Improving Urban Mobility Through Urban Analytics Using Electronic Smart Card Data

Mayuree Binjolkar, Daniel Dylewsky, Andrew Ju, Wenonah Zhang, Mark Hallenbeck
Data Science for Social Good-2017,University of Washington

Abstract

Fare collection technology has advanced significantly over the past few years and the electronic smart card fare collection system is a highly efficient way of collecting fare. It helps in reduction of costs for cash handling and processing and provides the transit users with a flexible and convenient way to pay for their trips. In addition, implementation of this technology helps in collection of valuable data that can been used to assess some of the transit related travel patterns in a geographical area. This study aims to use the electronic smart card data to understand various transit related travel patterns in Seattle. It uses the ORCA data to examine the true transit ridership behavior of the total population by studying the spatiotemporal incompleteness in the data itself. In addition, this data is used to analyze the transfer patterns of the transit users as this can help the transit agencies to improve their route planning for public transit and save operational costs. Several spatial analysis are also carried out for University District neighborhood in Seattle and the change in transit activity related patterns from 2015 to 2016 is studied.

Introduction

In 2016, TomTom ranked Seattle as the fourth most congested city in States. The extra travel time was 152 hours per year, which had grown at a pace of 3 percent from the year before. Because of the geographical features and expanding population in Seattle, moving more than seven hundred thousand people through a narrow strip of hilly land flanked by waters could be a world-class challenge. A sustainable and effective transportation solution is urgently in need, and public transportation can offer significant mitigation of the traffic blockage according to many urban planning studies.

Back to 2009, seven transit companies in the greater Puget Sound Region adopted a uniform electronic fare payment system called One Regional Card for All (ORCA), which provided an efficient and cost-saving alternative to the paper ticket. At the same time, this electronic media collected a massive volume of data, which directly reflected the transit riders' behavior. Only in last year, ORCA Board granted University of Washington the access to two, nine-week private datasets in the hopes that the Data Science for Social Goods Summer program can explore possible modifications on current public transit network. The ORCA data contained information on 23 millions actual boardings. Compared to the traditional survey method used in transportation system planning, this database presented a more comprehensive look at the transit

usage in Puget Sound region no matter whether a trip started in midnight or at a remote corner of the town. Even though many cities around the world have implemented automatic fare collection (AFC) systems, the generated information were usually unavailable for the transportation research centers in universities to take up academic studies, which resulted in a lack of mature methodologies on utilizing this type of modern public transit data metrics to model human travel behavior. In 2010, Korea Transport Institute analyzed travel time and transfer trip pattern using the data collected from their AFC system, which required both tap-on and tap-off from transit users. However, the ORCA data only required tap-ons and the AFC system could not register the exit location and time of a trip. One data processing method was developed to estimate the exit information from two consecutive boarding records, but many issues resided in the output. The DSSG ORCA team was eager to design a more robust algorithm to process the raw data into a standard and meaningful format for transportation research purposes, and since most AFC systems adopted the collecting method similar to ORCA, this algorithm could remove a huge barrier for other research groups on this topic. In addition, this project could also set an example of how public agencies effectively collaborate with universities on data analysis and manage some concerns involved in releasing government data , which do not often occur in transportation research field previously.

The social benefit from this ORCA project was immense if the team captured meaningful insights of the current transit system. A better transit service will encourage more people to take public transportation rather than driving cars, which would lead to less congested roads with reduced energy-cost and breathable air. The vision is to improve the service by scrutinizing two general topics: transfer analysis and bias analysis. Transfer analysis is important in transportation planning for reasons. It can identify any transfer location not included in the existing network, and this could save time and effort from transit riders, as well as it can point out any route that needs more buses to carry higher volume and this assist transit agencies in allocating resource more wisely. Bias Analysis compares two data sources, ORCA users versus all transit users which includes non-ORCA passengers. The bias analysis allows us to recognize the inherent difference between non-ORCA transit users and the ORCA users and predict one from the other. This can help us to understand how public transit serves all types of passengers and develop programs to invite those people to the electronic fare system. Besides, ORCA has established its reduced fare programs for low-income group, disabled and seniors, so from the ORCA data set, it is possible to inspect the transit service quality for those particular groups and make modifications to accommodate their needs.

Statistical Bias in ORCA

The ORCA data, though highly informative of transit ridership behavior, is limited by its scope. ORCA users make up 70% of all public transit riders, but their prevalence and usage is not uniform across the system and time periods. To truly understand transit ridership behavior of

the total population and not just ORCA users, the ORCA data must be accounted for how incomplete it is across different routes, stops, and times. To account for this bias, the ORCA data is compared with data from Automatic Passenger Count (APC) systems installed in buses that record the total number of passengers boarding and exiting the bus. This orthogonal data source is aligned with our primary data to see the ratio of ORCA passengers to total passenger counts for a given stop, route, and time. However, because the APC systems are installed on only a subset of buses, the APC data is much smaller than the ORCA data and available for far fewer trips and stops.

During the nine-week period in 2016 for which ORCA and APC data was available, there were roughly 31.7 million ORCA transactions recorded compared to roughly 11 million APC transactions. Of the 11 million observations for which there are both ORCA and APC transactions records, approximately 6 million or 54% are erroneous for either the APC or ORCA counts, leaving roughly 5 million bias ratio observations to work with. This data spans all stops and routes, with an average of 87 different observations for each route and stop combination. The table below is a sample of the bias table constructed, displaying the average difference between ORCA and APC counts for every stop for a particular given route and time. For this trip there are 35 or 34 samples for each stop, and is used to approximate what the difference would be in future trips or for other past trips for which there are no APC counts available.

TRIP_ID	RTE	DIR	STOP_ID	stop_name	med_stop_seq	sum_apc	sum_orca	count_diff	avg_count_diff	avg_bias_trip	bias_trip	num_trips
30935604	1	N	3600	S JACKSON ST & 12TH AVE S (3600)	0	35	11	24	0.685714286	0.572278912	0.52173913	35
30935604	1	N	1500	S JACKSON ST & 8TH AVE S (1500)	1	1	0	1	0.028571429	1	1	35
30935604	1	N	1510	S JACKSON ST & MAYNARD AVE S (1510)	2	8	1	7	0.205882353	0.8	0.77777778	34
30935604	1	N	1530	S JACKSON ST & 5TH AVE S (1530)	4	64	21	43	1.264705882	0.57043758	0.50588235	34
30935604	1	N	1610	PREFONTAINE PL S & YESLER WAY (1610)	5	25	11	14	0.4	0.751851852	0.38888889	35
30935604	1	N	538	3RD AVE & COLUMBIA ST (538)	6	23	13	10	0.285714286	0.277777778	0.27777778	35
30935604	1	N	558	3RD AVE & SENECA ST (558)	7	23	13	10	0.285714286	0.375	0.27777778	35
30935604	1	N	575	3RD AVE & PIKE ST (575)	8	26	19	7	0.2	0.219047619	0.15555556	35
30935604	1	N	600	3RD AVE & VIRGINIA ST (600)	10	39	11	28	0.8	0.627777778	0.56	35
30935604	1	N	605	3RD AVE & BELL ST (605)	11	20	12	8	0.228571429	0.297222222	0.25	35
30935604	1	N	1690	3RD AVE & VINE ST (1690)	12	7	4	3	0.085714286	0.466666667	0.27272727	35
30935604	1	N	2320	1ST AVE & BROAD ST (2320)	13	2	1	1	0.028571429	0	0.33333333	35
30935604	1	N	2330	1ST AVE N & DENNY WAY (2330)	15	7	4	3	0.085714286	0.375	0.27272727	35
30935604	1	N	2360	1ST AVE N & REPUBLICAN ST (2360)	16	3	0	3	0.085714286	1	1	35
30935604	1	N	2370	MERCER ST & QUEEN ANNE AVE N (2370)	17	12	0	12	0.342857143	1	1	35
30935604	1	N	2390	2ND AVE W & W ROY ST (2390)	19	1	0	1	0.028571429	1	1	35
30935604	1	N	2400	W OLYMPIC PL & 3RD AVE W (2400)	20	1	0	1	0.028571429	1	1	35

Table 1. Summary Statistics for Route 1

In addition to the average difference in count and percentage bias and ORCA counts, Transportation Analysis Zones (TAZ), discrete time intervals, continuous time, route direction, and destination are included as features to include in predicting total passenger count. All observations, in addition to their precise timestamp, are binned into six time intervals – Weekday AM peak (5 to 8 AM), Weekday PM peak (3 PM to 7 PM), Weekday Midday (9 AM to 2 PM), Weekday Night (8 PM to 4 AM), Weekend Midday (9 AM to 7 PM), Weekend Late (8 PM to 8 AM) – that are consistent with previous public transit analyses. The direction of the route

(inbound or outbound) as well as the route's end destination are included, as well as the bus stop's location. Though a given bus stop and route combination has on average a sample size of 87 observations, there are bus stop-route combinations with just a few observations. To generalize the model and to maintain geographic information while ensuring a large enough sample size, bus stops are binned into Transportation Analysis Zones (TAZ), a standard geographic unit model used commonly for transportation planning. The map below shows King County's various TAZs and the corresponding average percentage of the bus boarding counts that are not ORCA passengers for bus stops in that TAZ.

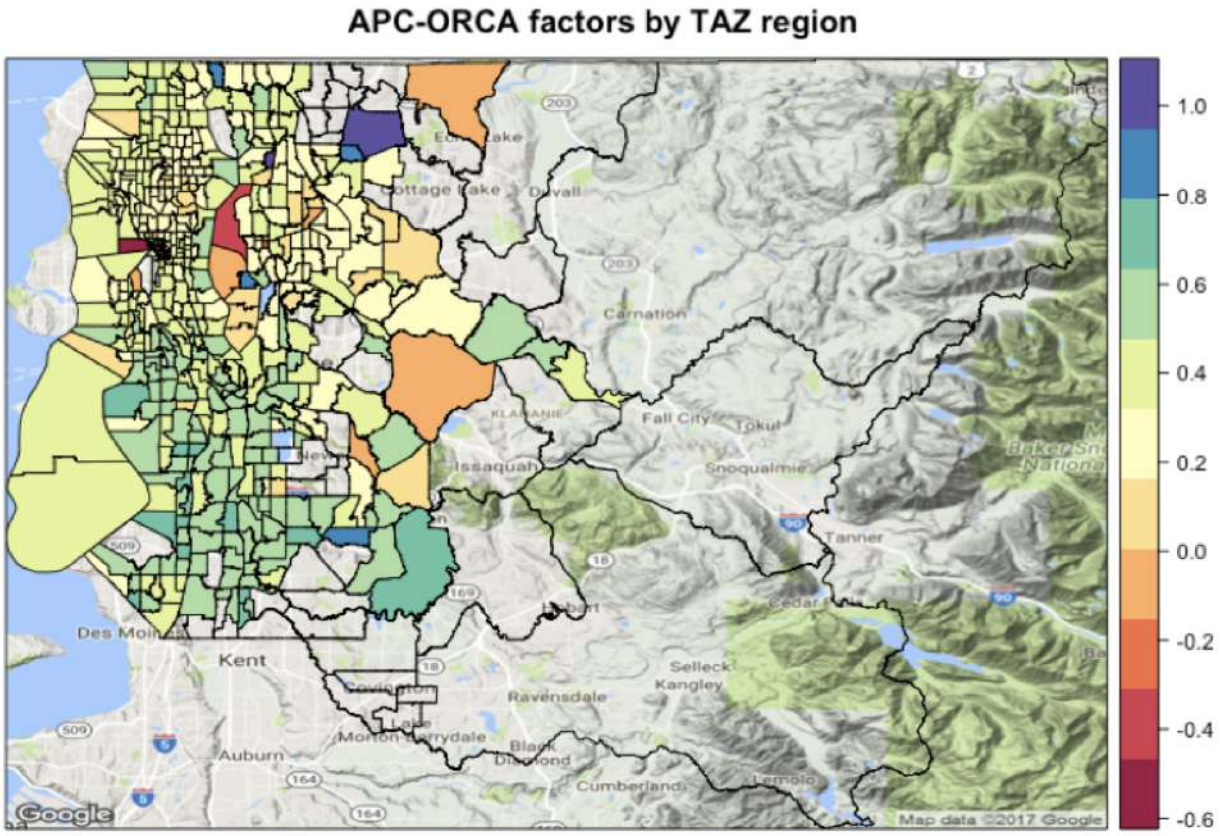

Figure 2

One of issues made clear by the TAZ bias map is the high level of noise and inaccurate data. Specifically, all bias levels below zero indicate incorrect data, as it is impossible to have fewer APC counts than ORCA counts. Though these are clear indicators of bad data either on the ORCA counts or APC counts, much of the data is plausible but unconfirmed. Indeed, there is no "ground truth" in this data. Even with rigorous data cleaning and processing, a level of noise and misclassification is inevitable and is a foundational fact that guides modeling decisions.

With this understanding and processing of the data, our team is currently exploring and refining two modeling approaches to infer total passenger count.

One approach is a zero-inflated negative binomial regression model with regularization. This mixed model approach is well suited for our problem given the distribution of the data, the dependent count variable with a high percentage of zero counts, and high level of noise in our data. A zero-inflated model combines both a logistic model to separate valid zero values from invalid zero values and a negative binomial regression on the remaining valid data. A shrinkage parameter is included in the regression model to select significant variables and reduce variance that result from the categorical variables such as TAZ and Routes with a high number of levels.

In addition to this regression approach, Hidden Markov Models are being developed to classify data as either accurate or inaccurate. Hidden Markov Models are well suited to this problem because of both the temporal nature of the data and the high level of noise in data. Specifically, the difference in APC and ORCA counts with their high level of noise and inaccuracy are the observable states that are dependent upon the true, accurate counts of ORCA and APC counts, or the "hidden" state. A multi-state Hidden Markov Model fit appropriately with covariates is being developed to identify accurate and inaccurate observations. The resulting data and Markov Model would be able to automate data cleaning in future data and would provide clean data to forecast with. The image below is a graphical representation of the transition probabilities between accurate counts (0) and inaccurate counts(1) on just a single route across time.

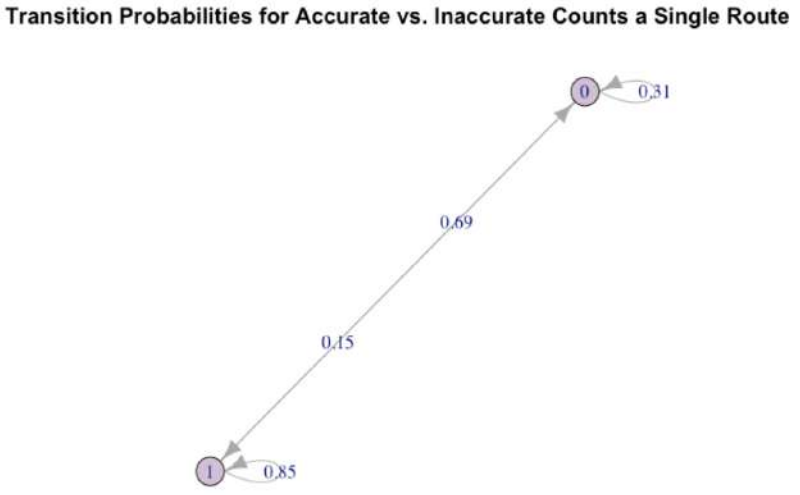

Figure 3

Transfer Analysis

This study aims to analyze the transfer related travel patterns in Washington State's Puget Sound region. The ORCA transaction data was used to recognize the transfers among all the trip boarding records. According to ORCA's definition of a transfer, within a 2-hour time window, a transit user has no limitations on the number of transfers that can be made. Roughly, 25 percent of the passenger-boardings in the ORCA transaction data table were recognized as transfers. As mentioned before, these transfer records consisted of the transfer boarding related information such as the transfer boarding time and the transfer boarding route. However, no information was available about the exit location of the transit user before the next boarding. This was formulated by performing quarter mile look-ups around the transfer boarding stop ID and a transfer data table was built.

The ORCA records were available for 2015 and 2016 and a transfer-data table was constructed for each year. Many different anomalies were found within these datasets as several data fields were estimated using completely different data files. For example, the exit location before the transfer boarding were built using the automatic vehicle location data table that was available from three different transit agencies, namely King County Metro, Community Transit and Pierce Transit. To have data that would comply with ORCA's definition of a transfer, several data cleaning tasks were performed. For example, transfer records that had transfer durations more than 2 hours and that had negative transfer time durations were identified and removed. These errors existed due to various reasons such as inconsistency in clock-time calibrations and problems in the GPS locations from the automatic vehicle location data.

Following these data cleaning tasks, several analytical questions were explored. The first question aimed at classifying different kinds of transfer. Transit users can carry out different activities such as shopping in the 2 hours transfer time duration and the task was to separate out these "financial" transfer records from the "real" transfers. For this purpose, it was also decided to add other explanatory variables to the transfer data set. For example, to know the walking time between the exit stop and the transfer boarding stop, walking times using HERE Maps API [1] were queried. Another variable that was introduced is the number of buses a transit user misses before boarding another bus and after alighting the previous bus. This variable was built by taking time difference between two consecutive time records for a bus route at a specific time of the day and fitting a function to it.

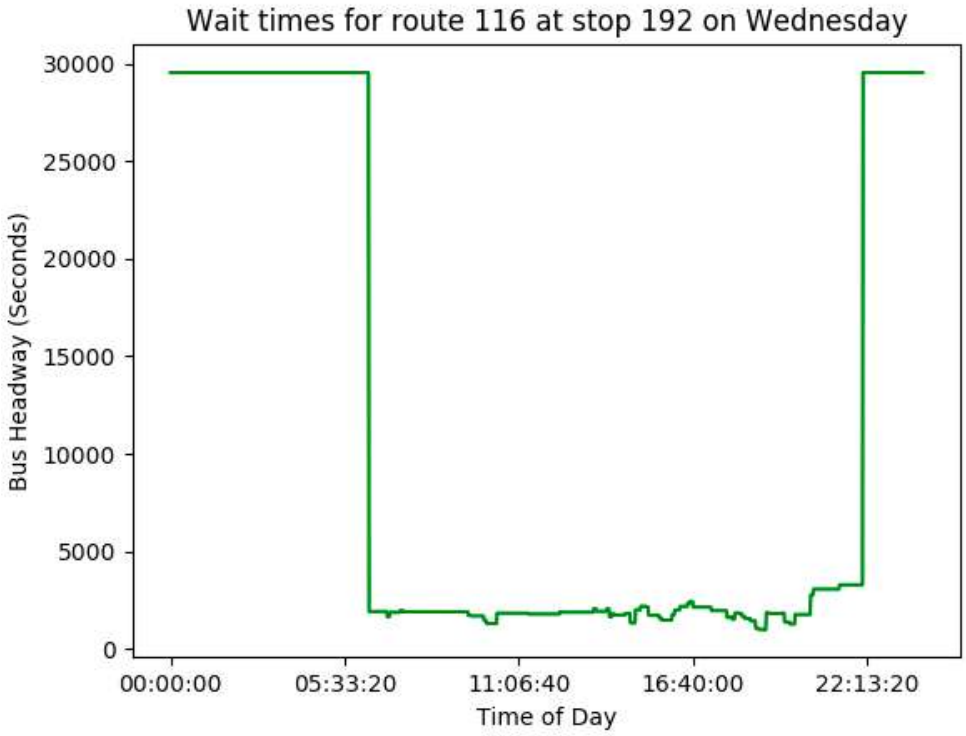

Figure 4

Another analytical question that was looked into is the change in transfer trip distributions between 2015 and 2016. For some areas in Seattle, such as the University District area, there have been several changes in the transit routes and transit stops. These changes have affected the transit related activities and new travel patterns can be seen in many areas. For example, the addition of light rail station in the University District in 2015 area made University District a more heavily-used transit area as the total trips count has increased by roughly 3500 daily from 2015, and the count did not include the trips go through the University District as its transfer location. The travel pattern has not dramatically changed in terms of number of transfers for each trip. However, about 97 percent of the trips starting from or ending in University District required zero or only one transfer, which indicates most trips were well served and the modified transit plan to adapt the link route was reasonable.

Trip Starting in University District				
Number of transfers	2015 count	Percentage	2016 count	Percentage
0	565912	80.75%	751113	81.69%
1	114736	16.37%	141975	15.44%
2	17185	2.45%	22516	2.45%
3	2603	0.37%	3218	0.35%
4	346	0.05%	513	0.06%
5	58	0.01%	85	0.01%
6	6	0.00%	20	0.00%
7	3	0.00%	2	0.00%
8	1	0.00%	0	0.00%
9	1	0.00%	0	0.00%
10	1	0.00%	0	0.00%
Trip Ending in University District				
Number of transfers	2015 count	Percentage	2016 count	Percentage
0	592654	81.20%	730357	79.87%
1	117652	16.12%	155251	16.98%
2	16764	2.30%	25160	2.75%
3	2407	0.33%	3062	0.33%
4	306	0.04%	490	0.05%
5	71	0.01%	88	0.01%
6	13	0.00%	10	0.00%
7	6	0.00%	2	0.00%
8	1	0.00%	1	0.00%
9	1	0.00%	0	0.00%
10	1	0.00%	0	0.00%

Table 2. University District Transfer Analysis

Conclusions and Future Work

As a part of the DSSG program, the ORCA team has been able to refine the existing database and has added explanatory power to the different ORCA related data-sets by restructuring them. As mentioned in the bias and the transfer analysis, different exploratory statistical tests have been performed to gain deeper insights into spatial factors affecting various transit activities in Seattle.

For the statistical bias analysis, the remaining three weeks will be spent on developing and refining the zero-inflated negative binomial regression and Hidden Markov Models. A shrinkage parameter will be added to the current regression model to reduce variance and dimension-size. The regression model will also be run with cross-validation on an Amazon Web Services virtual machine for further robustness. The Hidden Markov Model will also be further developed into a multistate model and fit with covariates. Deploying a Hidden Markov Model for the entire transit system will be compared against deploying a series of separate Hidden Markov Models for every individual route for effectiveness and predictability. It is likely that the Hidden Markov Models will also be run on cloud computing resources given its computational cost.

For the transfer analysis , different modelling techniques to automate the process of differentiating between real and “financial” transfers will be examined. This analysis will aim to identify whether a transfer record is a real transfer record as that will help in building up a refined origin- destination table that can be used to answer other questions by the transit agencies for improving their transit planning related operations. Also, we want to further use the transfer table to study the different transit related patterns in the University District Area.

References

[1] "REST APIs & Platform Extensions." *REST APIs - HERE Developer*. Web. 29 July 2017.