Exploring Crowdsourced Monitoring Data for Safety

March 2020 | Final Report

SAFE-D
SAFETY THROUGH DISRUPTION

Texas A&M Transportation Institute

San Diego State University
Leadership Starts Here

VT | TRANSPORTATION INSTITUTE
VIRGINIA TECH

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. TTI-Student-05	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Exploring Crowdsourced Monitoring Data for Safety		5. Report Date March 2020 6. Performing Organization Code:
7. Author(s) Shawn Turner Michael Martin Greg Griffin Minh Le Subasish Das Ruihong Wang Bahar Dadashova Xiao Li		8. Performing Organization Report No. Report TTI-Student-05
9. Performing Organization Name and Address: Safe-D National UTC Texas A&M Transportation Institute		10. Work Unit No. 11. Contract or Grant No. 69A3551747115/TTI-Student-05
12. Sponsoring Agency Name and Address Office of the Secretary of Transportation (OST) U.S. Department of Transportation (US DOT) State of Texas		13. Type of Report and Period Final Research Report 14. Sponsoring Agency Code

15. Supplementary Notes
This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program, and, in part, with general revenue funds from the State of Texas.

16. Abstract
This project included four distinct but related exploratory studies of data sources that could improve roadway safety analysis. The first effort evaluated passively gathered crowdsourced bicyclist activity data from StreetLight Data and found promising correlations (R^2 of 62% and 69% for monthly weekday and weekend daily averages) when the StreetLight data were compared to bicyclist counts from 32 locations in eight Texas cities, and even better correlation (R^2 of 94%) when compared with countywide Strava data expanded to represent total bicycling activity. The second effort evaluated the pedestrian counting accuracy of the Miovision system and found 15% error for daytime and 24% error for nighttime conditions. The third effort used INRIX trip trace data to determine origin-destination patterns and developed 40 decision rules to define the origin-destination patterns. The fourth effort analyzed crowdsourced Waze data (i.e., traffic incidents) and found it to be a reliable alternative to observed and predicted crashes, with the ability to identify high-risk locations: 77% of high-risk locations identified from police-reported crashes were also identified as high-risk in Waze data. The researchers propose a method to treat the redundant Waze reports and to match the unique Waze incidents with police crash reports.

17. Key Words Bicyclist counts, passive crowdsourced, pedestrian crossing, massive GPS data, origin-destination, crowdsourced incident data	18. Distribution Statement No restrictions. This document is available to the public through the Safe-D National UTC website, as well as the following repositories: VTechWorks, The National Transportation Library, The Transportation Library, Volpe National Transportation Systems Center, Federal Highway Administration Research Library, and the National Technical Reports Library.

19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 38	22. Price $0

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

Abstract

This project included four distinct but related exploratory studies of data sources that could improve roadway safety analysis. The first effort evaluated passively gathered crowdsourced bicyclist activity data from StreetLight Data and found promising correlations (R^2 of 62% and 69% for monthly weekday and weekend daily averages) when the StreetLight data were compared to bicyclist counts from 32 locations in eight Texas cities, and even better correlation (R^2 of 94%) when compared with countywide Strava data expanded to represent total bicycling activity. The second effort evaluated the pedestrian counting accuracy of the Miovision system and found 15% error for daytime and 24% error for nighttime conditions. The third effort used INRIX trip trace data to determine origin-destination patterns and developed 40 decision rules to define the origin-destination patterns. The fourth effort analyzed crowdsourced Waze data (i.e., traffic incidents) and found it to be a reliable alternative to observed and predicted crashes, with the ability to identify high-risk locations: 77% of high-risk locations identified from police-reported crashes were also identified as high-risk in Waze data. The researchers propose a method to treat the redundant Waze reports and to match the unique Waze incidents with police crash reports.

Acknowledgements

The authors acknowledge the following companies that participated in this project by providing data or access to their products:

- *StreetLight Data, Inc.*
- *Miovision*
- *INRIX*
- *Waze*

The authors gratefully acknowledge the review of this research report by Dr. Ipek Sener, Texas A&M Transportation Institute.

This project was funded by the Safety through Disruption (Safe-D) National University Transportation Center, a grant from the U.S. Department of Transportation – Office of the Assistant Secretary for Research and Technology, University Transportation Centers Program, and, in part, with general revenue funds from the State of Texas.

Table of Contents

List of Figures

List of Tables

Introduction

The "Exploring Crowdsourced Monitoring Data for Safety" project included four distinct but related exploratory studies of emerging data sources that could possibly improve roadway safety and exposure analysis:

1. Crowdsourced mobile device data to better quantify citywide bicycling activity.
2. Traffic signal system data to quantify pedestrian street crossings.
3. Crowdsourced GPS waypoint data to better quantify motor vehicle travel patterns.
4. Crowdsourced traffic incident data to better quantify overall motor vehicle crash risk.

This report summarizes the results of these four unique exploratory studies, each separately documented in the following sections.

Citywide Bicycling Activity from Crowdsourced Data

Some U.S. cities are spending millions of dollars to make bicycling a safe, convenient, and realistic transportation option for everyday trips. Yet, it is still a challenge to measure the impacts of these bicycling improvements at a citywide level. In particular, it is very difficult to accurately measure bicyclist trips and travel for large geographic areas. Comprehensive analyses of bicyclist safety rely on bicyclist exposure (i.e., trips and travel) to determine if the risk to bicyclists has been reduced.

Various sensor technologies (e.g., inductance loops, video machine vision) can be used to monitor bicyclist activity at specific locations, but it is not realistic or cost-feasible to install these sensors everywhere in a city. Some agencies use results from the American Community Survey (ACS) and National Household Travel Survey (NHTS) to quantify citywide bicycling activity, but both of these national surveys have major limitations. The ACS captures bicycling only when it is the primary mode of commuting, whereas the NHTS is intended for national statistics and has insufficient travel samples at a citywide level.

Some researchers and practitioners are exploring crowdsourced data to quantify bicycling activity. For example, Strava is a fitness tracking app and Web platform that sells aggregate activity data to government agencies via its Strava Metro division. However, several research efforts [1-5] have identified a recreational bias in the Strava Metro data. Also, bicyclists must initiate Strava tracking when a ride begins (called active sensing, as it requires an action by the user to begin tracking).

This subtask of the research project explored the use of passively gathered crowdsourced data as a possible means to quantify citywide bicycling activity. The bicycling activity data were obtained

for evaluation from StreetLight Data, a Big Data analytics company that processes trillions of geospatial data points to measure how pedestrians, bikes, and motor vehicles travel.

Evaluation Methods

Researchers obtained estimates of bicycling activity from StreetLight Data and compared them to two other existing estimates of bicycling activity:

1. Permanent bicycle counters installed at 32 locations
2. Expanded/adjusted estimates of total bicycling activity derived from Strava Metro [6].

For the 32 bicycle counter locations, the count data had been reviewed for quality and consistency in a previous project [7]. Therefore, the permanent count data were considered to be a suitable comparison benchmark. The expanded estimates of total bicycling activity derived from Strava Metro data do have error, but these comparisons were made at an aggregate (i.e., county) level. Even at a countywide level, the expanded Strava estimates are considered to be status quo measurements with known limitations and not a ground truth or authoritative benchmark. The estimates of bicycling activity obtained from StreetLight Data, called StreetLight Index values, represent a sample of bicycle trips starting in, passing through, or ending in defined zones.

Results

Comparisons to Permanent Counter Data

Permanent bicycle counts from 32 locations in eight cities across Texas were used to calculate the monthly weekday and weekend average bicycles per count location. A complete, uninterrupted set of 15-minute counts per day (96 total) for each month was required. If a particular month was missing 15-minute count data from any time period, then it was deemed incomplete and excluded from the analysis. Due to data availability, counts from April, May, and June were compared with StreetLight Index values. Table 1 provides the breakdown of counts by city and month. StreetLight Data provided unexpanded Index values based on location-based services (LBS) passive data for an average weekday (Monday–Thursday) and weekend broken out by month.

Table 1. Number of Bicycle Count Locations per City and Month

City	April	May	June
Allen	0	1	1
Austin	6	6	0
Dallas	10	11	0
Fort Worth	2	3	1
Houston	3	3	1
North Richland Hills	1	0	0
Plano	2	2	1
San Antonio	2	2	0

Figure 1 provides a comparison of the permanent bicycle counts to unexpanded StreetLight Index values. The correlations for both weekdays and weekends are moderately high, with R^2 values of 62% and 69%, respectively. This indicates that expanded StreetLight Index values are capable of estimating bicycle activity at specific locations with moderate accuracy despite the variance in the underlying LBS data sample sizes used to create the Index values.

Figure 1. Graph. Comparison of permanent bicycle counts to StreetLight Index values for May.

To determine the accuracy, average weekday and weekend expansion factors were developed using April data and applied to the StreetLight Index values for May. Accuracy is measured in terms of the mean absolute percentage error (MAPE) and mean absolute error (MAE). The percentage error in the expanded StreetLight Index values ranged from 21% to 76% for weekdays and 23% to 53% for weekends between cities, as shown in Table 2. The MAE shows that although the relative percentage values (i.e., MAPE) may be high in some cases, the absolute error in low-magnitude counts can be relatively small. The results illustrate again that StreetLight Index values have the potential to be expanded to estimate bicycle activity with moderate levels of accuracy. However, more count locations and a stratified expansion process could improve the error results by controlling for functional class and area type.

Table 2. Average Weekday and Weekend StreetLight Index Expansion Error Estimates by City

Average Type	City	Expanded StreetLight Index MAPE	Expanded StreetLight Index MAE
Weekday	**Austin**	76%	277
Weekday	**Dallas**	70%	410
Weekday	**Fort Worth**	21%	18
Weekday	**Houston**	55%	43
Weekday	**Plano**	38%	84
Weekday	**San Antonio**	47%	41
Weekend	**Austin**	26%	104
Weekend	**Dallas**	33%	311
Weekend	**Fort Worth**	57%	91
Weekend	**Houston**	553%	344
Weekend	**Plano**	23%	82
Weekend	**San Antonio**	69%	59

Comparisons to Expanded Strava Data

In this comparison, the Strava 2017 average daily bike-miles of travel (BMT) was calculated by multiplying the expanded Strava 2017 average annual daily bicyclist traffic by the corresponding link length, and then summed for each of the 254 counties in Texas. ("Expanded" means total bicyclist estimates derived from Strava bicyclist samples.) This countywide BMT value was compared to the StreetLight Index (which represents unexpanded bicyclist activity) for each of the 254 counties. In this case, the researchers did not expand the StreetLight Index because we wanted to test the relative correlations in the absence of any expansion process.

Figure 2 shows the comparison of countywide expanded Strava bicycling estimates to StreetLight Index values. The 25 counties with the most bicycling activity are shown here, since many of the less-populated counties in Texas have negligible bicycling activity. The correlation between the two measures of bicycling activity is quite high, with an R^2 value of 94%. The high level of correlation means that both methods of estimating bicycling activity are numerically similar, and that a simplified expansion process for the StreetLight Index value may be sufficient at aggregate levels (such as countywide or citywide).

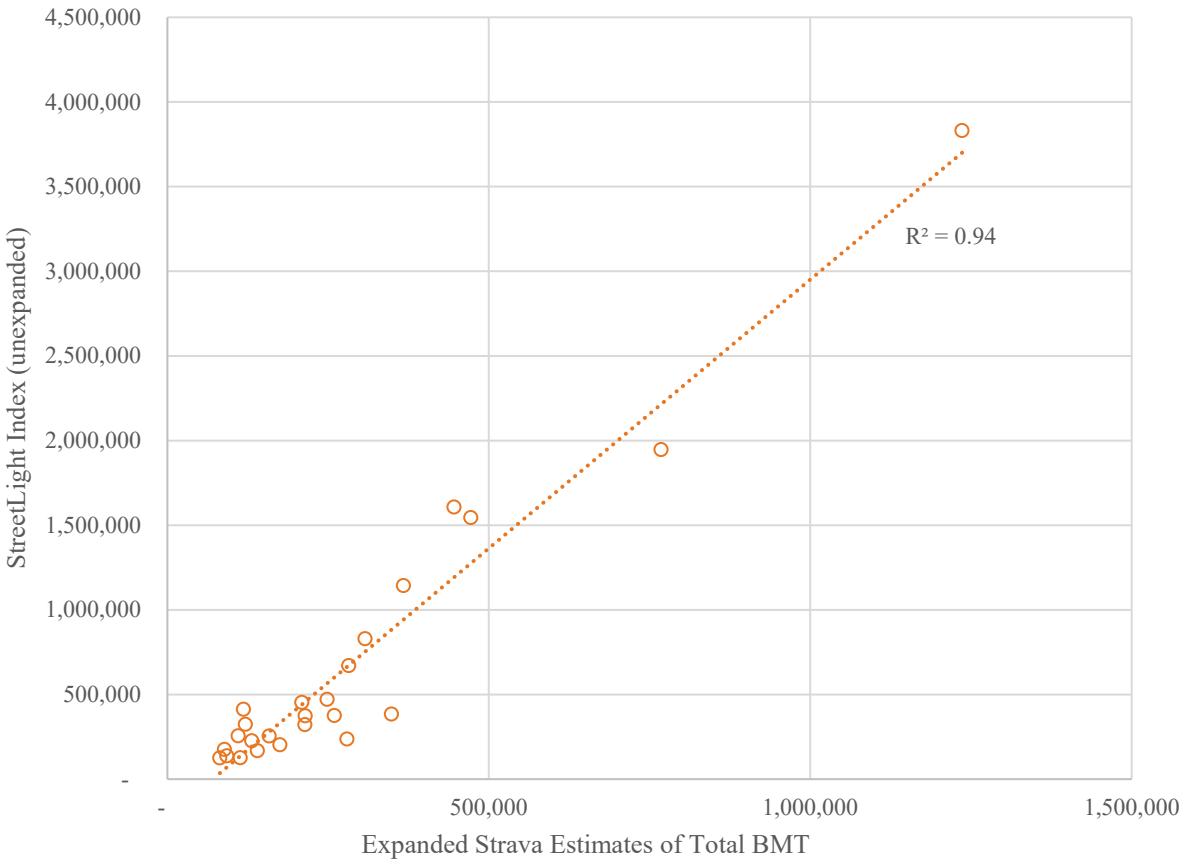


Figure 2. Graph. Comparison of county-level bicycling activity: top 25 counties in Texas.

As part of the data exploration, the researchers also compared several other bicycling exposure indicators to the StreetLight Index values. For example, population estimates are often used as a bicycling exposure indicator for aggregate (e.g., citywide, statewide, national) safety comparisons. Figure 3 shows a comparison of countywide population to countywide StreetLight Index values. The resulting correlation (R^2 = 99%) is even stronger than the previous countywide estimate of bicycling activity from expanded Strava estimates.

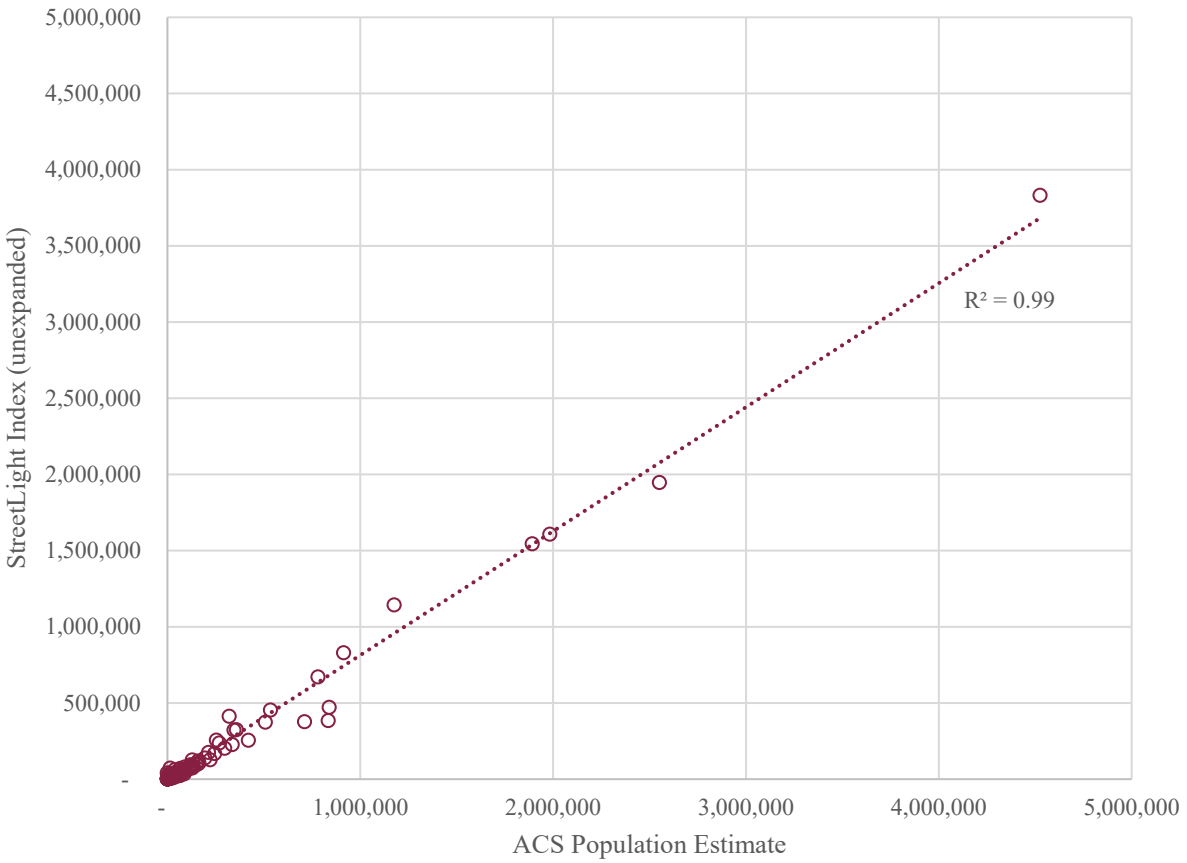


Figure 3. Graph. Comparison of county-level population to bicycling activity.

Conclusions

Overall, StreetLight Index values correlate well with permanent bicycle counts and very well with other bicycle activity estimates, like Strava and total population. StreetLight Index values appear to be affected by the variance in the underlying passive data samples over time and by location. This suggests that more detailed expansion methods that accommodate different functional classes and area types should be developed to more accurately estimate point location and areawide bicycle activity.

Pedestrian Volumes from Traffic Signal Systems

Evaluation Methods

The Texas A & M Transportation Institute (TTI) evaluated the accuracy of Miovision's TrafficLink Multimodal Detection and Counts by directly comparing their pedestrian crossing counts to benchmark (i.e., ground truth) counts established by TTI in a range of pedestrian traffic and lighting conditions. TTI established benchmark counts by using a double-blind manual counting process of source video collected by the Miovision detection system. The double-blind process worked as follows: two TTI employees manually and independently counted crossing

pedestrians in 15-minute intervals, each without the knowledge of the other's count result (i.e., double-blind). A TTI supervisor aggregated each 15-minute interval to 60-minute intervals and compared the TTI employees' independent counts. If the counts agreed within (i.e., less than or equal to) 3 and 10% tolerance, then the average of these two independent counts was considered the benchmark count against which the Miovision count was compared. If the two independent counts were not within (i.e., greater than) 3 and 10% tolerance of each other, then a recount was conducted until the two independent counts agreed within the acceptable tolerance. This ensured that the manual counts were a true and authoritative benchmark to compare to the Miovision counts. These thresholds were chosen because of the relative low pedestrian activity and the limited views of the source video (discussed next). Therefore, when the Miovision error is low (e.g., 10%–15%), that is approaching the error of the benchmark (10%), and in those cases, it is difficult to say with certainty that all error can be attributed to Miovision.

The source video used in the benchmark manual counts was from the Miovision video system, and therefore the exact times of the benchmark counts (derived from source video) were synchronized precisely with the times associated with the Miovision pedestrian counts. The source video was derived from Miovision's SmartView 360 "fisheye" camera. For example, Appendix A, Figure 8, shows five "split views" of the Martin Luther King Jr. Boulevard (MLK) at Guadalupe intersection. Some views were not as clear or as easy to reduce as others, such as the "Guadalupe South Leg" view.

Miovision and the City of Austin selected two intersection locations in Austin, Texas, for the evaluation. TTI only focused on the six defined pedestrian detection zones at the intersections: four intersection approaches (legs) at MLK and Guadalupe (Appendix A, Figure 9) and two intersection legs at Cameron and Cross Park (Appendix A, Figure 10). For example, the short crosswalk portion at the far west end of the North leg of the MLK and Guadalupe intersection was not counted by the Miovision system; therefore, this portion was not included in the evaluation.

TTI gathered 80 hours of benchmark source video at each intersection between June 18 and July 14, 2019. However, only 40 hours at each intersection were reduced and evaluated. Multiple hours at each intersection were evaluated during overnight periods (i.e., in dark periods with illumination only from streetlights).

TTI combined both crossing directions when making comparisons to the Miovision counts. TTI concluded that a count of total crossing pedestrians, by hour, would meet the requirements for most uses of historical pedestrian count data.

TTI calculated several different accuracy measures based on comparing the Miovision count values to the corresponding benchmark count values established by TTI.

- MAPE (Equation 1)
- Mean signed percent error, also known as bias (Equation 2)

Equation 1 $$Mean\ absolute\ percent\ error, MAPE\ (\%) = \frac{1}{n}\sum_{i=1}^{n}\frac{abs(x_i - \bar{x}_i)}{\bar{x}_i}$$

where $\bar{x}_i$ = TTI benchmark pedestrian count for the *i*th comparison
x_i = Miovision pedestrian count for the *i*th comparison
n = number of Miovision-to-benchmark comparisons

Equation 2 $$Mean\ signed\ percent\ error, or\ bias = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \bar{x}_i)}{\bar{x}_i}$$

where $\bar{x}_i$ = TTI benchmark pedestrian count for the *i*th comparison
x_i = Miovision pedestrian count for the *i*th comparison
n = number of Miovision-to-benchmark comparisons

Results

TTI summarized the accuracy measures across multiple dimensions to determine those factors that most influenced Miovision's accuracy, including intersection leg, ambient light (day or night), and level of pedestrian flow (low < 5 peds./hr., medium 6–31 peds./hr., and high >31 peds./hr.).

Table 3 shows the MAPE values for daytime conditions. The accuracy results are good with 11% error or less for the Cameron at Cross East leg and the MLK at Guadalupe South and North legs. However, the Cameron at Cross South leg and the MLK at Guadalupe East and West legs had 14% error or more.

Table 3. MAPE and Sample Size for Daytime Conditions

Intersection Leg	Low Pedestrian Flow: MAPE (sample size)	Medium Pedestrian Flow: MAPE (sample size)	High Pedestrian Flow: MAPE (sample size)
Cameron at Cross East	11% (22)	No data	No data
Cameron at Cross South	18% (21)	No data	No data
MLK at Guadalupe East	No data	37% (1)	34% (9)
MLK at Guadalupe South	0% (1)	8% (5)	11% (7)
MLK at Guadalupe North	No data	7% (9)	7% (5)
MLK at Guadalupe West	No data	28% (6)	14% (8)

Table 4 shows the MAPE values for nighttime conditions. The accuracy results were good with 3% error or less for the MLK at Guadalupe South leg (low pedestrian flow) and the North leg (high pedestrian flow) but those may be exceptions as they only had 2 and 1 samples, respectively. However, accuracy results were fair with 13%–17% error for the Cameron at Cross East leg and the MLK at Guadalupe East (medium and high pedestrian flow) and South leg (medium pedestrian flow). The remaining intersection legs had poor accuracy results of 21% or more. The poor accuracy may indicate a detection zone configuration and/or calibration issue. Note that the Cameron at Cross South (low pedestrian flow) was not reported because there were no pedestrians observed during the 14 hours evaluated.

Table 4. MAPE and Sample Size for Nighttime Conditions

Intersection Leg	Low Pedestrian Flow: MAPE (sample size)	Medium Pedestrian Flow: MAPE (sample size)	High Pedestrian Flow: MAPE (sample size)
Cameron at Cross East	14% (14)	No data	No data
Cameron at Cross South	NA*	No data	No data
MLK at Guadalupe East	33% (3)	17% (1)	13% (2)
MLK at Guadalupe South	0% (2)	13% (3)	No data
MLK at Guadalupe North	33% (3)	46% (3)	3% (1)
MLK at Guadalupe West	67% (3)	21% (4)	35% (1)

*Zero pedestrians observed on south leg (n = 14)

TTI summarized the accuracy results for daytime and nighttime, by intersection, as shown in Table 5. It shows 14% error for Cameron at Cross for both lighting conditions after removing the South leg nighttime hours as discussed previously. The accuracy for MLK at Guadalupe was 16% error for daytime and 28% for nighttime.

Table 5. Summary Accuracy Results and Sample Size for Miovision by Intersection

Intersection	MAPE (Daytime)	MAPE (Nighttime)
Cameron at Cross	14% (43)	14% (14)*
MLK at Guadalupe	16% (51)	28% (26)

*Does not include Cameron at Cross South leg (n = 14)

Table 6 shows that the accuracy for the combined intersection legs was 15% error for daytime and 24% for nighttime.

Table 6. Overall Accuracy Results and Sample Size for Miovision

Intersection	MAPE (Daytime)	MAPE (Nighttime)
Combined (All legs)	15% (94)	24% (40)*

*Does not include Cameron at Cross south leg (n = 14)

TTI also calculated the mean signed error to determine whether the Miovision counts were biased (consistently undercounting or overcounting pedestrians). Table 7 shows there was essentially no bias with 5% error for daytime and 2% for nighttime.

Table 7. Overall Accuracy Results and Sample Size for Miovision

Intersection	MAPE (Daytime)	MAPE (Nighttime)
Combined (All legs)	5% (94)	2% (40)*

*Does not include Cameron at Cross south leg (n=14)

Figure 4 is a correlation chart that shows each individual hourly comparison for all six intersection legs for the hours evaluated (minus nighttime hours at the Cameron at Cross South leg). This figure provides visual confirmation the mean error results of Miovision were basically random.

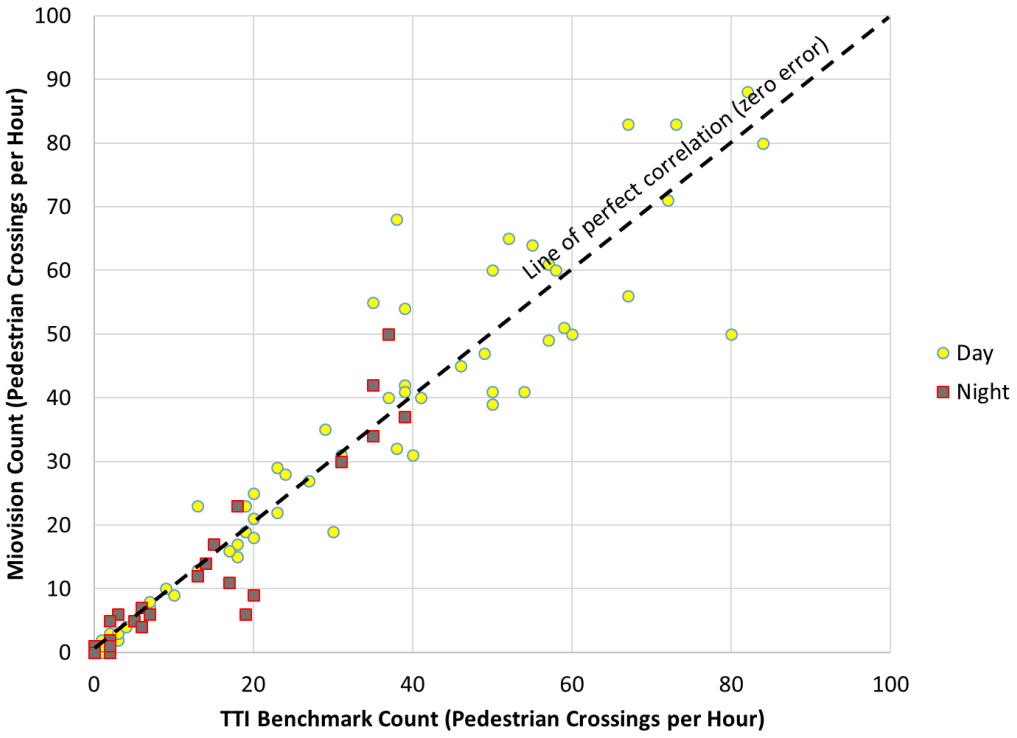


Figure 4. Graph. Visual comparison of hourly Miovision and TTI benchmark pedestrian counts.

Conclusions

The Miovision system for counting crossing pedestrians at intersections is accurate enough to be used by practitioners, even during hours of low light. The use of automated counting systems like the Miovision system has the potential to dramatically improve the quantity and quality of pedestrian exposure data available for safety analyses.

Travel Patterns using GPS Waypoint Data

Roadside Interview Survey (RSI) is a common method to determine roadway origin-destinations (O-D). Crowdsourced data from private companies can assist in developing new algorithms to understand O-D measures. Although these data sources suffer from limitations, they offer valuable

insight into many traffic operational exposure measures. The current study acquired four months (February, June, July, and October of 2015) of INRIX waypoint data that include vehicle trips from various data aggregator sources. To determine the relationship between O-D patterns and other area-specific variables, the research team acquired Census block-group-level data from the ACS and block-level economic data from the Longitudinal Employer-Household Dynamics (LEHD) program. This component of the project analyzed the relationship between demographic information and O-D patterns by Census spatial unit block group. The research team applied classification-based association rules mining to determine significant rules.

Literature Review

Many recent studies have aimed at determining O-D patterns using crowdsourced data. Some key studies in this area are briefly described in this section.

Sana et al. [8] used information from the Google Aggregated and Anonymized Trips (AAT) to develop a machine learning model and generate San Francisco Bay area hourly O-D demand matrices. They found that the developed model could effectively predict dynamic O-D person trip matrices by using both existing and future versions of AAT information. Ma et al. [9] developed a data-driven structure to estimate daily dynamic O-D using high-granular traffic frequency and speed data spanning many years. The developed framework employed t-Distributed Stochastic Neighbor Embedding (t-SNE) and k-means techniques to statistically cluster regular traffic data into typical traffic models. Fan et al. [10] conducted a study in Guangzhou City, China; they developed an O-D assessment methodology for systematic transit travelers. The researchers used smart card bus transportation data to improve the trip-chain O-D estimation algorithms. The results of the study are useful for real-world work associated with the O-D estimation. Ge et al. [11] used aggregated data of mobile phone traces to estimate work-related trips and develop a method to estimate O-D matrices based on the maximum entropy principle. The researchers calculated the trip production and attraction by using a nonlinear programming problem; they then used a matrix fitting problem to distribute trips to each O-D pair. Furthermore, two recent studies used Maryland INRIX waypoint data; one study estimated vehicle miles traveled [12] and the other determined the reliability of truck drivers' routing decisions [13].

Evaluation Methods

The primary objective of this study was to perform exploratory data analysis on the INRIX waypoint data as a part of Big Data analysis and to determine potential areas for future exploration.

Data Sources

The research team used three separate databases to perform the analysis:

- Maryland INRIX waypoint trip data (a waypoint is a geospatial point recorded by GPS devices that represents the location of the recorded vehicle)
- Census block-group-level ACS 2013–2017 data
- Census block-level LEHD 2015 data

Maryland INRIX Trip Data

The research team members collected Maryland INRIX trip data to perform the analysis. The data, covering four months in 2015 (February, June, July, and October) contain three types of monthly files in comma-separated values (CSV) format:

- TripRecordsReport (trip data)
- TripRecordsReportWaypoints (waypoint data)
- TripsRecordsReportProviderDetails (information on trip data providers)

The acquired dataset contains:

- 19,690,402 trips
- 1,376,720,203 waypoints
- 5,451,095 unique device identifications
- 148 data providers (45 providers)
    - 3 providers account for 52% of trips
    - 18 providers account for 99% of trips
- Four types of vehicle driving profiles: (1) consumer vehicles, (2) taxi/shuttle/town car service fleets, (3) local delivery fleets, and (4) for-hire/private trucking fleets
- Three vehicle weight classes: (1) light duty truck/passenger vehicle (0–14000 lbs.), (2) medium duty truck/vans (14001–26000 lbs.), and (3) heavy-duty truck (> 26000 lbs.)

ACS Data

The ACS data, acquired from the U.S. Census Bureau's Decennial Census Program, provide demographic, social, housing, and economic estimates for different spatial area units, including Census tract and block groups. The research team used ACS five-year (2013–2017) estimates for Maryland. The Maryland dataset contains 3,926 block groups. The research team collected a wider list of variables in the preliminary analysis.

LEHD Data

The research team also used the Census block data from the LEHD Origin-Destination Employment Statistics (LODES) 2015 data. The LODES files contain data for Residential Area Characteristics (RAC) and Work Area Characteristics (WAC). The block-level data were merged into the block group level to perform the analysis.

Data Integration

The data integration work included several steps:

- The team members acquired 2013–2017 ACS five-year estimates for Maryland from the U.S. Census to represent calendar year 2015 (i.e., for maximum ACS sample size). ACS data contain different tables such as age, gender, income, and household. The research team compiled the data for the key demographic and relevant data.

- The team members collected block-level 2015 LEHD data for Maryland. The block-level RAC and WAC data were merged into the block group level.
- The research team used the QGIS tool to spatially merge O-D data with census block groups. Later, several separate databases were developed based on the following:
    - Monthly O-D data by block group
    - Daily O-D data by block group
    - Hourly O-D data by block group
    - O-D data by vehicle type

Methodology

This study applied two methods to extract insights from O-D data: (1) a random forest algorithm to perform variable selection, and (2) rules mining to extract significant rules. The significant variables (block group level) selected by the random forest algorithm for model development include average O-D measure, total population (Popu), households (HH), households with family (HH_F), total WAC jobs (Total_Jobs_WAC), and household median income (HH_MedInc). This study used classification-based association rules to develop the rules associated with O-D patterns. The association rule can be represented as Antecedent (A) → Consequent (B), where both of them are disjoint itemsets.

$$Support\ of\ antecedent, S(A) = \frac{\sigma(A)}{N};$$

$$Support\ of\ consequent, S(B) = \frac{\sigma(B)}{N};\ \text{and}$$

$$Suppport\ of\ rule\ or\ S(A \rightarrow B) = \frac{\sigma(A \cap B)}{N}.$$

The measure of reliability for a generated rule is known as confidence: $C(A \rightarrow B) = \frac{S(A \rightarrow B)}{S(A)}$.

The lift, $L(A \rightarrow B) = \frac{S(A \rightarrow B)}{S(A).S(B)}$, is a measure that represents the ratio of confidence and expected confidence. A lift value greater than one shows positive interdependence between A and B, while a value smaller than one indicates a negative interdependence. A value of one signifies independence.

Results

Descriptive Statistics

The research team developed several data visualization plots from the O-D patterns. Figure 5 illustrates the O-D trips by block groups. A lighter color in a certain area indicates a lower number of trips, while darker colors indicate a higher number of trips per block group.

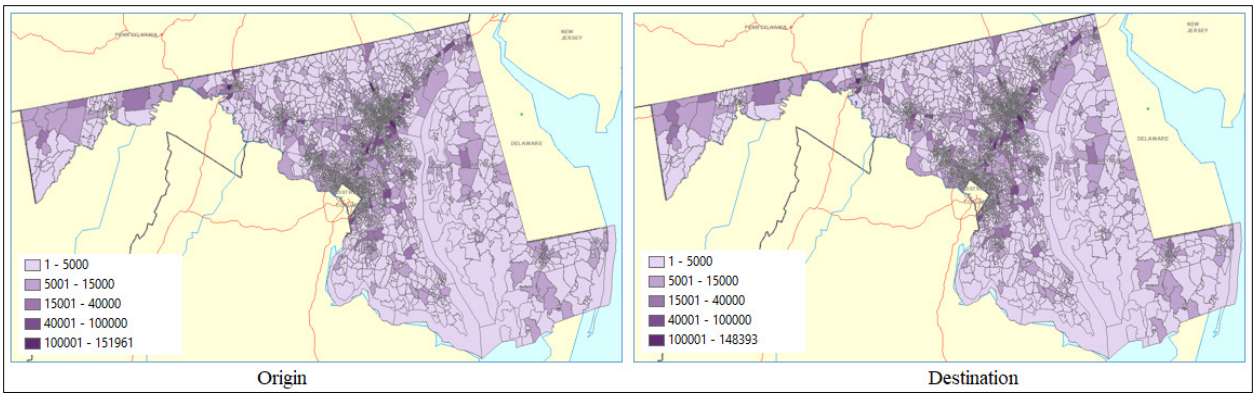


Figure 5. Maps. Number of trips by block groups based on origin and destination.

It is important to determine the key O-D matrices for different scenarios. Three different temporal patterns were used for analysis: all 24 hours, morning peak (6–10 a.m., Monday–Friday), and evening peak (4–8 p.m., Monday–Friday). In Appendix B, Figure 11 through Figure 14 show the O-D pairs for four vehicle types: consumer vehicles, delivery fleets, trucking fleets, and taxi fleets. Tables 9-12 list the top 20 O-D pairs with the number of trips in each pair for 24 hours.

Figure 15 provides a visualization of the number of trips by hour for different fleet types. In Figure 16, the top eight O-D generator block groups are plotted by hour for each day of the four months.

It is also important to show the trends of the O-D distribution for the top O-D generator block groups, and a chord diagram is an excellent tool to show these patterns. The research team developed four chord diagrams (by month) as shown in Figure 17 through Figure 20.

Rules Mining

The top 10 rules with high lift values are listed in Table 8 (a list of 40 rules is shown in Table 13). Average O-D measures per block group are divided into five classes based on the quantile percentages: TQ = 1 [1%–20%], TQ = 2 [21%–40%], TQ = 3 [41%–60%], TQ = 4 [61%–80%], and TQ = 5 [81%–100%]. For example, TQ = 5 indicates the block groups with the top 20% of the O-D trips. The findings show that higher population, WAC, and number of households are associated with a higher number of O-D trips. The rules provide several breakpoints of the variable clusters to determine the top rules.

Table 8. Top 10 Rules

Antece.	Conse.	S	C	L	Counts
Total_Jobs_WAC=(> 3778.5], HH_F=(386.5; 659.5]	TQ=5	0.0	1.0	5.0	35
Total_Jobs_WAC=(> 3778.5], Popu=(1366.5; 1978.5]	TQ=5	0.0	1.0	5.0	34
Total_Jobs_WAC=(> 3778.5], Popu=(1978.5; 3311.5]	TQ=5	0.0	1.0	5.0	31
Total_Jobs_WAC=(> 3778.5], HH_MedInc=(> 98467]	TQ=5	0.0	1.0	5.0	28
Total_Jobs_WAC=(> 3778.5], HH_MedInc=(53774;	TQ=5	0.0	0.9	4.9	61
Total_Jobs_WAC=(> 3778.5]	TQ=5	0.0	0.9	4.8	109
Popu=(> 3311.5], HH=(> 1252.5], HH_F=(> 934.5]	TQ=5	0.0	0.9	4.5	29

Antece.	Conse.	S	C	L	Counts
Total_Jobs_WAC=[< 21.5], HH=[< 264.5]	TQ=1	0.0	0.9	4.5	75
Total_Jobs_WAC=[< 21.5], HH_F=[< 230.5]	TQ=1	0.0	0.8	4.4	157
Total_Jobs_WAC=(833; 3778.5], HH=(> 1252.5]	TQ=5	0.0	0.8	4.4	22

Notes: Antece.= Antecedent, Conse.= Consequent, S = Support, C = Confidence, and L = Lift

Conclusions

The current study framework was applied to 19.8 million raw GPS vehicle trajectories collected in 2015 in Maryland. The results showed that the distribution of O-D measures for different vehicle types varies by hour and by month. The top 20 block groups with high average O-D measures contribute approximately 8.5% of trips. The generated top 40 rules provide several breakpoints of the key variables for appropriate rule development. The generated rules can be used for developing appropriate strategies in a way to improve the travel experiences of the roadway users.

The current study is not without limitations. The rules mining is limited to monthly O-D data. Future studies should explore the rules at more granular levels of data (e.g., daily or hourly). Also, this study is limited to only four months of data from 2015. A more comprehensive dataset is required to develop a robust framework of O-D measures.

Real-Time Crash Risk Prediction Using Waze Data

Identifying traffic crash hot spots is important for improving roadway safety. Traditionally, police crash reports (PCRs) have been the primary source of crash data in safety studies. However, using PCRs as the sole source of information has several drawbacks. For example, crashes that do not cause extensive property damage are mostly underreported. Underreporting can significantly influence the effectiveness of data-driven safety analysis and prevent safety analysts from reaching statistically meaningful results. Crowdsourced traffic incident data, such as Waze, can add a new dimension to the traditional safety analysis by providing real-time crash and traffic incident data. However, using these data sources also has some challenges. One of the major problems is data redundancy; many people may report the same event. This project investigated the potential of using crowdsourced Waze incident reports (WIRs) to better assess traffic risks on freeways. The researchers analyzed four weeks of WIRs and PCRs obtained from the I-35 corridor in North Texas. The researchers collected a whole week of data from four different months: August, October, November, and December 2016. First, the authors developed a new method to reduce data redundancy and obtain unique Waze incidents (unique WIRs). The researchers then matched the unique WIRs with the observed crashes and compared their spatial and temporal distributions. In addition, the researchers estimated predicted crashes through safety performance functions (SPFs) and crash modification factors (CMFs) to assess whether the WIR data can be used as a reliable surrogate of these safety measures (i.e., observed crash frequency and predicted crashes) for identifying high-risk locations.

Evaluation Methods

Figure 6 illustrates the flow chart of the research methodology used in this paper. The researchers utilized three data sources: PCRs, WIRs, and roadway inventory shapefiles. The researchers first selected freeway crashes from PCRs and WIRs by removing crashes on frontage roads and ramp exits and entrances. Then, the duplicate WIRs were eliminated to identify unique Waze incident events (unique WIRs). A similar process was performed to match the unique WIRs with PCRs to create a merged dataset (PCRs + WIRs). Meanwhile, the researchers calculated the predicted crash frequency using freeway SPFs and CMFs. Finally, the researchers created four safety datasets: WIRs, PCRs, the merged dataset, and predicted crashes.

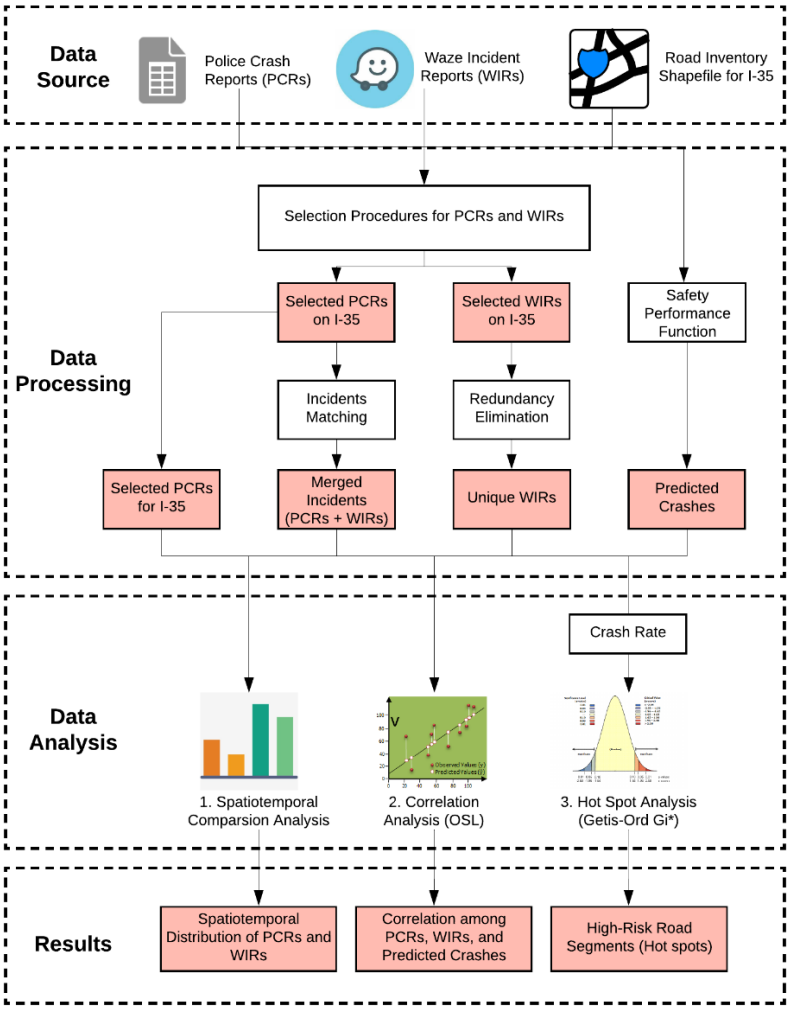


Figure 6. Flow chart. Research methodology.

To better explore the potential of WIRs in road safety analysis, three analyses were conducted:

1) *Spatiotemporal comparison analysis*: characterize the spatiotemporal distributions of PCRs and WIRs.

2) *Correlation analysis*: investigate the relationship between PCRs, WIRs, and predicted crashes to test further if WIRs could be used as a surrogate safety measure when PCRs are unavailable.
3) *Hot spot analysis*:

    a. Calculate crash rates for each road segment using PCRs, unique WIRs, the merged dataset, and predicted crashes.
    b. Perform hot spot analysis (Getis-Ord Gi*) using different crash rates to identify high-risk road segments. This analysis evaluated if WIRs could capture traffic risks that are ignored by conventional crash datasets (e.g., PCRs).

This study also assessed the performance of WIRs for identifying high-risk road segments. The researchers first calculated crash rates for each road segment using four different data sources, including PCRs, unique WIRs, the merged dataset, and predicted crashes. Then, Getis-Ord Gi* statistics were conducted based on different crash rates to identify hot spots, defined as high-valued road segments surrounded by high-valued neighboring segments. This study compared hot spots detected from different data sources in different months to investigate if the distribution of hot spots varied from month to month. The researchers also examined the monthly results with the hot spots detected from four-month datasets to identify constant hot spots. This study defines *constant hot spot* as a segment, or its neighboring segments (within ± 1 mile), that (1) was determined to be a hot spot in more than two different months and (2) also was identified as a hot spot in the four-month dataset.

Results

Result for WIR Redundancy Elimination and Matching with PCRs

The "true" incident, the PCR, was used as the starting point. Different combinations of spatial and temporal thresholds were tested to remove redundant WIRs corresponding to the same PCR and to match unique WIRs with the PCRs. The researchers hypothesized that when spatial and temporal "distances" from the true, PCR incident to the surrogate, WIR, incident reach their optimal value, the number of matched WIRs should experience a significant increase since more redundant WIRs can be captured. After the optimal threshold is attained, the number of matched WIRs should not be significantly different from the optimal number of matched WIRs. As the results of these analyses, the researchers determined the optimal spatial and temporal thresholds for identifying the redundant WIRs as:

- Spatial threshold: a 2,250-meter radius.
- Temporal thresholds: 90 minutes (−20 to 70 minutes).

By applying these thresholds, 1,807 WIRs were finally consolidated into 381 unique WIRs. A similar process was conducted to match unique WIRs with PCRs. In this study, only 13 out of 177 PCRs (7.34%) were matched with the unique WIRs (13 out of 381). Finally, the researchers created

a merged database by combining PCRs with unmatched unique WIRs. This dataset contains 545 traffic incidents and crashes.

Spatiotemporal Comparison Analysis Result

Through the spatiotemporal comparison of PCRs and WIRs, the researchers found that these two data sources show a very similar spatial distribution. However, the temporal comparison shows a significant difference between them. In this study, PCRs were reported during daytime, while WIRs were more intensively reported during nighttime. It is also worth noting that 60.24% of the road segments in the study site received more WIRs than PCRs, and 27.1% received the same number of WIRs and PCRs. This implies that unreported traffic incidents were more prevalent on most of the road segments. These traffic incidents should be considered in road safety studies.

Correlation Analysis Result

This approach investigated the relationship between PCRs, unique WIRs, and the estimated crashes through predictive models to statistically test if WIRs could be used as a surrogate data source or safety measures in the absence of crash data. The correlations among these three datasets are detailed in Figure 7. This figure illustrates that PCRs are better correlated with WIRs (0.63) than with predicted crashes (0.57), although this correlation is not significant enough to suggest multicollinearity. It also suggests that WIRs can better represent the predicted safety risk than PCRs (0.70 vs. 0.57). The researchers also developed an ordinary least squares (OLS) regression model to further investigate the relationship between the three safety measures PCR, WIR, and predicted crashes. Two regression models were constructed. One uses unique WIRs alone as the independent variable. The other uses both WIRs and predicted crashes as independent variables. The regression models suggest that both WIRs and predicted crashes are significant predictors for estimating PCRs. However, using WIRs alone may not be capable enough since the model performance is relatively unsatisfying with an R^2 of 0.4.

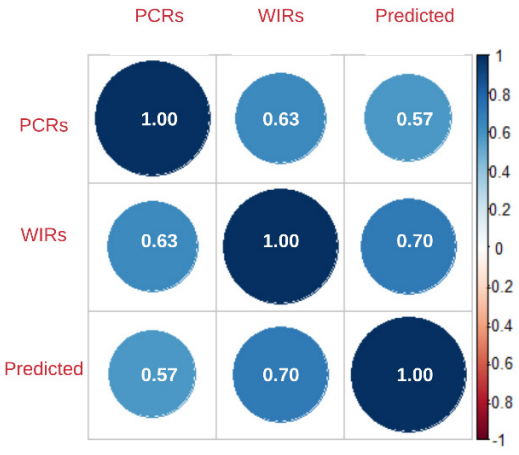


Figure 7. Illustration. Correlations among PCRs, unique WIRs, and predicted crashes.

Hot Spot Analysis Result

The result demonstrates that the hot spots may vary in different months; however, there are still some constant hot spots which may be considered as true high-risk segments. By combing PCRs with WIRs, more high-risk road segments could be identified (14 miles) compared to the results generated from PCRs (8 miles), unique WIRs (13 miles), and predicted crashes (5 miles). Most of the hot spots detected from PCRs (75%), unique WIRs (77%), and predicted crashes (100%) could be identified from the merged data. Therefore, it can be concluded that integrating WIRs and PCRs can better capture traffic risks and discover more unidentified high-risk road segments.

Conclusions

This study shows that overall Waze is an invaluable source of data for safety researchers that is tremendously useful for capturing unreported traffic incidents. However, there are still some gaps that were not adequately addressed by this study. Although the findings are promising, the researchers used Waze data only from an interstate corridor, which is generally assumed to generate more Waze reports. This gap may also affect some of the findings; for example, the temporal and spatial thresholds for consolidating the WIRs and matching them with PCRs may not be applicable to other facility types. Future research will focus on these areas.

Additional Products

The Education and Workforce Development (EWD) and Technology Transfer (T2) products created as part of this project can be downloaded from the Project Page on the Safe-D website. The final project dataset is located on the Safe-D Collection of the VTTI Dataverse.

Education and Workforce Development Products

Education and workforce development products include:

- TTI Assistant Research Scientist Lingtao Wu, Ph.D., helped in developing a replicable framework for the spatial join, which is available as open source code on Github: https://github.tamu.edu/wulingtao/Post_GIS_Spatial_Join
    - Eight students assisted with this research: Greg Griffin (graduate in Community and Regional Planning), Xiao Li (graduate in Geography), Dennis Mbaka (undergraduate in Civil Engineering), Viviana Rodriguez (undergraduate in Civil Engineering), Raquel De La Torre (undergraduate in Art), Ly-Na Tran (undergraduate), Ruihong Wang (graduate in Electrical and Computer Engineering), and Manya Umamahesh (graduate in Information Technology and Management).

Technology Transfer Products

The following products were or will be generated:

- Wang, R., S. Das, and A. Mudgal. "Patterns of Origin Destination Distributions: Rules Mining using Massive GPS Trajectory Data." Proceedings of UDS'20: First International Conference on Urban Data Science, January 20-21, Madras, India.
- Li, X., B. Dadashova, S. Turner, and D. Goldberg. "Rethinking Highway Safety Analysis by Leveraging Crowdsourced Waze Data." Presented at the 99th TRB Annual Meeting, Washington, DC, January 12-16, 2020.
- Turner, S. "Making Sense of Emerging Data for Nonmotorized Transportation." Presented at the 99th TRB Annual Meeting, Washington, DC, January 12-16, 2020.
- Le, M. "Video Analytics for Counting Pedestrians." To be presented at the Texas Trails & Active Transportation Conference, San Antonio, TX, March 25, 2020.
- Turner, S. "Emerging Sources of Pedestrian and Bicyclist Count Data." To be presented at the National Travel Monitoring Exposition and Conference, Raleigh, NC, June 2020.
- A research results briefing was held with Miovision (about 8 participants) and StreetLight Data company representatives (about 5 participants) in September 2019.

Data Products

Three datasets were provided for the research described in this report. These datasets were uploaded to the Safe-D Collection on the VTTI Dataverse and are described below:

- Evaluation of StreetLight Data Bicycle Count Estimates: https://doi.org/10.15787/VTT1/OBV82F. These data represent monthly average weekday and weekend bicycle counts and count index values at specific points locations across eight cities in Texas: Allen, Austin, Dallas, Fort Worth, Houston, North Richland Hills, Plano, and San Antonio.
- Evaluation of Miovision Pedestrian Count Data: https://doi.org/10.15787/VTT1/351GZJ. The data represent one week of selected hourly weekday and weekend pedestrian counts at two intersections in Austin, Texas.
- Travel Patterns using GPS Waypoint Data: https://doi.org/10.15787/VTT1/81SKJW. The dataset is a monthly sample dataset including U.S. Census Block group specific monthly origin-destination data with other demographic variables.

References

1. Boss, D., T. Nelson, M. Winters, and C. J. Ferster. Using Crowdsourced Data to Monitor Change in Spatial Patterns of Bicycle Ridership. *Journal of Transport & Health*, March 2018. https://doi.org/10.1016/j.jth.2018.02.008.
2. Conrow, L., E. Wentz, T. Nelson, and C. Pettit. Comparing Spatial Patterns of Crowdsourced and Conventional Bicycling Datasets." *Applied Geography*, Vol. 92, 2018, pp. 21–30. https://doi.org/10.1016/j.apgeog.2018.01.009.
3. Griffin, G. P., and J. Jiao. Crowdsourcing Bicycle Volumes: Exploring the Role of Volunteered Geographic Information and Established Monitoring Methods. *URISA Journal*, Vol. 27, No. 1, 2015, pp. 57–66.
4. Hochmair, H. H., E. Bardin, and A. Ahmouda. Estimating Bicycle Trip Volume for Miami-Dade County from Strava Tracking Data. *Journal of Transport Geography*, Vol. 75, January 2019, pp. 58–69. https://doi.org/10.1016/j.jtrangeo.2019.01.013.
5. Roy, A., and T. Nelson. Bias Correction in Geolocated Crowdsourced Data from Strava Using Machine Learning Based Linear Models. In *Geophysical Research Abstracts*, Vol. 20, 2018.
6. Dadashova, B., G. P. Griffin, S. Das, S. Turner, and M. Graham. Guide for Seasonal Adjustment and Crowdsourced Data Scaling. Report FHWA/TX-18/0-6927-P6. Texas A&M Transportation Institute, December 2018.
7. Turner, S., R. Benz, J. Hudson, G. Griffin, P. Lasley, B. Dadashova, and S. Das. Improving the Amount and Availability of Pedestrian and Bicyclist Count Data in Texas. Report FHWA/TX-19/0-6927-R1. Texas A&M Transportation Institute, February 2019.
8. Sana, B., J. Castiglione, D. Cooper, and D. Tischler. Using Google's Passive Data and Machine Learning for Origin-Destination Demand Estimation. *Transportation Research Record*, Vol. 2672, No. 46, 2018, pp. 73–82.
9. Ma, W., and Z. Qian. Estimating Multi-Year 24/7 Origin-Destination Demand Using High-Granular Multi-Source Traffic Data. *Transportation Research Part C: Emerging Technologies*, Vol. 96, 2018, pp. 96–121.
10. Fan., W, and Z. Chen. Estimation of Origin-Destination Matrix and Identification of User Activities Using Public Transit Smart Card Data. Center for Advanced Multimodal Mobility Solutions and Education, 2018.
11. Ge, Q., and F. Daisuke. Updating Origin–Destination Matrices with Aggregated Data of GPS Traces. *Transportation Research Part C: Emerging Technologies*, Vol. 69, 2016, pp. 291–312.
12. Fan, J., C. Fu, K. Stewart, and L. Zhang. Using Big GPS Trajectory Data Analytics for Vehicle Miles Traveled Estimation. *Transportation Research Part C*, Vol. 103, 2019, pp. 298–307.
13. Kong, X., W. Eisele, Y. Zhang, and D. Cline. Evaluating the Impact of Real-time Mobility and Travel Time Reliability Information on Truck Drivers' Routing Decisions.

Transportation Research Record: Journal of the Transportation Research Board. Vol. 2672, 2018, pp. 164–172.

Appendix A: Pedestrian Volumes from Traffic Signal Systems

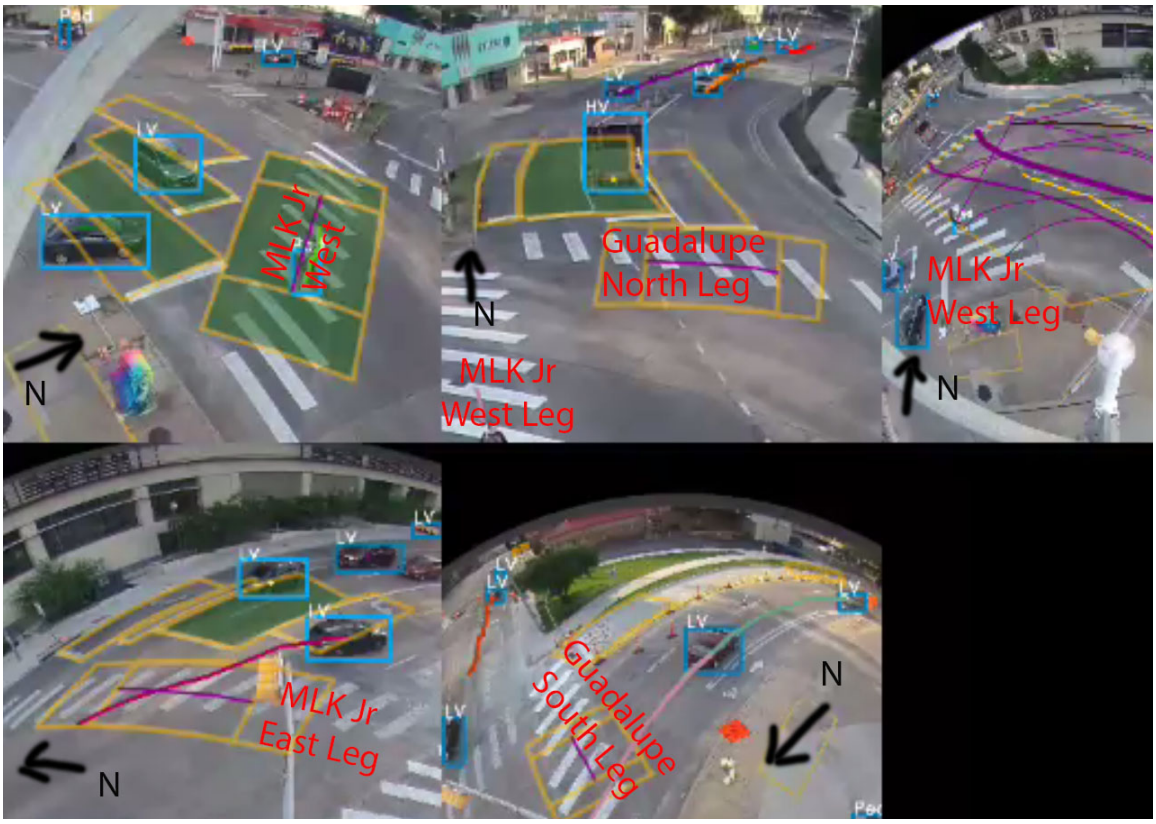


Figure 8. Screen captures. Example of Miovision source video used to develop benchmark counts (showing MLK at Guadalupe).

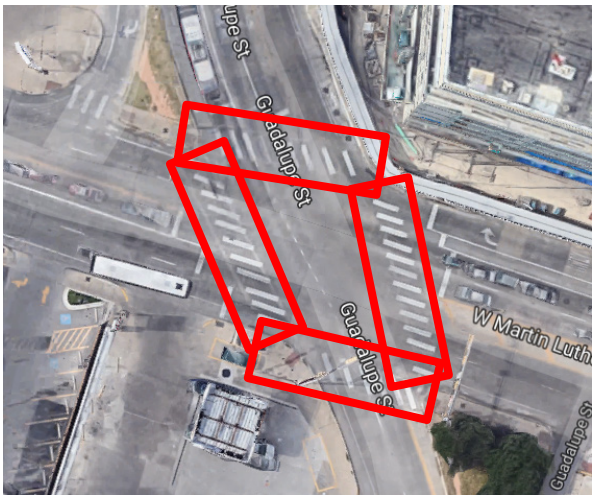


Figure 9. Screen capture. Crosswalks evaluated at MLK and Guadalupe.

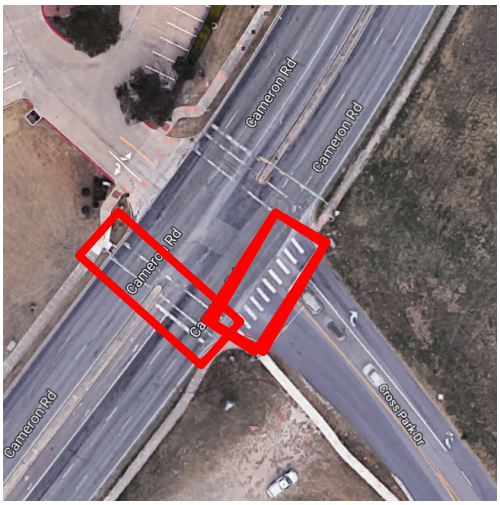
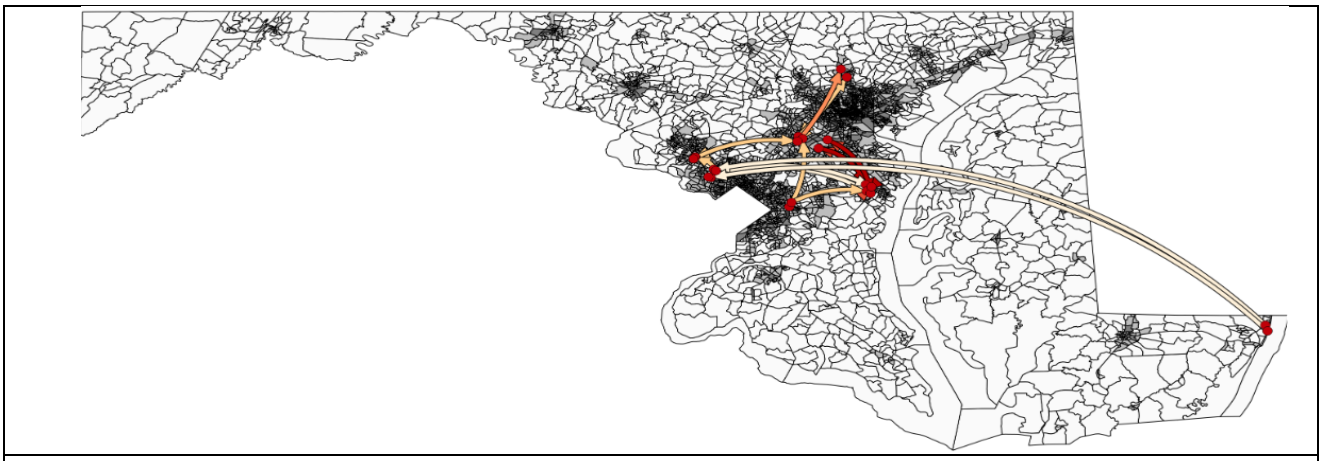
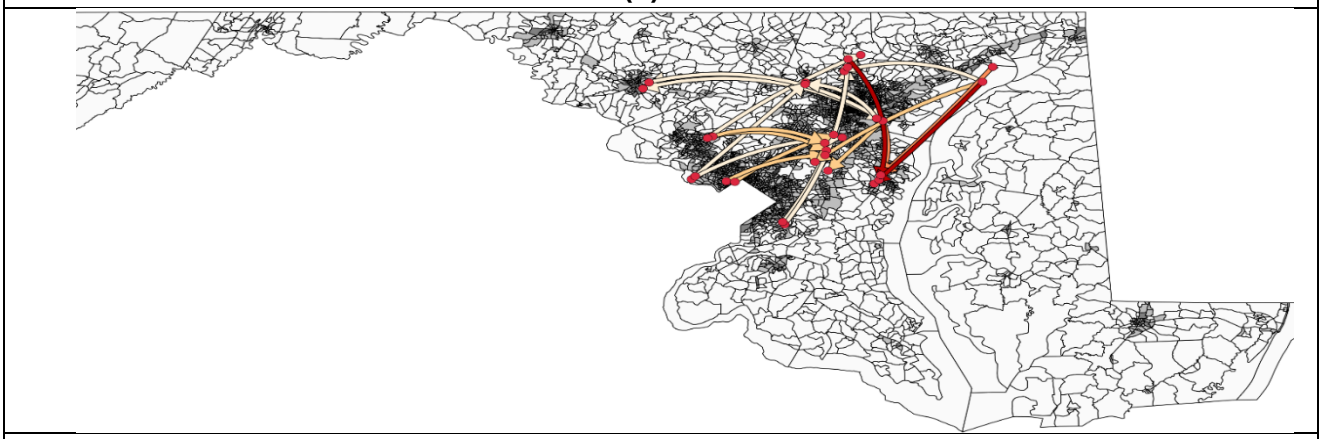


Figure 10. Screen capture. Crosswalks Evaluated at Cameron & Cross Park

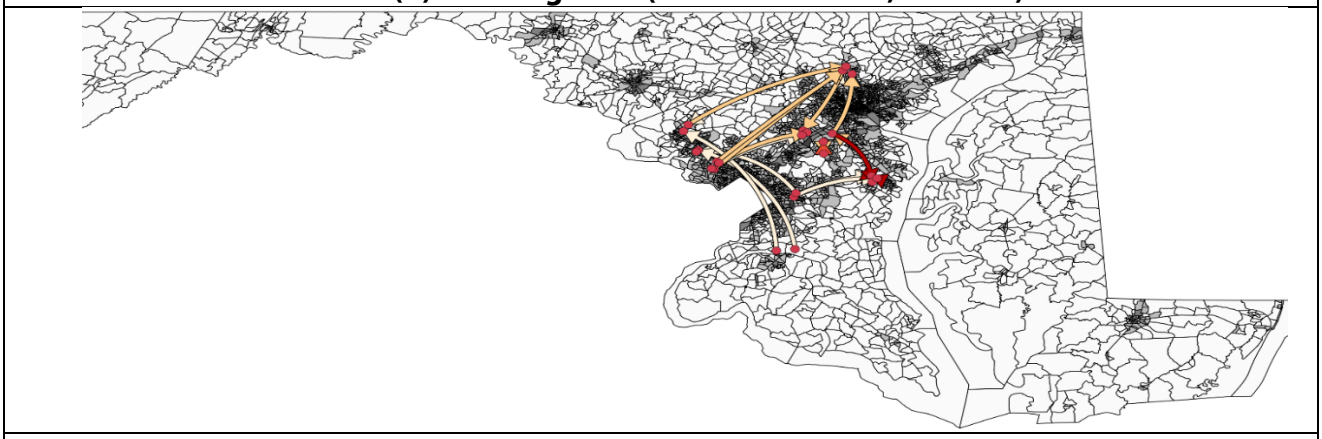
Appendix B: Travel Patterns using GPS Waypoint Data – Supporting Figures and Tables



(a) 24 hours

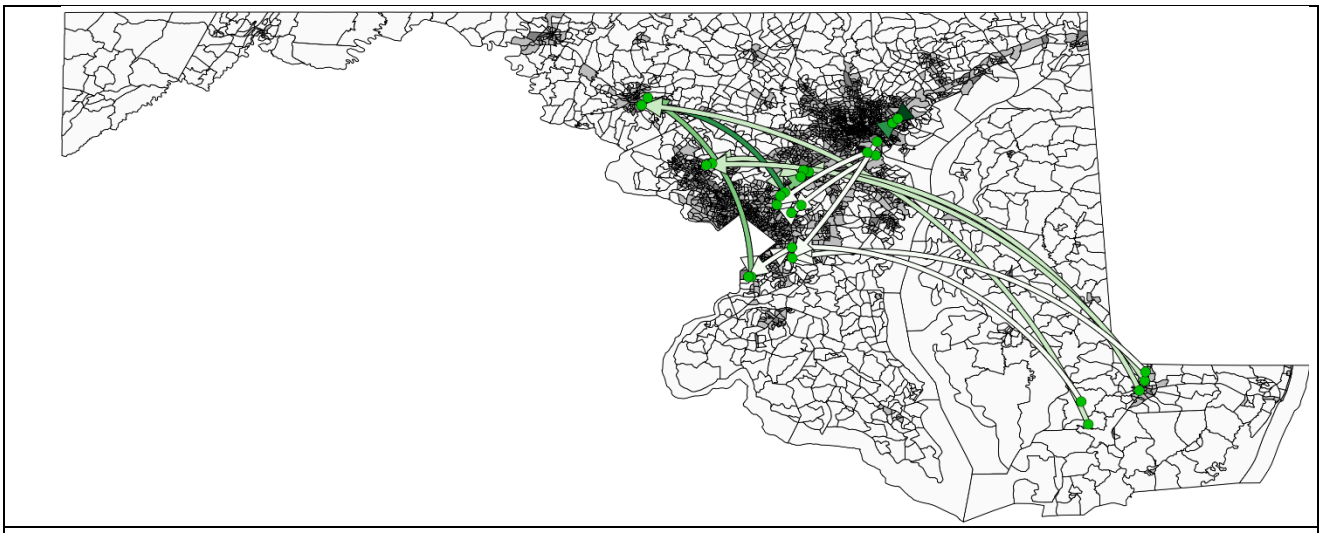


(b) Morning Peak (6 a.m. to 10 a.m., Mon-Fri)

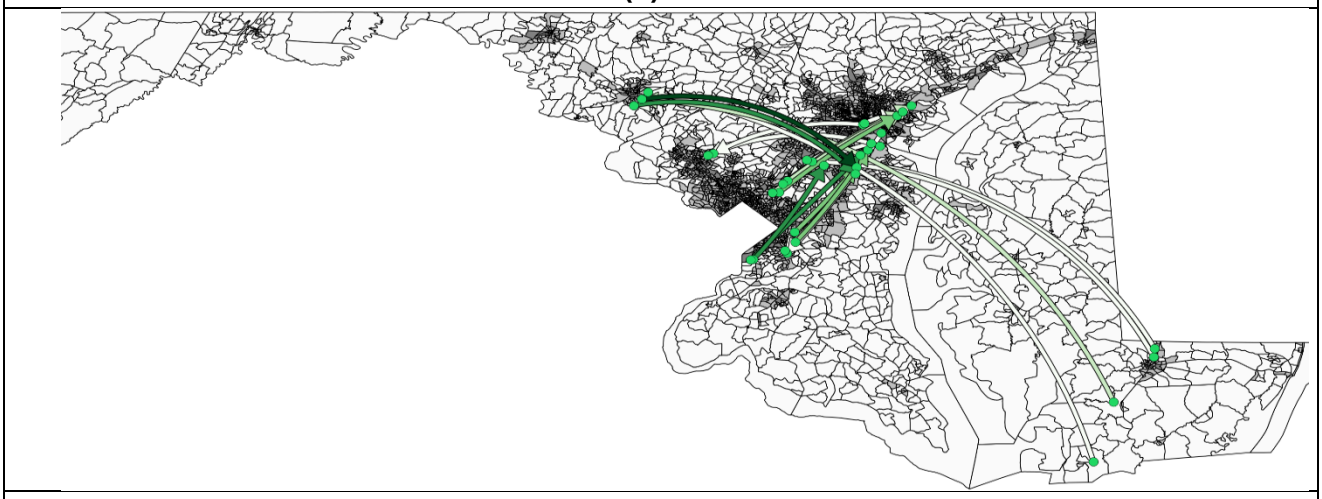


(c) Evening Peak (4 p.m. to 8 p.m., Monday through Friday)

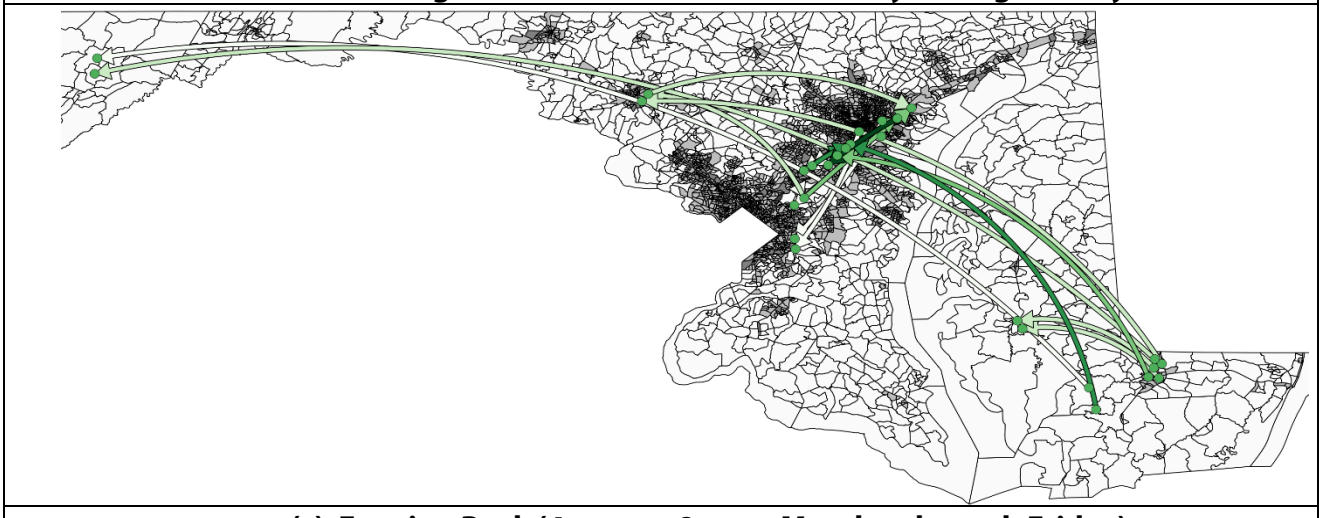
Figure 11. Maps. Top 20 O-D pairs for consumer vehicles.



(a) 24 hours

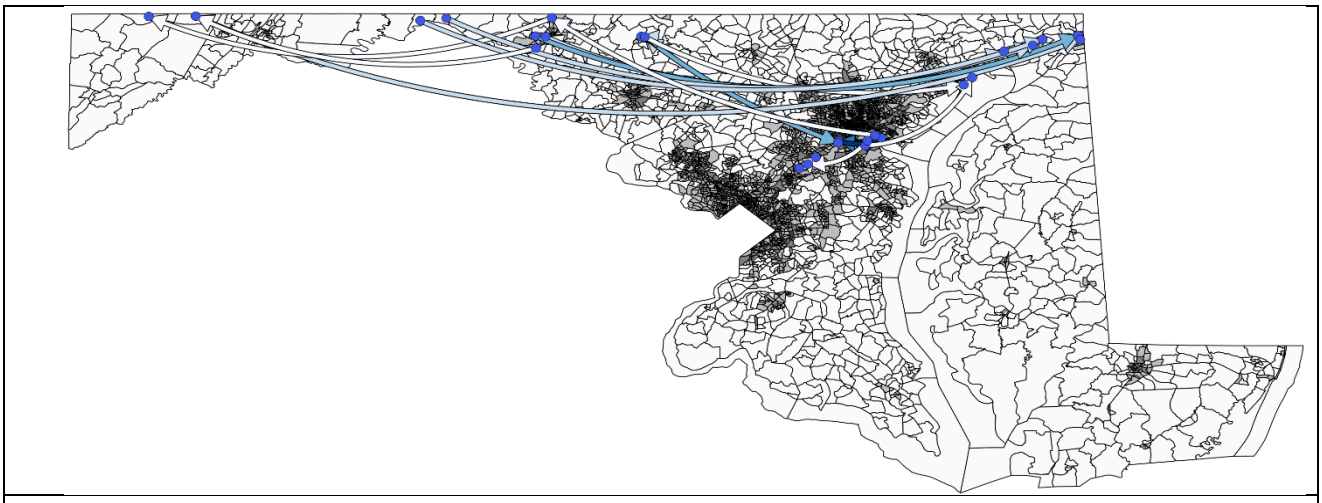


(b) Morning Peak (6 a.m. to 10 a.m., Monday through Friday)

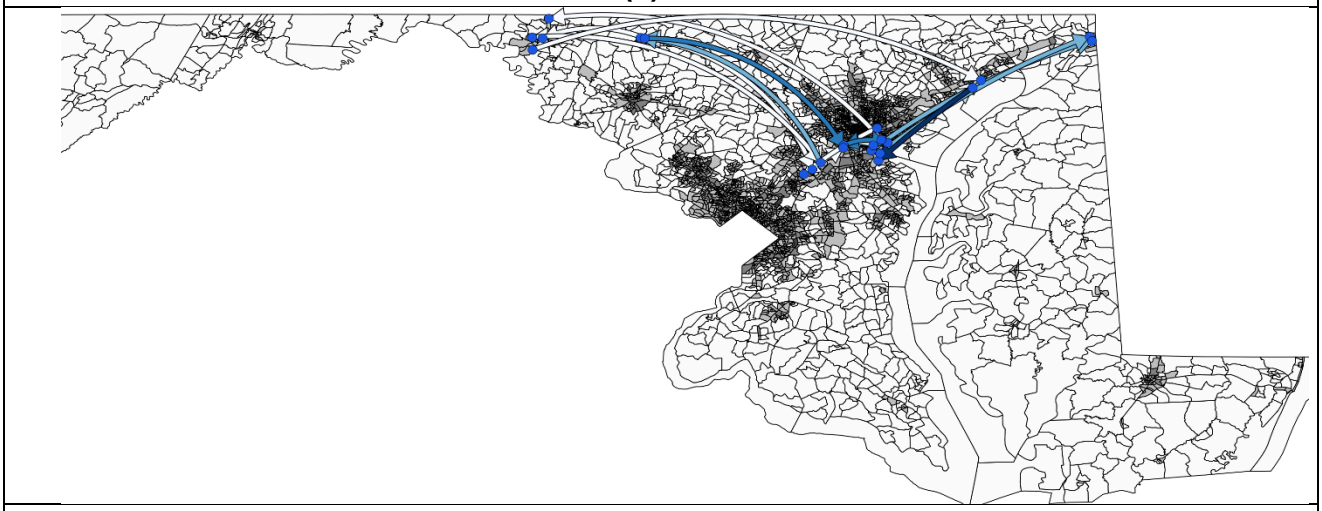


(c) Evening Peak (4 p.m. to 8 p.m., Monday through Friday)

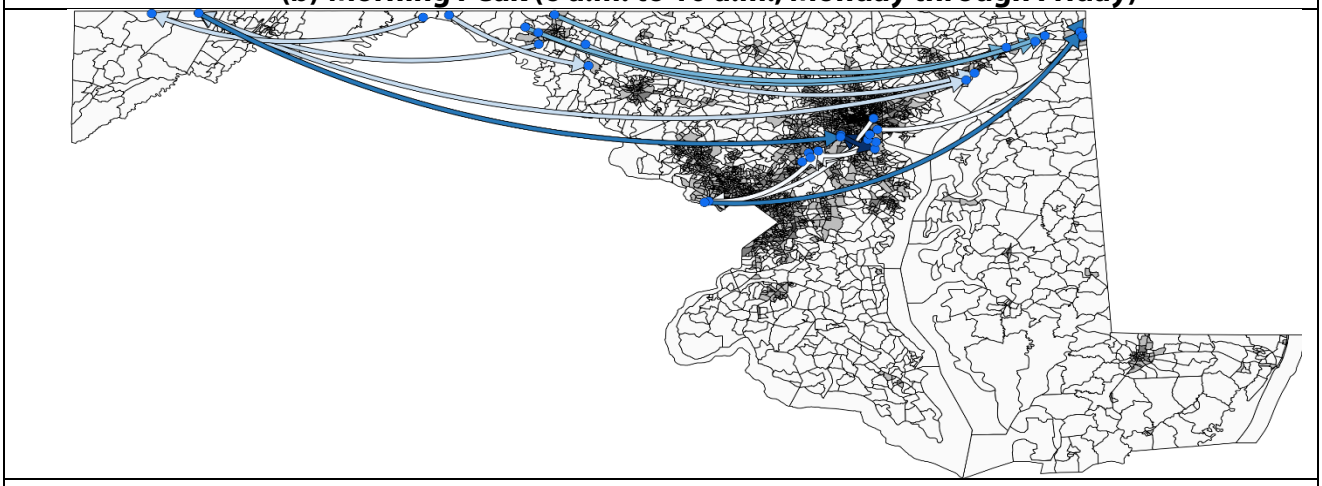
Figure 12. Maps. Top 20 O-D pairs for field service/local delivery fleets.



(a) 24 hours

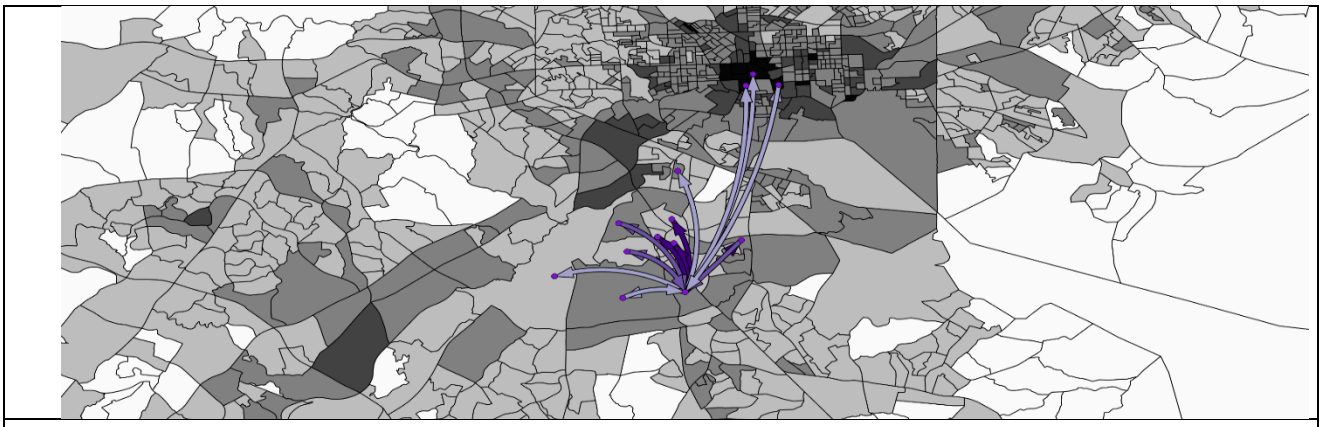


(b) Morning Peak (6 a.m. to 10 a.m., Monday through Friday)

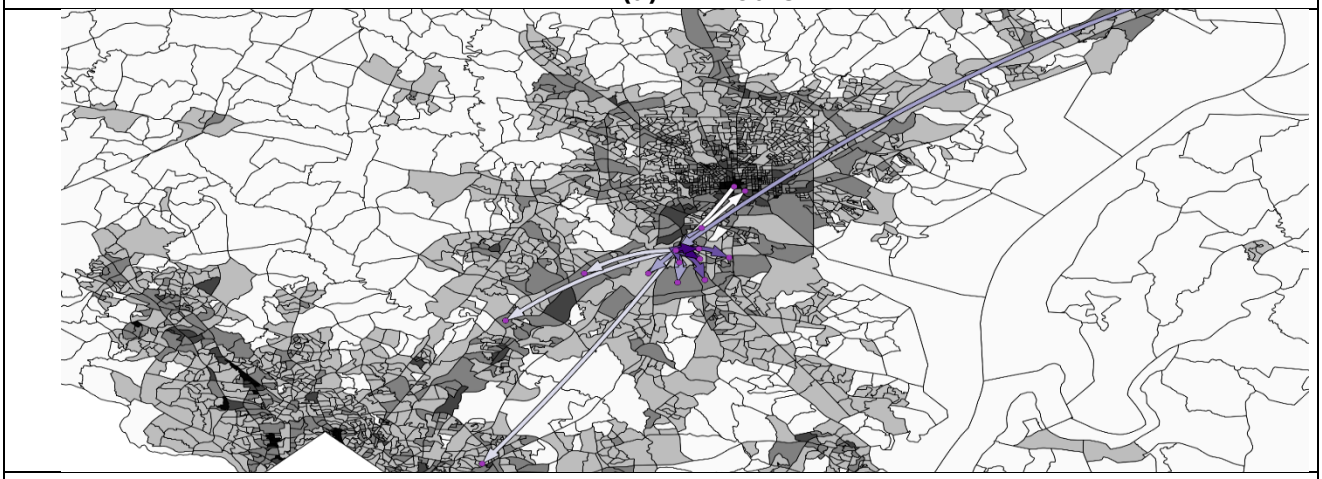


(c) Evening Peak (4 p.m. to 8 p.m., Monday through Friday)

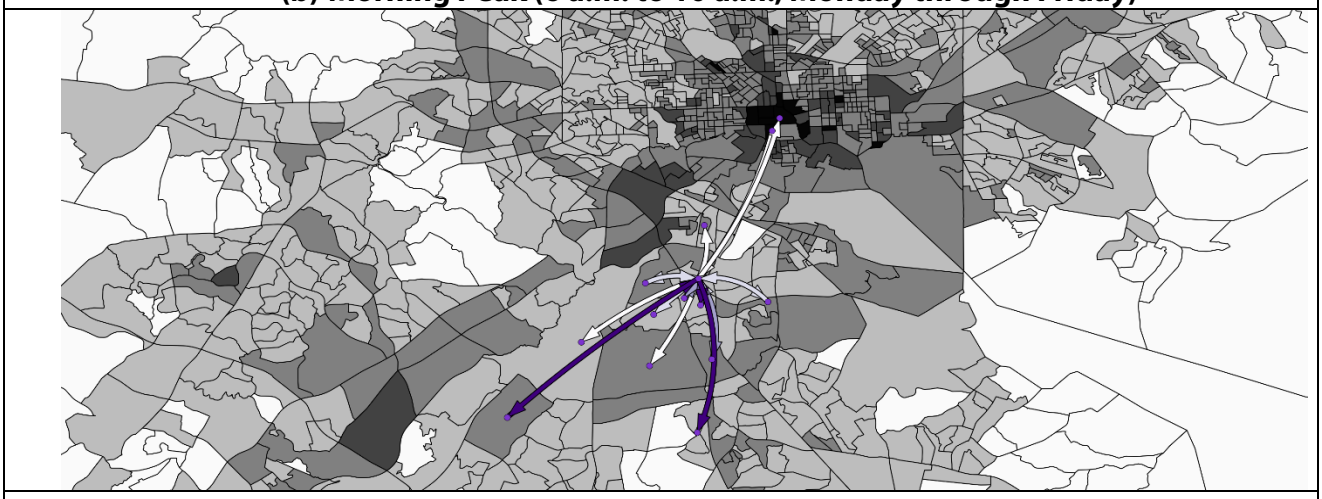
Figure 13. Maps. Top 20 O-D pairs for hire/private trucking fleets.



(a) 24 Hours



(b) Morning Peak (6 a.m. to 10 a.m., Monday through Friday)



(c) Evening Peak (4 p.m. to 8 p.m., Monday through Friday)

Figure 14. Maps. Top 20 O-D Pairs for taxi/shuttle/town car service fleets.

Table 9. O-D Pair Rank for Consumer Vehicles (Four months of 2015)

Rank	Origin GEOID	Destination GEOID	Total Trips
1	US 240037401022	US 240037512003	2143
2	US 240037024022	US 240037027011	2060
3	US 240317009041	US 240317012183	1925
4	US 240037027011	US 240037024022	1883
5	US 240037512003	US 240037401022	1870
6	US 240317012183	US 240317009041	1757
7	US 240276067061	US 240276067071	1369
8	US 240054085052	US 240054084001	1257
9	US 240054084001	US 240054085052	1200
10	US 240479501001	US 240479501002	1143
11	US 240037024022	US 240037025001	1104
12	US 240276067071	US 240276067063	1096
13	US 240276067071	US 240276067061	1086
14	US 240317012053	US 240317060123	1048
15	US 240276067063	US 240276067071	1016
16	US 240317008171	US 240317008173	998
17	US 240338035141	US 240338035211	993
18	US 240317060123	US 240317012053	972
19	US 240037027011	US 240037024021	934
20	US 240037024021	US 240037024022	919

Table 10. O-D Pair Rank for Field Service/Local Delivery Fleets (Four months of 2015)

Rank	Origin GEOID	Destination GEOID	Total Trips
1	US 240037508014	US 240039800001	1482
2	US 240037512002	US 240039800001	1447
3	US 240037512003	US 240039800001	778
4	US 240039800001	US 240037512002	716
5	US 240039800001	US 240037512003	513
6	US 240037504002	US 240039800001	442
7	US 245100401001	US 240039800001	399
8	US 240037512001	US 240039800001	389
9	US 240039800001	US 240037504002	389
10	US 240039800001	US 245100401001	361
11	US 240039800001	US 240037512001	329
12	US 240054301011	US 240039800001	270
13	US 245100302002	US 240039800001	267
14	US 240037508033	US 240039800001	261
15	US 240039800001	US 245102201002	255
16	US 240039800001	US 245100302002	252
17	US 240039800001	US 240037508014	244
18	US 245102201002	US 240039800001	242
19	US 240037504003	US 240039800001	230
20	US 240037503002	US 240039800001	224

Table 11. O-D Pair Rank for Hire/Private Trucking Fleets (Four months of 2015)

Rank	Origin GEOID	Destination GEOID	Total Trips
1	US 240338004111	US 240338067121	3423
2	US 240338067121	US 240338004111	2837
3	US 240338074083	US 240338074042	2825
4	US 240217510033	US 240217722002	2542
5	US 240317001051	US 240317001033	2386
6	US 240338074083	US 240338002122	2006
7	US 240399301022	US 240450108003	1997
8	US 240338022044	US 240338022012	1941
9	US 240338022012	US 240338022044	1912
10	US 245102505002	US 245102505001	1905
11	US 240338074042	US 240338074083	1895
12	US 240276069053	US 240276069012	1783
13	US 240217722002	US 240217510033	1691
14	US 245102606056	US 240054501003	1685
15	US 240450101012	US 240450107026	1673
16	US 240054501003	US 245102606056	1593
17	US 240054501003	US 240054512001	1576
18	US 240276069012	US 240276069013	1552
19	US 240338014111	US 240338014112	1508
20	US 240450101012	US 240450004001	1501

Table 12. O-D Pair Rank for Taxi/Shuttle/Town Car Service Fleets (Four months of 2015)

Rank	Origin GEOID	Destination GEOID	Total Trips
1	US 240059802001	US 240054304003	7315
2	US 240054304003	US 240059802001	4625
3	US 240430108022	US 240430105004	3892
4	US 245102505002	US 245102506001	3536
5	US 240276069053	US 240276069012	3275
6	US 240217530022	US 240217530024	2921
7	US 240150305062	US 240150305033	2726
8	US 240054211022	US 245102606056	2634
9	US 240430009003	US 240430105004	2598
10	US 240430105004	US 240430108022	2593
11	US 240150312021	US 240150309063	2500
12	US 240276069012	US 240276012032	2477
13	US 240430103001	US 240430105004	2362
14	US 240230003001	US 240230002001	2269
15	US 240253024001	US 240253024002	2235
16	US 240430107002	US 240430107001	2216
17	US 240253024002	US 240253024001	2148
18	US 245102606056	US 240054211022	2117
19	US 240150309063	US 240150309064	2088
20	US 240276069012	US 240276069053	2088

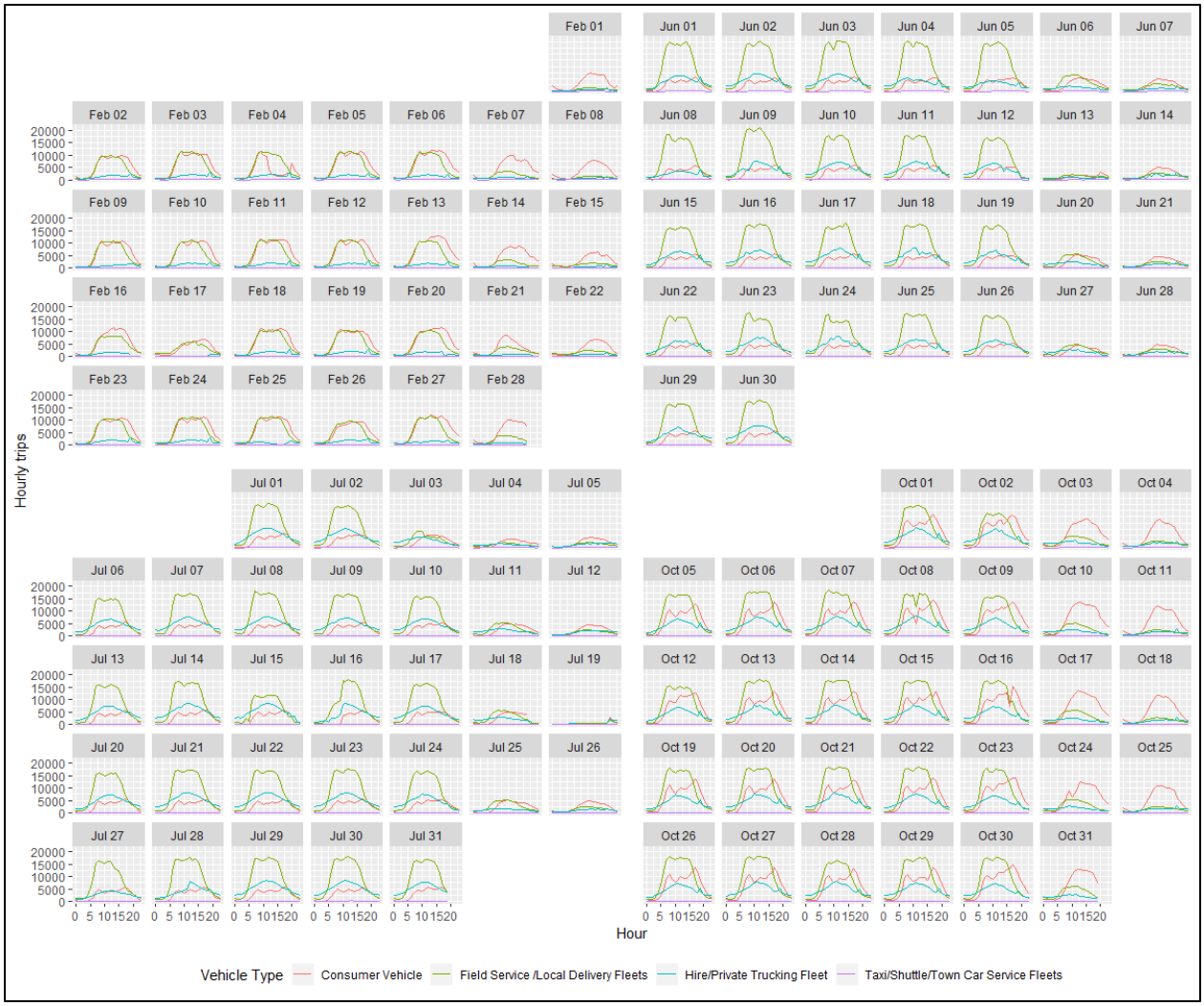


Figure 15. Graphs. Trip distribution by hour for different vehicle types.

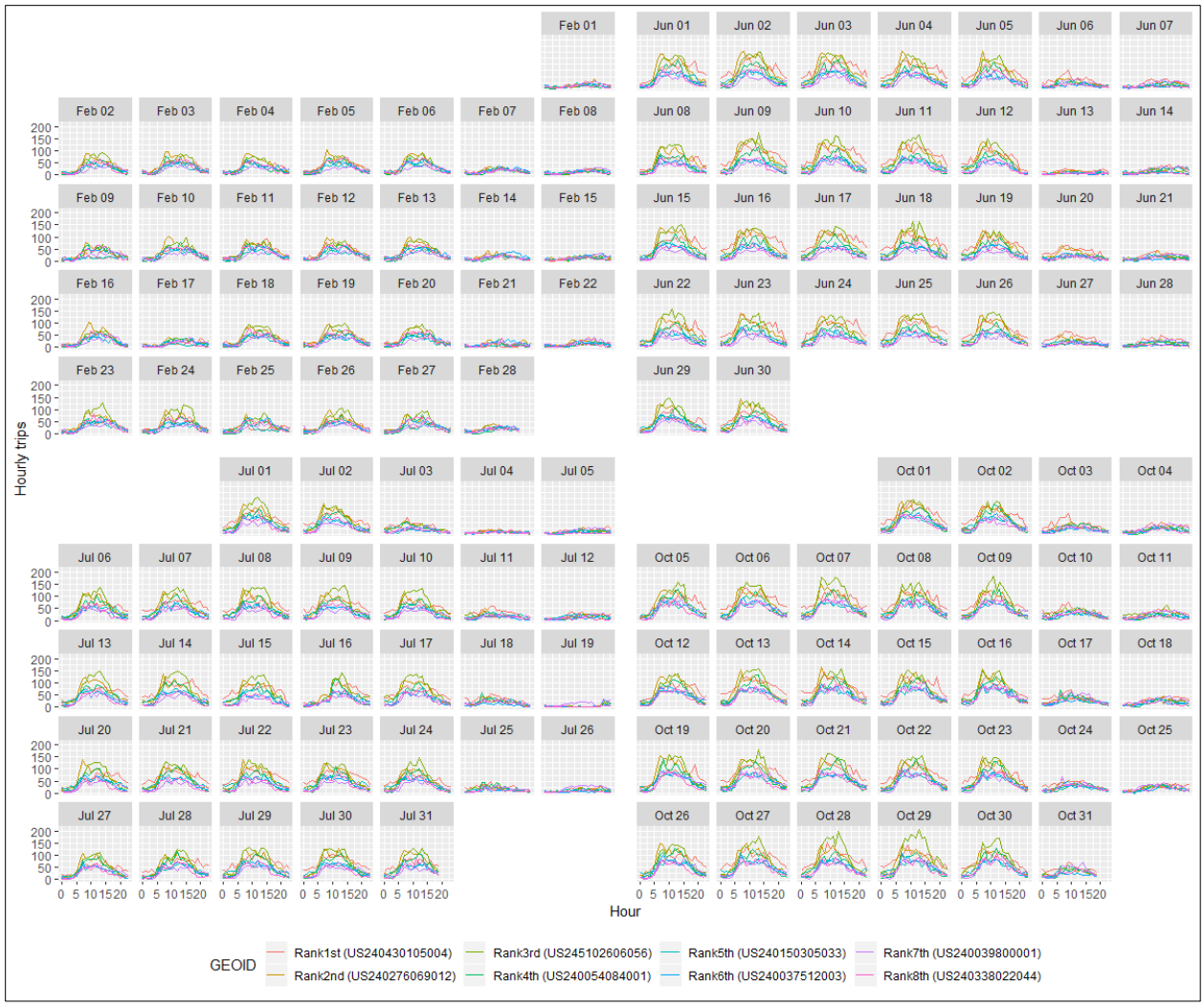


Figure 16. Graphs. Trip distribution by hour for top eight O-D generator block groups.

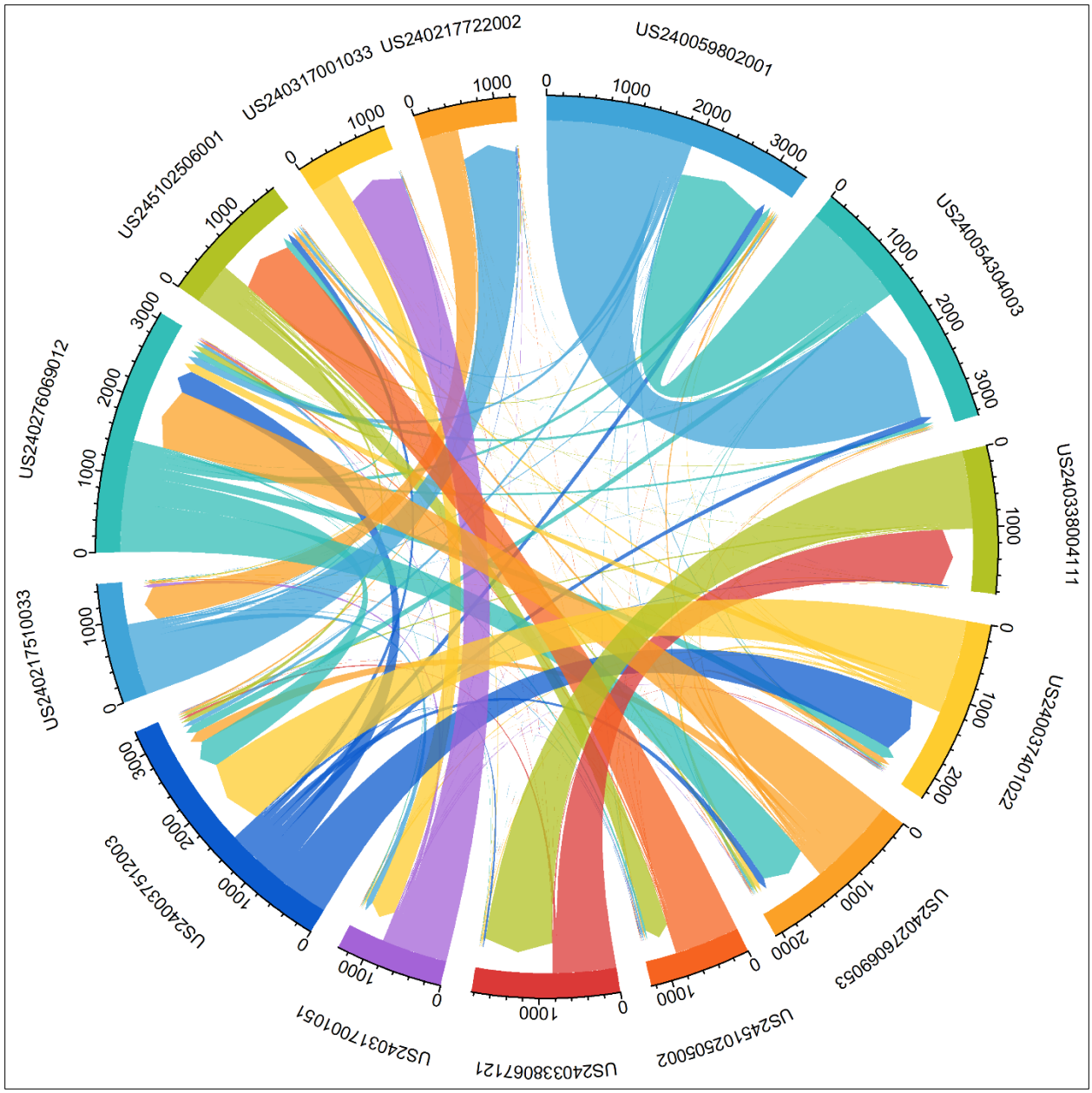


Figure 17. Chord diagram.[1] Top O-D generator block groups (February 2015).

[1] A chord diagram illustrates the interrelationships between individuals (for this study, each block group is considered as an individual spatial unit). The associations between individuals are used in displaying commonality of information or interest. A chord diagram is useful in comparing the similarities and patterns within a dataset. Nodes are arranged in a circular form, with the associations between points connected to each other either with arcs or curves. The assigned values to each connection are represented proportionally by the size of each arc. The color is used in grouping the data into different categories that aid in making comparisons and distinguishing groups.

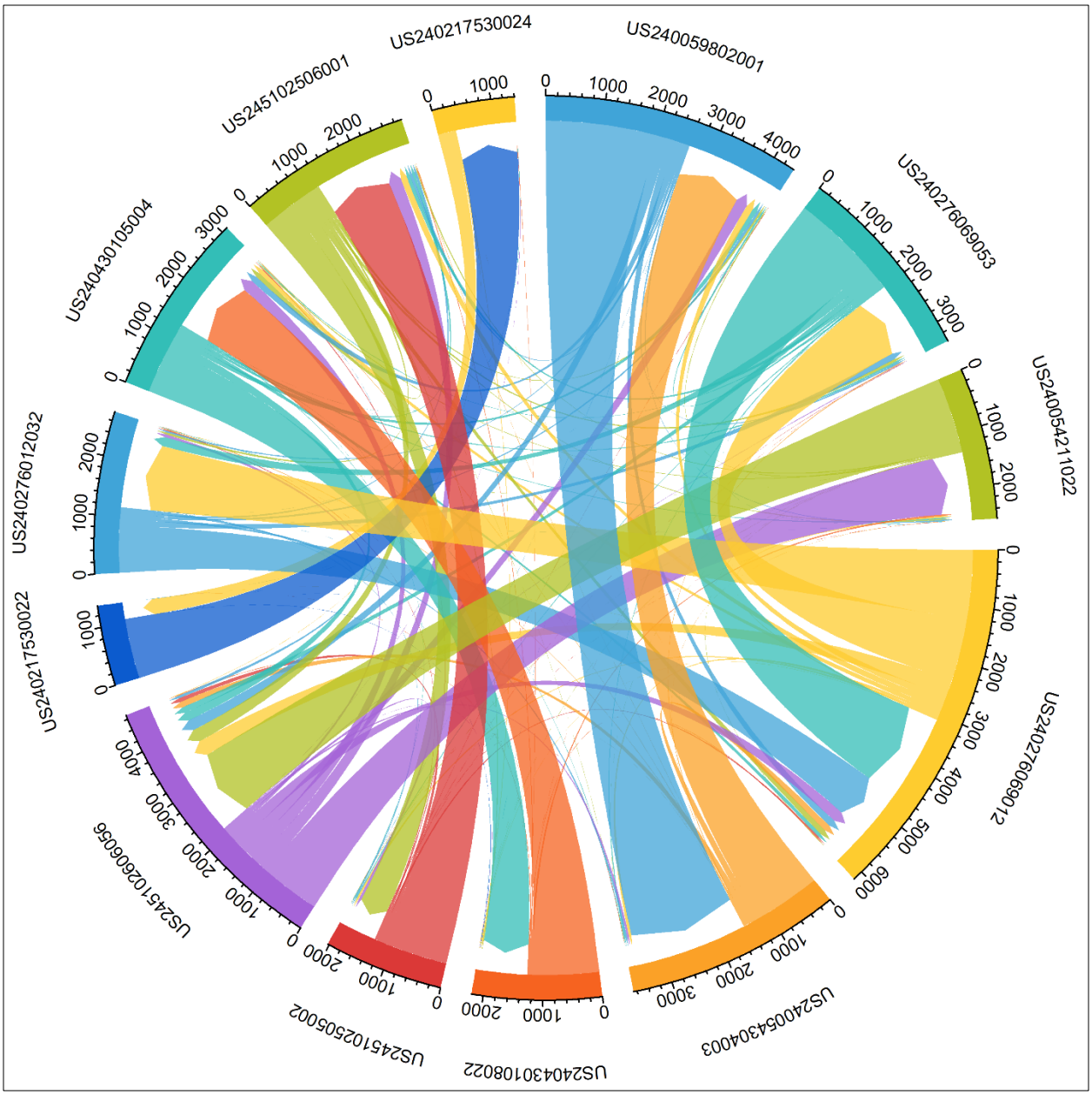


Figure 18. Chord diagram. Top O-D generator block groups (June 2015).

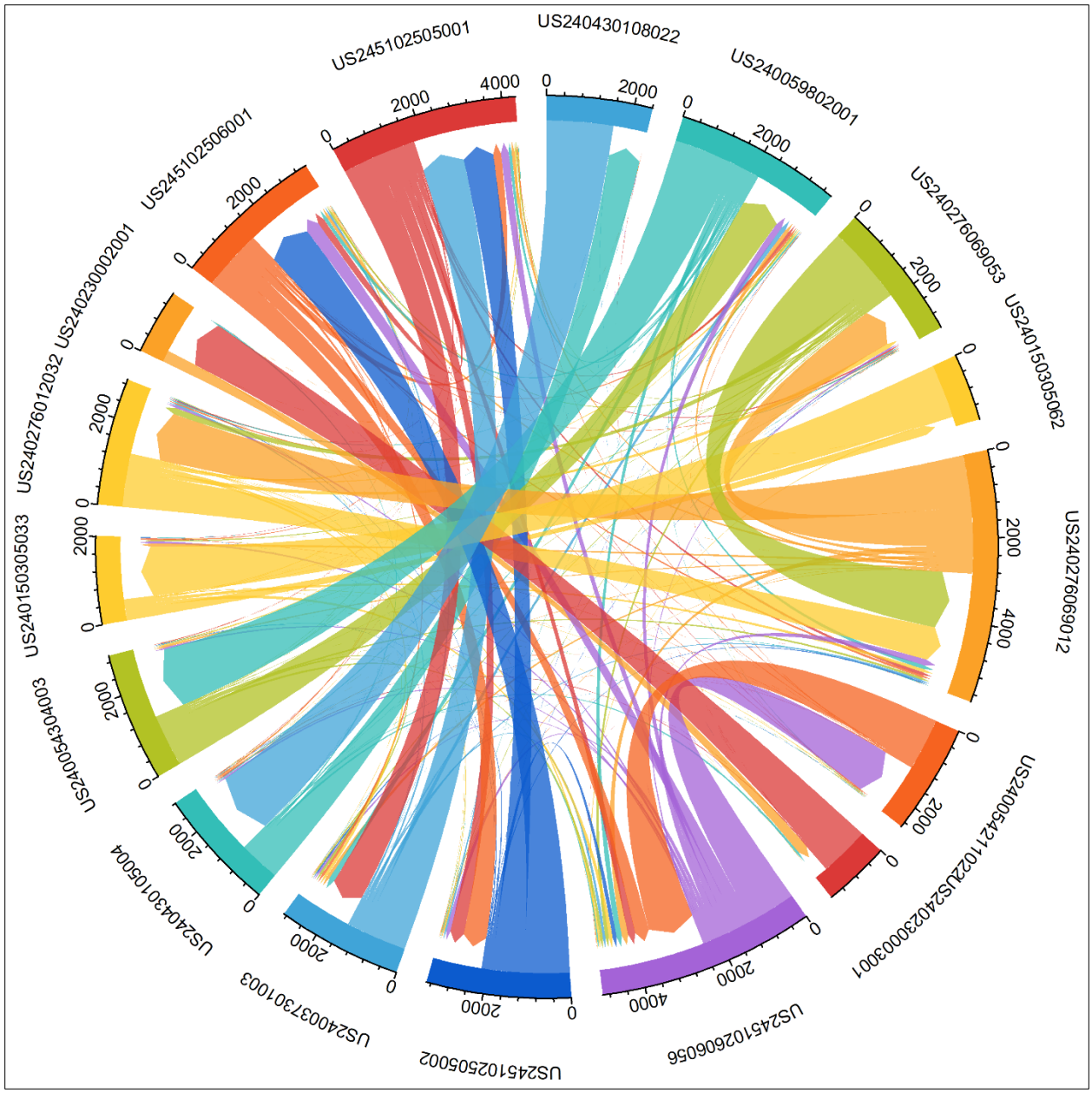


Figure 19. Chord diagram for the top O-D generator block groups (July 2015).

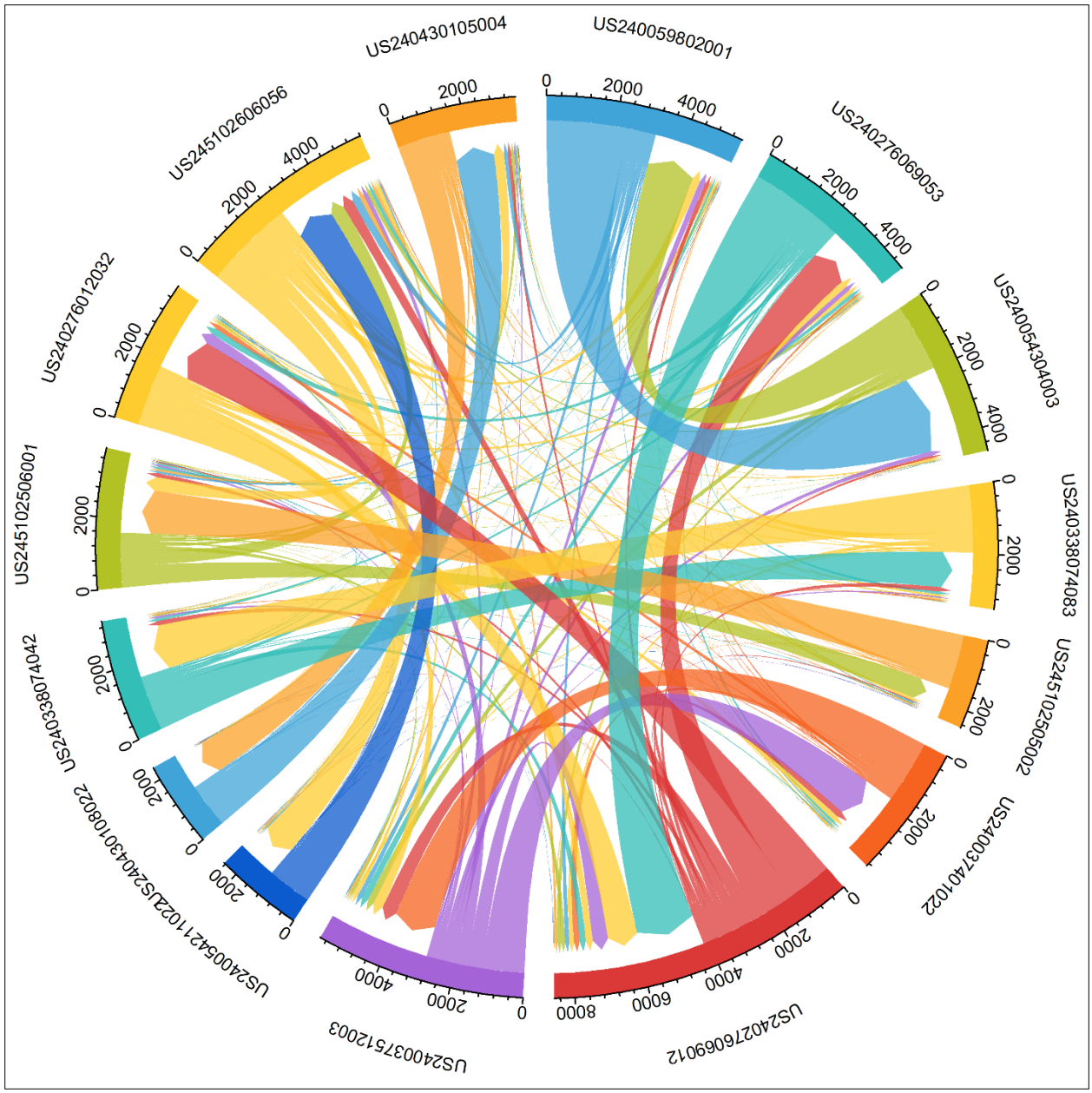


Figure 20. Chord diagram. Top O-D generator block groups (October 2015).

Table 13. Top 40 Rules

Antece.	Conse.	S	C	L	Counts
Total_Jobs_WAC=(> 3778.5], HH_F=(386.5; 659.5]	TQ=5	0.01	1.00	5.00	35
Total_Jobs_WAC=(> 3778.5], Popu=(1366.5; 1978.5]	TQ=5	0.01	1.00	5.00	34
Total_Jobs_WAC=(> 3778.5], Popu=(1978.5; 3311.5]	TQ=5	0.01	1.00	5.00	31
Total_Jobs_WAC=(> 3778.5], HH_MedInc=(> 98467]	TQ=5	0.01	1.00	5.00	28
Total_Jobs_WAC=(> 3778.5], HH_MedInc=(53774; 98467]	TQ=5	0.02	0.98	4.92	61
Total_Jobs_WAC=(> 3778.5]	TQ=5	0.03	0.96	4.82	109
Popu=(> 3311.5], HH=(> 1252.5], HH_F=(> 934.5]	TQ=5	0.01	0.91	4.53	29
Total_Jobs_WAC=[< 21.5], HH=[< 264.5]	TQ=1	0.02	0.90	4.52	75
Total_Jobs_WAC=[< 21.5], HH_F=[< 230.5]	TQ=1	0.04	0.88	4.41	157
Total_Jobs_WAC=(833; 3778.5], HH=(> 1252.5]	TQ=5	0.01	0.88	4.40	22
HH=(> 1252.5], HH_MedInc=(> 98467]	TQ=5	0.01	0.88	4.38	21
Total_Jobs_WAC=[< 21.5], HH_MedInc=[< 53774]	TQ=1	0.04	0.82	4.08	137
Popu=(> 3311.5], HH_F=(> 934.5]	TQ=5	0.01	0.81	4.05	34
Total_Jobs_WAC=[< 21.5], Popu=(623.5; 1366.5]	TQ=1	0.05	0.81	4.03	171
Total_Jobs_WAC=(833; 3778.5], HH_F=(659.5; 934.5]	TQ=5	0.01	0.80	4.00	52
Total_Jobs_WAC=[< 21.5], HH=(264.5; 477.5]	TQ=1	0.04	0.80	3.99	155
Total_Jobs_WAC=(833; 3778.5], HH=(972.5; 1252.5]	TQ=5	0.02	0.79	3.97	58
Popu=(> 3311.5], HH=(> 1252.5]	TQ=5	0.01	0.79	3.95	45
Total_Jobs_WAC=(21.5; 61.5], Popu=[< 623.5]	TQ=1	0.01	0.78	3.91	36
Total_Jobs_WAC=(833; 3778.5], HH_MedInc=(> 98467]	TQ=5	0.03	0.78	3.89	102
Total_Jobs_WAC=(61.5; 120.5], Popu=[< 623.5]	TQ=1	0.01	0.78	3.89	28
Total_Jobs_WAC=(833; 3778.5], Popu=(> 3311.5]	TQ=5	0.01	0.76	3.82	26
HH_F=(> 934.5]	TQ=5	0.01	0.76	3.80	35
Total_Jobs_WAC=(833; 3778.5], Popu=(1978.5; 3311.5]	TQ=5	0.03	0.76	3.80	123
Total_Jobs_WAC=(833; 3778.5], HH_F=(386.5; 659.5]	TQ=5	0.04	0.75	3.77	141
Popu=(> 3311.5], HH_MedInc=(> 98467]	TQ=5	0.01	0.73	3.66	30
HH=(> 1252.5]	TQ=5	0.02	0.72	3.59	61
Total_Jobs_WAC=[< 21.5]	TQ=1	0.07	0.72	3.58	275
Total_Jobs_WAC=(833; 3778.5], HH_MedInc=(53774; 98467]	TQ=5	0.05	0.71	3.55	179
Total_Jobs_WAC=(833; 3778.5], HH=(609.5; 972.5]	TQ=5	0.03	0.71	3.54	131
Total_Jobs_WAC=(21.5; 61.5], HH=[< 264.5]	TQ=1	0.01	0.71	3.54	46
Total_Jobs_WAC=(833; 3778.5], Popu=(1366.5; 1978.5]	TQ=5	0.03	0.71	3.54	99
Total_Jobs_WAC=(833; 3778.5]	TQ=5	0.10	0.69	3.45	362
Total_Jobs_WAC=(61.5; 120.5], HH=[< 264.5]	TQ=1	0.01	0.67	3.33	30
Popu=[< 623.5], HH_MedInc=[< 53774]	TQ=1	0.02	0.66	3.31	94
Total_Jobs_WAC=(21.5; 61.5], HH_F=[< 230.5]	TQ=1	0.02	0.64	3.18	89
Popu=[< 623.5], HH=[< 264.5]	TQ=1	0.03	0.62	3.08	122
HH=[< 264.5], HH_MedInc=[< 53774]	TQ=1	0.03	0.60	3.01	98
Total_Jobs_WAC=(405.5; 833], HH=(972.5; 1252.5]	TQ=4	0.01	0.58	2.91	25
HH=(972.5; 1252.5], HH_MedInc=(> 98467]	TQ=5	0.01	0.56	2.81	32

Notes: Antece.= Antecedent, Conse.= Consequent, S= Support, C=Confidence, and L=Lift