5 pm	5- 6 pm	6- 7 pm	7- 8 pm	8- 9 pm	9-10 pm	10-11 pm	Total
Monroe Ave NE	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	0	4	11	7	3	13	0	0	9	20	0	26	13	1	107
	Wed	0	4	8	6	9	19	8	5	12	8	11	19	14	25	15	19	0	0	182
	Thu	0	0	0	0	0	3	4	12	13	4	10	12	18	16	5	0	12	0	109
	Fri	0	3	6	7	6	5	7	4	8	14	6	27	9	1	0	0	0	0	103
	Sat	0	0	0	0	3	0	22	25	6	0	19	18	21	21	11	5	7	1	159
	Sun	0	1	3	11	13	23	18	10	19	16	21	36	17	16	25	0	2	0	231
N Park St NE	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	0	0	6	24	25	13	17	0	0	36	46	0	34	17	0	218
	Wed	0	12	10	8	31	35	36	28	23	16	27	33	51	55	45	42	0	0	452
	Thu	0	0	0	0	0	11	31	25	18	19	22	25	45	44	50	0	21	0	311
	Fri	0	12	10	11	11	14	27	21	30	28	24	43	33	3	0	0	0	0	267
	Sat	0	0	0	0	0	0	112	60	69	0	81	68	73	40	32	17	14	0	566
	Sun	0	4	9	11	29	33	44	33	70	65	54	66	40	49	45	0	8	0	560
Walker Ave NW	Mon	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	3	1	4	4	2	5	2	5	3	3	13	7	4	0	0	56
	Wed	0	4	3	3	2	5	1	3	7	3	0	4	8	4	0	0	0	0	47
	Thu	0	0	0	0	2	2	1	3	3	4	2	1	4	9	1	5	0	0	37
	Fri	0	0	3	5	1	3	3	5	5	3	8	7	5	1	0	0	0	0	49
	Sat	0	0	0	0	0	2	9	5	3	4	2	5	4	1	3	0	1	1	40
	Sun	0	1	1	1	2	5	9	2	8	9	4	1	6	1	5	2	1	0	58

Site	Weekday /Hour	5- 6 am	6- 7 am	7- 8 am	8- 9 am	9-10 am	10-11 am	11am-Noon	Noon-1 pm	1- 2 pm	2- 3 pm	3- 4 pm	4- 5 pm	5- 6 pm	6- 7 pm	7- 8 pm	8- 9 pm	9-10 pm	10-11 pm	Total
White Pine Trail	*Mon*	0	0	0	0	0	0	0	0	0	0	4	13	2	1	0	4	0	1	25
	Tue	0	0	1	4	0	15	17	22	13	22	4	19	24	33	35	0	12	2	223
	Wed	2	0	14	6	9	0	12	19	0	13	9	0	36	42	34	35	13	0	244
	Thu	1	6	9	7	12	8	23	8	12	9	0	23	18	33	27	24	10	2	232
	Fri	0	4	3	1	14	2	0	15	0	0	7	7	11	1	4	0	1	0	70
	Sat	0	0	4	6	7	0	0	7	15	12	16	10	5	11	2	5	1	0	101
	Sun	0	1	5	2	3	6	4	10	16	20	30	16	23	39	29	19	5	0	228
Oxford Street Trail	*Mon*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Tue	0	0	0	2	1	5	0	1	4	4	1	4	5	9	6	5	0	0	47
	Wed	0	0	1	3	3	2	1	0	2	4	3	3	9	5	12	5	0	0	53
	Thu	0	0	1	2	1	3	5	4	2	3	4	1	5	12	4	16	1	0	64
	Fri	0	0	1	0	3	3	3	5	6	4	0	1	7	3	5	7	5	0	53
	Sat	0	0	1	2	9	1	4	5	1	6	2	0	1	2	4	6	0	0	44
	Sun	0	0	0	3	5	8	7	8	2	7	3	9	2	2	1	0	1	0	58
Kent Trail	*Mon*	2	5	11	5	17	7	9	7	10	8	12	3	5	12	2	6	2	1	124
	Tue	0	8	12	13	9	17	27	25	10	14	21	47	47	44	51	22	12	0	379
	Wed	0	3	22	11	11	23	30	26	19	32	28	24	32	57	60	38	2	1	419
	Thu	0	7	13	14	8	16	22	18	11	27	31	26	31	48	44	21	6	1	344
	Fri	0	7	18	11	14	19	12	15	11	18	24	18	17	19	25	26	8	2	264
	Sat	0	5	16	30	29	57	55	54	44	46	52	42	26	46	31	12	9	1	555
	Sun	0	2	3	13	28	41	70	59	49	55	56	53	40	57	31	16	8	0	581

8.5 Survey of cyclists at bicycle parking areas (racks and hoops)

Survey location	

Survey duration: (1-2minutes)

Goal of the survey

This survey is conducted by Western Michigan university to gather the following information

(1) Demographics and cycling behavior of the cyclists
(2) Characteristics of the trip(s) made by the cyclists
(3) The proportion of cyclists using the fitness and health apps to track their cycling activities

A. Demographic and cycling behavior

1. Age
   - ☐ <16
   - ☐ 16-24
   - ☐ 25-34
   - ☐ 35-44
   - ☐ 45-54
   - ☐ 55-64
   - ☐ >65
2. Sex
   - ☐ Male
   - ☐ Female
   - ☐ Prefer not to say
3. How do you describe your biking skills level?
   - ☐ Beginner
   - ☐ Intermediate
   - ☐ Expert
4. Do you use STRAVA app to keep track of your cycling activities?
   - ☐ Yes
   - ☐ No
5. Do you use any other fitness app(s) to keep track of your cycling activities?
   - ☐ Yes
   - ☐ No
6. How often do you bike?
   - ☐ Always (*Daily*)
   - ☐ Often (*Few times a week*)
   - ☐ Sometimes (*Several times a month)*
   - ☐ Seldom (*Few times a year*)

B. Characteristic of the trip (s) a cyclist made/about to make

7. What is/was the purpose of your trip?

- ☐ Commute to/from work or school
- ☐ Recreational
- ☐ An errand

8. What are the bicycle facilities available along the route for your trip?
   - ☐ Bike lane
   - ☐ Shared lane
   - ☐ Trail
   - ☐ Paved shoulder
   - ☐ Sidewalk

9. How will you rate the quality of bicycle facility available for your trip?

Poor	Fair	Good	Excellent

10. When traveling a segment with a bicycle facility such as bike lane, shared lane, or paved shoulder, would you prefer riding on a sidewalk if available?
    - ☐ Never
    - ☐ Only if the roadway is busy or it is high speed road
    - ☐ Always
    - ☐ No preference

8.6 Tool for estimating hourly bicycle volume

Location: https://trclc.shinyapps.io/BikeExposure/

This interactive tool estimates hourly bicycle volume at a given roadway segment using Random Forest(RF) model. The user needs to download a template file and fill in the required information. Once the data is uploaded, the user can view results by clicking the “Table” or “Visualization” tab. The estimates can then be downloaded.

The data needed are as follows:

VARIABLE	DESCRIPTION	INSTRUCTIONS
site	Site identification number	Fill-in the site identification number. It should be an integer
strava	Hourly Strava Count	Fill-in the Strava counts for your hour of interest. It should be an integer
humidy	Average Hourly Relative humidity (%)	Fill-in the humidity for your hour of interest. It should range from 0 to 100
hour_adjAM.hrs..6am.9.59am	Hourly Adjustment Factor: 6am-9: 59am	Put 1 if your hour of interest falls within this time of the day otherwise 0
hour_adjEvening.hrs..8pm.11.59pm	Hourly Adjustment Factor: 8pm-11:59pm	Put 1 if your hour of interest falls within this time of the day otherwise 0
hour_adjMid.Day.hrs..10am.2.59pm	Hourly Adjustment Factor: 10am-2:59pm	Put 1 if your hour of interest falls within this time of the day otherwise 0
hour_adjPeak.PM.hrs..3pm.7.59pm	Hourly Adjustment Factor: 3:00pm-7:59	Put 1 if your hour of interest falls within this time of the day otherwise 0
bikefacilityBike.Lane	Bike Facility: Bike Lane	Put 1 if the road segment has bike lane otherwise 0
bikefacilityShared.Lane	Bike Facility: Shared Lane	Put 1 if the road segment has shared lane otherwise 0
bikefacilityTrail	Bike Facility: Trail	Put 1 if it is a trail facility otherwise 0
landuseIndustrial	Landuse: Industrial	Put 1 if the road segment passes mainly through the industrial area otherwise 0
landuseInstitutional	Landuse: Institutional	Put 1 if the road segment passes mainly through the

VARIABLE	DESCRIPTION	INSTRUCTIONS
		institutional(campus) area otherwise 0
landuseResidential	Landuse: Residential	Put 1 if the road segment passes mainly through the residential area otherwise 0
pro.male	% of males in the population in the census block	Fill-in the percentage of males in a population at a given census block. It should range from 0 to 100
pro.white	% of white in the census block population	Fill-in the percentage of white race in a population at a given census block. It should range from 0 to 100
pro.bike	% of bikers (workers) in the census block population	Fill-in the percent of bikers commuting to work in a population at a given census block. It should range from 0 to 100