Patterns of Origin Destination Distributions

Rules Mining using Massive GPS Trajectory Data

Ruihong Wang
Texas A&M Transportation Institute
Bryan, TX 77807, USA
r-wang@tti.tamu.edu

Subasish Das
Texas A&M Transportation Institute
Bryan, TX 77807, USA
s-das@tti.tamu.edu

Abhisek Mudgal
Indian Institute of Technology (BHU)
Varanasi, India
abhisek.civ@iitbhu.ac.in

Abstract

Understanding the spatio-temporal distribution of road traffic conditions is imperative to the design of suitable countermeasures for congestion reduction. The amount of crowd sourced data from private mobility companies is rapidly expanding to include data sources that were previously not available; this data can assist in developing new approaches and algorithms to understand travel patterns and origin-destination (O-D) distributions. We acquired four months (February, June, July, and October of 2015) of INRIX Waypoint (global positioning system or GPS trajectories) data that includes vehicle trips from various sources. To determine the association between demographic, economic, and land use information and O-D patterns, we acquired Census block group level data from American Community Survey (ACS), and block level economic data from Longitudinal Employer-Household Dynamics (LEHD). We provided insights on the relationship between demographic information and O-D patterns by Census spatial unit 'block group'. Based on multi-source data, we used classification-based association rules mining to determine key association patterns.

1 Introduction

Roadside interview survey has been considered as the common method to determine the measures of roadway origin and destinations (O-D). The crowd sourced data from the private companies can assist in developing new algorithms to understand O-D measures. Although these data sources are not without limitations, they are proving valuable insights in determining many traffic operational exposures. Many studies performed analysis on the O-D patterns using different regression models and machine learning algorithms. In general, regression models ignore subgroup or clusters in the data sets. On the other hand, machine learning models can provide better predicted O-D demands with little or no valuable insights about the data. Rules based models can identify the subgroup effects from complex and large data sets without imposing any prior assumptions. With the increasing attention on the O-D related measures, a new study design is proposed in this paper for the identification interesting patterns in O-D distributions.

Third part data providers (e.g., INRIX, HERE, Strava) are increasingly producing real-time traffic flow data for both motorized and non-motorized trips. These real-time traffic flow trajectories are collected from many digital sources (with inclusion of location and time-stamp of the trajectory) that travelers intentionally or unintentionally share through smart cell phones, navigation devices, fleet management systems, Bluetooth devices, and other regular cellular services. To mitigate the current research gap, we acquired four months (February, June, July, and October of 2015) of INRIX Waypoint or vehicle trajectory data that includes vehicle trips in Maryland. We acquired various area level demographic and economic information from different data sources such as ACS and LEHD. The analysis was conducted at Census block group level. The classification based associated rules mining was applied to identify the hidden trends from the complex and large O-D data set of Maryland.

1.1 Research Questions

We developed several key research questions in our study design:

- **RQ1** How do O-D patterns differ by different temporal units (e.g. hour) and vehicle types?
- **RQ2** How can we improve the conventional O-D visualizations?
- **RQ3** What are key contributors for different O-D patterns?

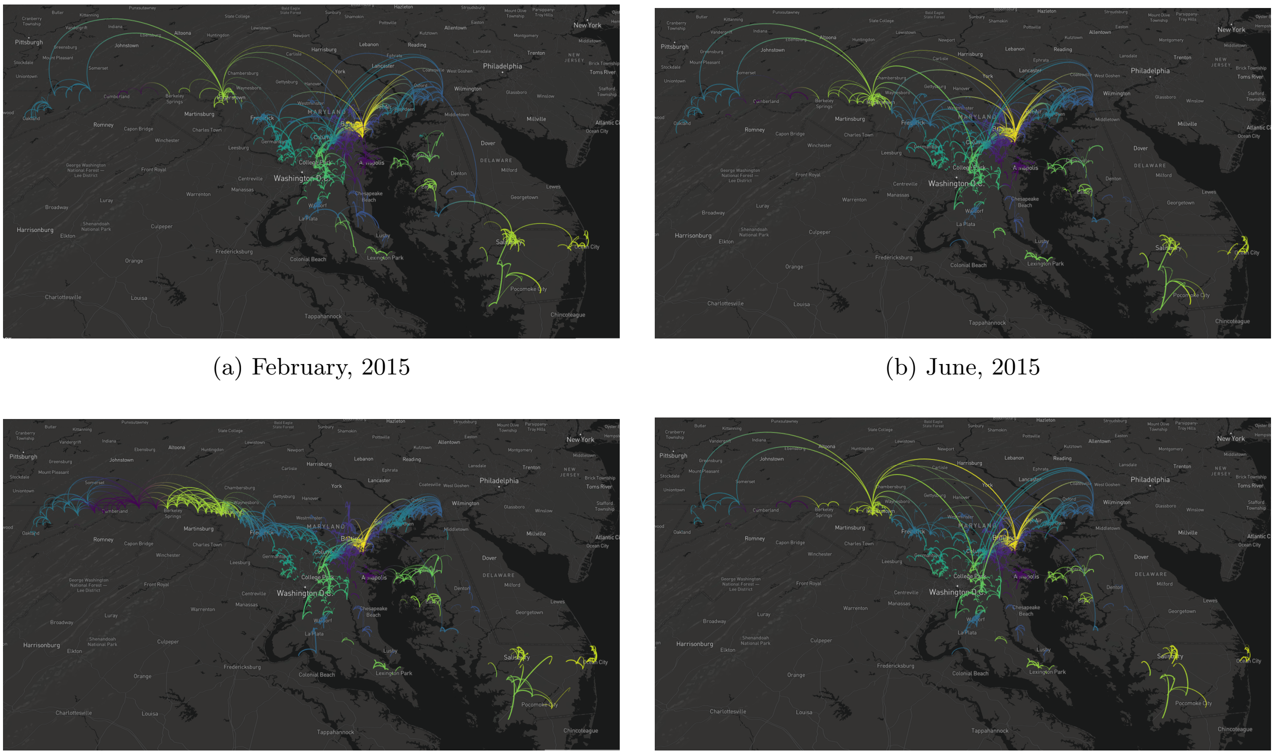
2 Literature Review

Many recent studies have aimed to determine O-D patterns using crowd sourced or vendor provided real-time data. Some key studies in this area are briefly described below.

In a study conducted by Sana et al. [Sana18], researchers used information from the Google Aggregated and Anonymized Trips (AAT) to develop a machine learning model and generate the 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. In another study, Ma et al. [Ma18] developed a data-driven structure to estimate daily dynamic O-D. To accomplish this, researchers used high-granular traffic frequency and speed data spanning over 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. [Fan18] 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 were useful for the real-world work associated with the O-D estimation. In another study, Ge et al. [Ge16] used aggregated data of mobile phone traces to estimate work-related trips and developed a method to estimate O-D matrices based on maximum entropy principle. The researchers calculated the trip production and attraction by using a non-linear programming problem; they then used a matrix fitting problem to distribute trips to each O-D pair. Iqbal et al. [Herijanto05] developed a combined fluid analogy and singly constrained gravity model to predict the trip distribution in developing country. The new modified gravity model outperformed the conventional singly constrained gravity model and was suitable for the limited quality and quantity of data.

There are a large number of studies regarding association rules mining in transportation. Geurts et al. [Geurts03] researched on the one-year traffic accident data in Flanders, Belgium. The result showed that human and behavioral aspects play an important role in understanding the traffic safety problems. Lipan et al. [Lipan10] designed a data acquisition system for the GPS data from the bus network of Cluj-Napoca, Romaniato. By conducting an association rule mining study, the relation between road features, time information, bus load etc. and the speed variations was mined. Lee et al. [Lee16] investigated the round-trip car sharing system in the city of Cagliari, Italy. The outcome illustrated the connection between various combination of round-trip operations characteristics. Das et al. [Das19] studied the hit and run pattern by association rules mining, finding the key features which contributes most to the occurrences of these crashes.

Furthermore, two recent studies used Maryland INRIX Waypoint data; one study estimated vehicle miles traveled [Fan19, Nasri19] by the big data analysis engine Apache Spark and another determined the reliability of routing decisions made by the truck drivers [Kong18]. However, application of data mining in O-D measures are very limited.



(a) February, 2015

(b) June, 2015

(c) July, 2015

(d) Oct, 2015

Figure 1: OD patterns (top 2000) by Month

3 Approach

3.1 Data Sources

To better understand the patterns of O-D measures, we examined a large array of possible data sources. We finally selected three separate databases to perform the analysis:

- Maryland INRIX Waypoint trip data
- Census block group level ACS 2013-2017 data
- Census block level LEHD 2015 data

3.1.1 Maryland INRIX trip data

We collected Maryland INRIX trip data to perform the analysis. The acquired data for four months (February, June, July and October 2015) contain three types of monthly files in comma separated value (CSV) format: TripRecordsReport (trip data), TripRecordsReportWaypoints (Waypoint data), and TripsRecordsReportProviderDetails (information on trip data providers). The acquired data set contains:

- 19,690,402 trip information
- 1,376,720,203 Waypoint (geospatial point recorded by GPS devices that represents the location of the recorded vehicle) information
- 5,451,095 unique device identifications
- 148 data providers (45 providers)
  - 3 providers account for 52 percent of trips
  - 18 providers account for 99 percent of trips

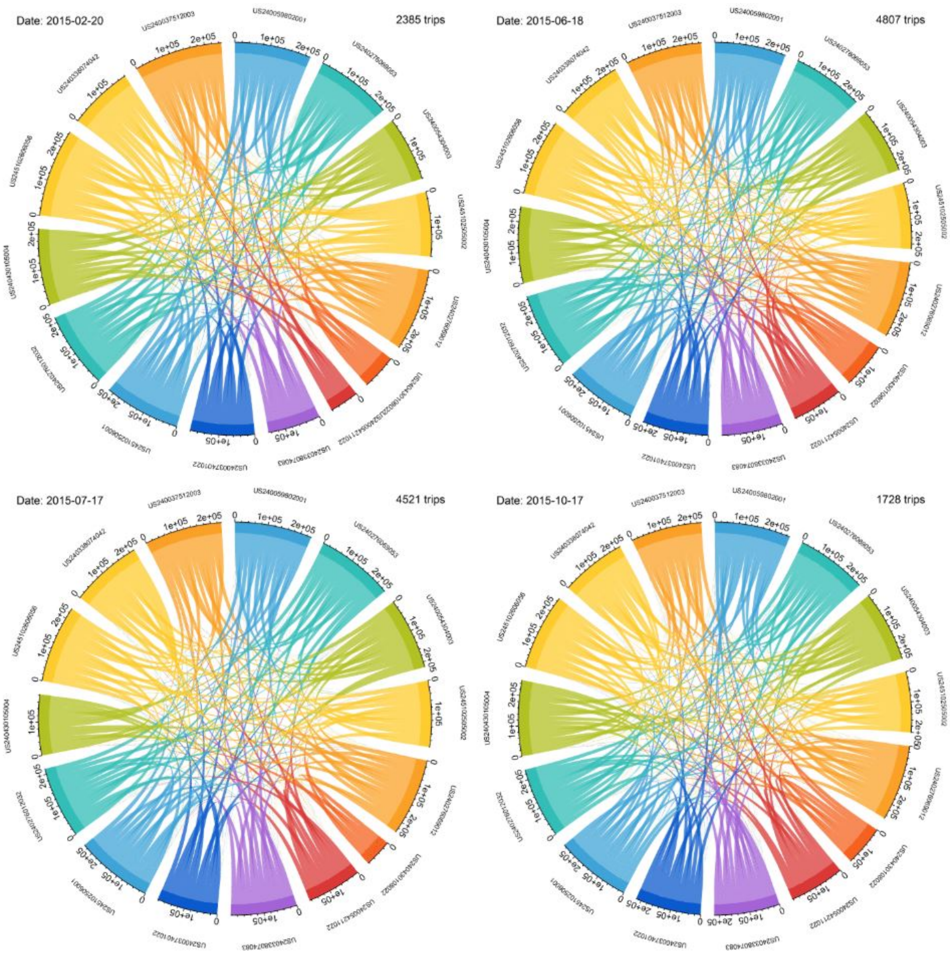


Figure 2: O-D patterns by Day

- 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 to 14000 lbs.), 2) medium duty truck/vans (14001 to 26000 lbs.), and 3) heavy-duty truck (greater than 26000 lbs.).

3.1.2 ACS Data

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

3.1.3 LEHD Data

We also used the Census block data of LEHD Origin-Destination Employment Statistics (LODES) 2015 data. The LODES files contain data for Residential Area Characteristics (RAC) and Work Area Characteristics (WAC). This data set provides insights on economic condition of the residents in Census block. As our analysis was limited to block group level, these measures are calculated for the block group level.

3.2 Data Integration

We used ArcGIS and several open source software R packages to develop the required database for analysis. The data integration steps are following:

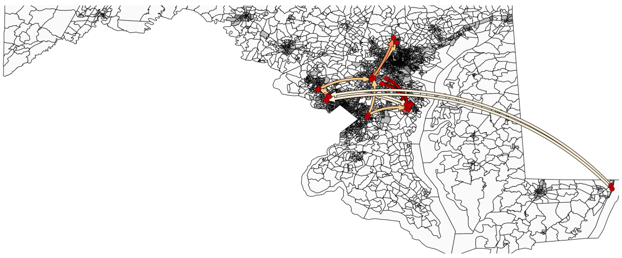
- The collected ACS data contains different tables such as age, gender, income, and household. We selected key demographic and relevant information for this analysis.
- The block level RAC and WAC data were merged into block group level.
- We used QGIS tool to spatially merge Waypoint O-D data with census block groups level information. Later, several separate databases were developed based on the following: 1) Monthly O-D data by block groups, 2) Daily O-D data by block groups, 3) Hourly O-D data by block groups, and 4) O-D data by vehicle types.

4 Evaluation

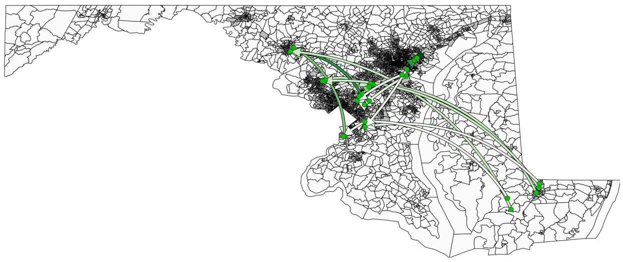
4.1 O-D Distribution

Insights on OD patterns and demand can improve transportation planning and management. Holidays, game days, crashes or incidents, events, land use pattern changes, and many other factors can affect the OD measures that varies by hour, day, month, weekend-weekday, or any other related temporal patterns. It is important to determine the key O-D matrices for different scenarios. Three different temporal patterns were used in this analysis: 1) all 24 hours, 2) morning peak (6-10 am, Mon-Fri), and 3) evening peak (4-8 pm, Mon-Fri).

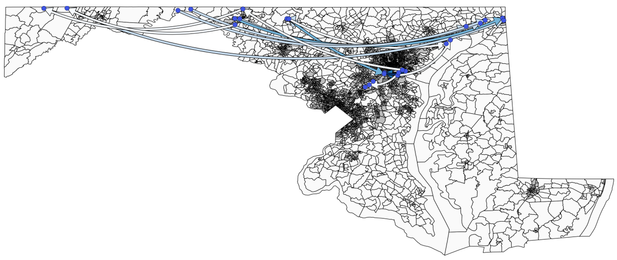
To answer RQ1 and RQ2, we developed several different data visualization tools to provide additional insights on the data. Figure 1 illustrates the O-D patterns (top 2000) by month. The color of the nodes are based on the distances. The overall trends look similar for three months (February, June, and October). However, the patterns of July are significantly different from the other three months. Instead of a conventional two-dimensional plots, three-dimensional plots can show the complex nature of O-D patterns in an easier manner.



(a) Consumer vehicles (24 hours)

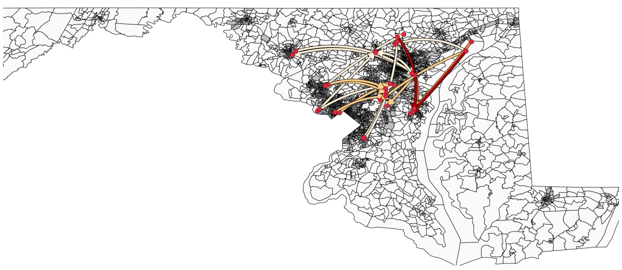


(b) Local delivery fleets (24 hours)

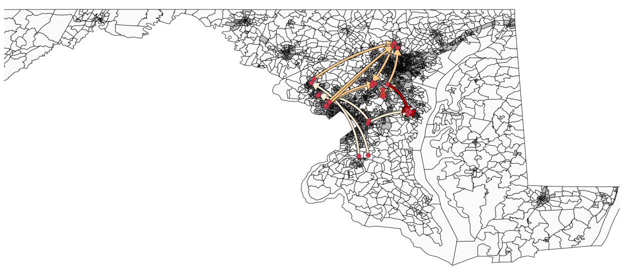


(c) For-hire/private trucking fleets (24 hours)

(d) Taxi/shuttle/town car service fleets (24 hours)



(e) Morning Peak for consumer vehicle (6am- 10am, Mon-Fri)



(f) Evening Peak for consumer vehicle (4pm- 8pm, Mon-Fri)

Figure 3: OD patterns by vehicle type and hour

A chord diagram, resourceful in comparing the similarities and patterns within a data set, illustrates the interrelationships between individuals (for this study, each block group is considered as an individual spatial unit). The associations between individual block groups are used in displaying commonality of information or interest. Nodes (for this study, block groups) are arranged in a circular form, with the association between points connected to each other either with arcs or curves. The assigned values to each of the 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 2, the static version of the interactive chord diagram tool, indicates that the in-betweenness between the O-D pairs is significant. However, an interactive version of this tool is more useful due to its ability to show the O-D change patterns from one block group to another block group in real-time. This kind of graphics are useful in deciphering more insights from complex O-D patterns. For example, if a particular block group has several significant temporal trends, necessary steps can be taken to alleviate the congestion or other relevant issues. The is anticipated that the inclusion of chord diagrams of O-D patterns in the network fundamental diagram (NFD) can increase the performance in traffic monitoring and control.

Transportation networks are complex in nature due to the heterogeneous nature of vehicle flow. One unique feature of the acquired database is that it contains four vehicle types: consumer vehicles, local delivery vehicles, and for-hire or private trucking fleets, and, taxi/shuttle/town car service fleets. Figure 3 shows the top 20 O-D pairs for the four different vehicle type. The O-D patterns of For-hire/private trucking fleets show the highest distances between the origins and destinations. Taxi or shuttle trips are rather short and mostly in between the major urban locations. The top O-D pairs also differ by different temporal clusters as shown in Figure 3 e and f.

5 Results

5.1 Rules Mining

Although many transportation research studies used association rules mining [Das19, Das19, Saunders19], usage of this technique in O-D pattern analysis was not conducted before. We used classification-based association rules to develop the rules that show patterns of land use and associated factors relating to O-D distribution. The association rule can be represented as $Antecedent(A) \rightarrow Consequent(B)$, where both of them are disjoint itemsets. Here,

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

;

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

and Support of the rule can be written as

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

The measure of reliability for a generated rule is known as confidence

$$C\left(A \rightarrow B\right) = \frac{S\left(A \rightarrow B\right)}{S(A)}$$

.

The lift is a measure that represents the ratio of confidence and expected confidence.

$$L\left(A \rightarrow B\right) = \frac{S\left(A \rightarrow B\right)}{S\left(A\right).S(B)}$$

A lift value greater than one shows positive interdependence between A and B, while a value smaller than one refers negative interdependence, and a value of one signifies independence.

We selected several significant variables by performing random forest algorithm. The selected variables (block group level) for the 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). The top 40 rules with high lift values are listed in Table 1. Average O-D measures per block group are divided into five classes based on the quantile percentages of the O-D frequencies: TQ= 1 [1-20 percent], TQ= 2 [21-40

Table 1: Top 40 Rules.

Antecedent	Consequent	Support	Confidence	Lift	Counts
Total_Jobs_WAC=(>3778.5], HH_F=(386.5; 659.5]	TQ=5	0.01	1	5	35
Total_Jobs_WAC=(>3778.5], Popu=(1366.5; 1978.5]	TQ=5	0.01	1	5	34
Total_Jobs_WAC=(>3778.5], Popu=(1978.5; 3311.5]	TQ=5	0.01	1	5	31
Total_Jobs_WAC=(>3778.5], HH_MedInc=(>98467]	TQ=5	0.01	1	5	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	0.9	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	4.4	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	0.8	4	52
Total_Jobs_WAC=[<21.5], HH=(264.5; 477.5]	TQ=1	0.04	0.8	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	3.8	35
Total_Jobs_WAC=(833; 3778.5], Popu=(1978.5; 3311.5]	TQ=5	0.03	0.76	3.8	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	0.1	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	0.6	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

percent], TQ= 3 [41-60 percent], TQ= 4 [61-80 percent], and TQ=5 [81-100 percent]. For example, TQ=5 indicates the block groups with top 20 percent percent of the O-D trips. The findings show that higher number of populations, WAC, and households are associated with higher number of O-D trips (TQ=5). In several top rules, it was found that block group with high median household income is associated with high number of O-D trips. The rules provide several breakpoints of the variable clusters to determine the top rules.

Below, we discuss each of the research questions presented in Section 1.1.

RQ1: We observe in Figure 1 that the majority of the monthly O-D patterns are similar. The daily O-D patterns by the top 20 O-D generators indicate that the range of daily trips in the block groups differs in values to some extent.The trip distribution by vehicle types also differ in patterns. The top For-hire/private trucking fleets O-D generators have the most distance measures than the other three vehicle type.

RQ2: Data visualization is an integral part of data-driven analysis. A carefully developed data visualization delivers a powerful cognitive advantage to its viewers. We aimed to develop well-designed and interactive visualization to provide the end users a spontaneous and reactive impression and help them to cut through the clutter of complex data structures. We used different data visualization libraries such as D3.js and deck.gl to develop O-D visualizations with important matrices. Conventional static visualization often limits the scope of understanding the massive patterns of these OD patterns. The interactive patterns of the visualizations are able to provide more flexibility to understand the trends and patterns over different temporal units. For example, the users can limit the temporal units to identify the clusters of major OD generators from these visualizations. This study focuses on insightful visual graphics, which has been rarely studied due to the limited data and advanced graphical tools. Incorporation of these tools in the state-of-the-art traffic flow management tools can improve traffic monitoring and control in real-time.

RQ3: We observe that rules mining provided top 40 association rules that do not need any kind of prior assumptions. It is important to note that additional rules can be examined based on the requirements. The rule numbers are limited here for easier interpretation. As earlier studies did not explore rules mining on O-D data sets, the findings from this study can shed some lights in this unexplored research domain. The current analysis is limited to only 40 top rules. However, more insightful rules can be developed by using different thresholds of support and confidence. Some of the key insights from these rules are following:

- High WAC and population and housing units are associated with top 20 percent of the O-D trips. For example, Rule 1: Total_Jobs_WAC=(>3778.5], HH_F=(386.5; 659.5] has a lift value of 5 with 35 events. The lift value of Rule 1 infers that the proportion of trips associated with this rule was 5 times the proportion of other antecedents associated with top 20 percent of the O-D trips.

- Block groups with lower WAC and housing units are associated with low number of trips (TQ=1). It is intuitive because trips are usually associated with number of jobs in an area.

- Block groups with higher median incomes are also associated with top 20 percent of the O-D trips. This finding indicates that income thresholds are associated with the O-D trip patterns and frequencies.

6 Conclusion and Future Work

Real measures associated with O-D patterns is an indispensable component for modeling transportation networks. Third party data providers are increasingly producing real-time traffic flow data. With the help of vehicle trajectory data, it is possible to provide intuitive measures related to O-D patterns. Our study design was applied to 19.8 million raw GPS vehicle trajectories collected in 2015 in Maryland. The results showed that distribution of O-D measures for different vehicle types vary by hour and by month. The top 20 block groups with high average O-D measure contribute approximately 8.5 percent of trips. The generated top 40 rules provide several breakpoints of the key variables. The findings show that number of jobs, population, housing units, and median income are associated with high number of trip generations. For many cases, city planners do not have adequate resources in translating massive data sets into important performance measures. The breakpoints determined in the rules can be easily used for developing appropriate strategies in a way to improve the travel experiences of the roadway users. One unique contribution of the current study is that it used a data mining method that was less explored in O-D analysis studies. Additionally, this study has developed a methodology that can be replicated in other trajectory and O-D measure data sets to determine intuitive knowledge more efficiently.

Our study is not without limitation. The rules mining applied in this study is limited to monthly O-D data. Future studies should explore the rules with more granular levels of data (for example, daily or hourly). Additionally, our study is limited to only four months of data from 2015. A more comprehensive data is required to develop a robust framework of O-D measures.Future studies can explore the current limitations to determine additional insights from O-D databases.

References

[Das19] S Das, A Dutta, R Avelar, K Dixon, X Sun, M Jalayer. Supervised association rules mining on pedestrian crashes in urban areas: identifying patterns for appropriate countermeasures. International Journal of Urban Sciences, 2019.

[Das19] S Das, A Dutta, X Sun. Patterns of rainy weather crashes: Applying rules mining. Journal of Transportation Safety & Security, 2019, 1-23.

[Das19] Das S., X. Kong, and I. Tsapakis Hit and run crash analysis using association rules mining. Journal of Transportation Safety & Security(2019), 1—20.

[Fan18] 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.

[Fan19] 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.

[Ge16] 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.

[Geurts03] Geurts k., G. Wets, T. Brijs, and K. Vanhoof Profiling of high-frequency accident locations by use of association rules. Transportation research record 1840, 1 (2003), 123—130.

[Herijanto05] Herijanto W. and N. Thorpe Developing the Singly Constrained Gravity Model for application in developing countries. Journal of the Eastern Asia Society for Transportation Studies (2005),1708—1723.

[Kong18] Kong X., W. Eisele, Y. Zhang, and D. Cline. Evaluating the Impact of Real-time Mobility and Travel Time Reliability Information on Truck Driver Routing Decisions. Transportation Research Record: Journal of the Transportation Research Board. Vol. 2672, 2018, pp. 164—172.

[Lee16] Lee D., L. Quadrifoglio , I. Meloni Discovering relationships between factors of round-trip car sharing by using association rules approach. Procedia engineering161 (2016), 1282—1288.

[Lipan10] Lipan F., and A. Groza Mining traffic patterns from public transportation GPS data. In Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing. IEEE, 123—126.

[Ma18] 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.

[Nasri19] Nasri A., L. Zhang , J. Fan , K. Stewart , H. Younes , ChengFu, and Jessberger S. Advanced Vehicle Miles Traveled Estimation Methods for Non-Federal Aid System Roadways Using GPS Vehicle Trajectory Data and Statistical Power Analysis. Transportation Research Record (2019),0361198119850790.

[Sana18] Sana, B., J. Castiglione, D. Cooper, and D. Tischler. Using Google Passive Data and Machine Learning for Origin-Destination Demand Estimation. Transportation Research Record, Vol 2672, Issue 46, 2018, pp 73—82.

[Saunders19] William John Saunders, Saleh R. Mousa, Julius Codjoe. Market basket analysis of safety at active highway-railroad grade crossings. Journal of Safety Research, 2019.