Analyzing Manually-Collected Uber Data to Maximize Driver Earnings in Salt Lake City

GEOG 6000 - Final Project

By Erik Neemann

13 December 2017

Abstract

Over the last several years, ridesharing apps have become increasingly popular in the United States. However, companies have maintained a close hold on the collected data, meaning it's not available to help drivers make decisions on when/where to drive. This paper analyzes manually-collected Uber data from Salt Lake City during the fall of 2017 to help drivers maximize earnings. The Uber data are explored using a variety of statistical methods from summary statistics to regression models, decision trees, principal component analysis (PCA), geostatistics, and spatial point patterns. A typical Uber trip lasts around 12.5 minutes, covers 4-5 miles, and results in a total earning of $5-6. There is no observed difference in trip distance, tips, and driver earnings based on rider gender, but trips ending at the airport cover significantly larger distance and result in greater tips and earnings than non-airport trips. Regression models provide useful prediction of total earnings based on a handful of variables, while probabilistic methods add value by assessing confidence. Trip start and end points are shown to cluster and co-occur in space. Principal component and geostatistical analyses, along with local autocorrelation tests can help Uber drivers identify good locations to find efficient, high-earning trips.

Introduction

As Uber and other rideshare platforms continue to grow in popularity, the companies amass tremendous amounts of data from the thousands of rides logged every day in major cities (Dogtiev 2017). This data would be extremely useful for municipality planners and engineers, as well as

Uber drivers alike. However, Uber understands the power of its data and keeps a close hold on it, although it has recently unveiled Uber Movement, a site that can be used to analyze traffic patterns in cities. Unfortunately, this information has only been shared for 7 cities worldwide, 2 of which are in the United States (Boston and Washington D.C.). Furthermore, the data itself isn't publicly available, only a limited look into insights that can be gained from the data, such as average travel times (https://movement.uber.com/cities?lang=en-US).

This paper takes a more detailed look at Uber data that was manually-collected in Salt Lake City from August to November of 2017. The goal is to help answer questions, such as:

- Where should drivers position themselves to get the highest-earning rides?
- How accurately can a trip's earnings be predicted based on a few variables?
- Do trip distances, tips, or earnings differ by rider gender?

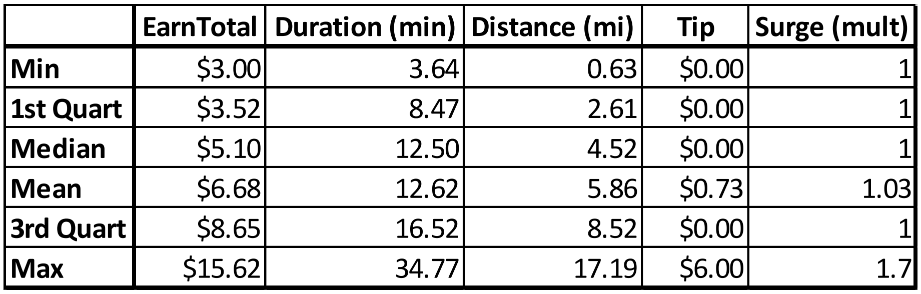
A variety of data analysis methods were employed to interrogate the data, including: summary statistics, linear regression, generalized linear models (GLMs), regression trees and random forests, principal component analysis (PCA), geostatistics, and spatial point patterns. These techniques have been applied in a variety of ways throughout the scientific literature over the past few decades. PCA is frequently used to reduce the dimensionality of data and explain covariances among multiple variables. In the field of climate, these covariances are often referred to as teleconnections and may link weather patterns in different regions. Smith et al. (2015) used PCA on detrended monthly sea surface temperature data to characterize the Pacific Decadal Oscillation for climate studies of Great Basin precipitation. Similarly, Strong and Davis (2008) used principal components to correlate Northern Hemisphere jet stream variability with the Arctic Oscillation, which is based on monthly mean 1000 millibar height anomalies between 20-90°N.

Random forests have gained traction in the last decade as a technique for forecasting hard-to-predict weather phenomena. Different studies have used a wide variety of observed, remotely sensed, and simulated weather variables to forecast the occurrence and severity of aviation turbulence (Williams 2013) and thunderstorm initiation (Williams et al. 2008). In some cases, a combination of regression trees and geostatistics have been used to improve interpolation. Balk and Elder (2000) modeled snow depths in Colorado’s Loch Vale watershed by using binary decision trees to capture large-scale variations and kriging (or co-kriging) for small-scale variations. These two methods combined to produce better results than either method on its own. All of the well-known techniques discussed above have shown success in previous studies and were useful in examining the Uber data.

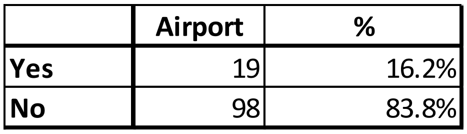
Data and Methods

Data was collected from 125 Uber trips between August and November 2017. Of these 125 trips, the analysis data set was trimmed down to those trips starting in Salt Lake County (117), while the remaining trips (7) were used for model verification. One outlier was also removed, which had a trip distance and total earning that was more than double any other trip. Data was primarily gathered on Friday mornings between 5:30-9:30am. The spatial data were collected with a smartphone GPS application and the attributes were gathered from the Uber app and website. These attributes include trip date, start/end time, duration, distance, driver earning total, surge multiplier, tip, rider gender, latitude, longitude, and an airport trip identifier. Some analyses were conducted on the trip’s total earnings, while other methods used a corrected (standardized) value with the surge multiplier and tips removed. This was done in an effort to remove the seemingly random components of the total earnings. Polyline features from the GPS traces were collected, but the geographical analysis presented here is limited to the trip starting points (and ending points) and their associated attributes (Figure 1).

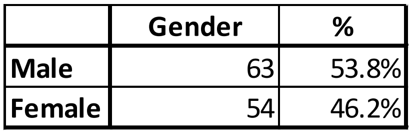
A sampling of standard statistics for the trips are summarized in the tables below. A typical trip lasted around 12.5 minutes, covering 4-5 miles with no tip or surge multiplier, resulting in a total driver earning of $5-6. While this trip is considered "typical," there are a wide variety of trip total earnings, corrected earnings (tip & surge multiplier removed), distances, durations, and tips, as shown in the histograms (Figures 2-6). Trip earnings and tips tend to follow an exponential distribution, while distance and duration are closer to normal or gamma distributions. The differences in these distributions created challenges in analyzing the data, resulting in some caveats for interpretation.



	EarnTotal	Duration (min)	Distance (mi)	Tip	Surge (mult)
Min	$3.00	3.64	0.63	$0.00	1
1st Quart	$3.52	8.47	2.61	$0.00	1
Median	$5.10	12.50	4.52	$0.00	1
Mean	$6.68	12.62	5.86	$0.73	1.03
3rd Quart	$8.65	16.52	8.52	$0.00	1
Max	$15.62	34.77	17.19	$6.00	1.7



	Airport	%
Yes	19	16.2%
No	98	83.8%



	Gender	%
Male	63	53.8%
Female	54	46.2%

While only 16.2% of Uber trips end at the airport, airport trips result in higher driver earnings (Figure 7); the mean airport trip earns $11.09, compared to $5.83 for non-airport trips. A T-test confirms that this is a statistically significant result at greater than 99.9% confidence. Additionally, males make up 53.8% of Uber riders, but differences in driver earnings based on rider gender are not statistically significant. Driver total earnings most closely correlated with trip distance, and Figures 8 and 9 show a scatterplot of total and corrected earnings by distance. These plots also break out gender by color with males (females) in blue (pink), and airport trips by size where non-airport trips (airport trips) are small dots (large dots).

Numerous points within the data set were duplicates, resulting from pick-ups or drop-offs at the same location multiple times (e.g., the airport). This presented problems within the statistical analysis program, R, as several functions were unable to handle duplicate points (building neighborhoods, kriging, etc.). To combat this issue, R's "jitter" function was used to add a small amount of random noise to the coordinates in order to permit their use in all analysis methods. This noise was on the order of meters, so it is not expected to have impacted any of the results.

Several methods that were used to analyze the Uber data will be described in the following paragraphs. First, a handful of prediction models will be discussed, followed by geostatistical and spatial point pattern analysis. A trip's total earnings are of most interest to Uber drivers, so a handful of models were built in an effort to predict total earnings from other variables within the data. However, as mentioned previously, the distribution of total earnings is closer to exponential than Gaussian (see Figure 2), creating some challenges in modeling and interpreting the results. Nonetheless, total earnings were still predicted instead of corrected earnings or corrected earnings above the minimum, since total earnings are of greatest interest to drivers.

The first two prediction models were built with linear regression. One was a very simple model that relied purely on distance to predict total earnings -- distance is known to be the largest component of Uber's actual driver earning calculation. The second model used stepwise automatic selection starting from a null model and with a total scope of 8 variables (distance, duration, surge, tip, gender, airport, longitude, latitude). This resulted in an "optimum" model with 5 variables (distance, duration, tip, surge, gender) that minimized the AIC score. The third model was a GLM that used a Poisson distribution and log link function. However, in order to utilize this model, the total earnings distribution was discretized into bins by rounding to the nearest dollar. This permitted the use of the Poisson distribution and transformation back into dollars during the prediction. The fourth and fifth models used regression trees, the fourth of which used a single tree, pruned to a moderately aggressive level of complexity (Figure 10). The fifth model employed a random forest of 500 trees, with 5 variables considered at each split. As shown in Figure 11, the random forest easily determined distance to be the most important variable, followed by duration and tip. All other variables are much less important, and this also aligns with the variables used in the single regression tree (Figure 10). Finally, principal component regression was used for the sixth model. This applied 4 components resulting from the PCA that accounted for 98.4% of the variance in the data. Additional information on PCA will be provided later in the Discussion section. Once all models were built, they were used to predict total earnings and log-earnings from the 7 data points that fell outside of Salt Lake County.

Next, standard Moran's I, Monte Carlo simulations, and Getis-Ord (G*) tests examined global and local spatial autocorrelation in the start and end point data. For these tests, neighborhoods were built using the k-nearest neighbors method (k=5) with row-standardized weights. Randomization sampling was also used and log-values of corrected earnings were employed to account for the exponential distribution of trip earnings.

The third set of analyses consisted of geostatistical prediction and simulation. Ordinary kriging was initially done to estimate total earnings across Salt Lake City based on trip starting points. This was followed by conditional kriging simulations done to estimate the probability of earnings above a certain threshold. Both start and end points were used for kriging simulations, but the end point data proved much more difficult to fit with variograms, resulting in limited success.

It the final group of analyses, spatial point patterns and clustering were examined for the start and end points using Ripley's K. The start and end points were also examined for co-occurrence with the k-cross function. In order to perform these analysis, the start and end points were clipped down to a smaller region of denser points, encompassing Salt Lake City's primary interstate loop and extending west to Bangerter Highway.

Results

The prediction model results showed a variety of performance skill in estimating the total earnings of the 7 Uber trips that fell outside of Salt Lake County. Using root mean squared errors (RMSE) as the primary metric, the optimized linear model performed the best, followed by the PCR model and random forest model (Figures 12 and 13). The single regression tree performed the worst, while the simple linear model and the GLM produced similar, poor results. For the predictions based on log-earnings, all of the model showed improved performance, except for the PCR model (Figures 14 and 15). In this case, the random forest model was the best, followed by the single regression tree and the optimum linear model. The fact that nearly all models performed better with log-earnings, demonstrates that using log-earnings as the dependent variable is probably the best choice for regression modeling, given the distribution mentioned previously. It's not immediately clear why the PCR model's performance didn't noticeably improve in the predictions based on log-earnings, like the other models. However, the components in the PCA were derived independent of the response variable and the scaling that was used in the "prcomp" function may have also nullified any differences based on the dependent variable.

Spatial autocorrelation tests showed that the log-earnings of the Uber start points did not demonstrate autocorrelation. The Moran's I value was 0.068 and p-values were insignificant. End points, on the other hand, did exhibit spatial autocorrelation with a Moran's I of 0.484 and a pvalue of 0.001 based on a Monte Carlo test with 999 simulations. This is evidence to reject the null hypothesis and conclude that spatial autocorrelation does exist with Uber trip end points. A Getis-Ord (G*) test was also performed and z-scores were calculated for the start and end points. Each plot showed interesting patterns and clusters of local autocorrelation. The start points (Figure 16) have clusters of low values south of downtown and between downtown and the university, while a cluster of relatively high values is seen on the east side just north of I-80. For the end points (Figure 17), the airport has a strong cluster of high values and a cluster of low values is observed between downtown and the university.

A plot of total earnings that were the basis for kriging on start points is shown in Figure 18. These total earnings were used to interpolate a surface of values across the Salt Lake Valley with an ordinary kriging technique (Figure 19). The kriging results identify 3 primary areas of higher-earning start points (Uber trips near $10): the east side near 2100 S and 2100 E, west of the I-215 loop near Kearns, and to the south along I-15 near South Jordan. Of these three higher-earning locations, the east side appears to be the most robust, as the other two locations are interpolated from only a few (or 1) nearby points. Further, the ordinary kriging results showed relatively high cross-validation errors (not shown) even though the residuals appeared random. A kriging simulation of 50 iterations was conducted across the region to predict the probability of total earnings greater than $8 (Figure 20). This value was chosen subjectively because it represents approximately the top third (31%) of total earnings. The simulation identified similar regions as the ordinary kriging, with the addition of a fourth high-probability region between downtown and the airport. The highest probabilities, approaching 80%, are shown near Kearns and on the east side north of I-80. Total earnings for the Uber end points (Figure 21) were also used for kriging simulations. Figure 22 shows the results, with the airport identified by high probabilities of earnings greater than $8 at 90%. Another, broad region of high probabilities covers the southeast quadrant of the Salt Lake Valley, where only a few data points exist. However, the small sample size and poor variogram fit (not shown) for the end point data make this result highly suspect. The area between downtown Salt Lake City and the university also stands out as trips ending there have very low probabilities (less than 10%) of exceeding $8. Additional kriging was done on other variables, such as corrected earnings and corrected earnings above the minimum fare, but poor variogram fits resulted in output that was not trustworthy.

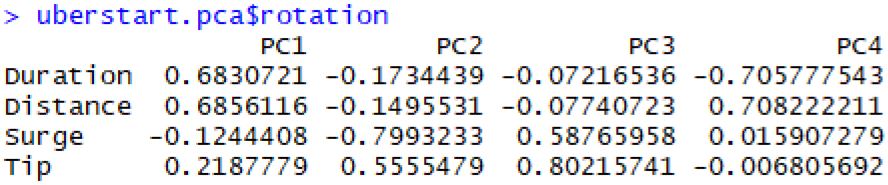
To examine spatial patterns, the Ripley's K was calculated using Monte Carlo simulations (99 iterations) for both start and end points. In both cases, the points were highly clustered (not shown) when compared a completely spatially random theoretical distribution. A third test for marked point patterns using R's "kcross" function was also conducted with a 99-simulation envelope. These results also demonstrate clustering and co-occurrence between Uber trip start and end points (Figure 23).

Discussion

In Salt Lake city, an Uber driver's earnings are based on a well-defined combination of trip distance ($0.7215/mile), duration ($0.0825/min), wait time ($0.0825/min), surge multiplier (between 1.0 and 3.0, multiplied by distance and duration), and tip (cash added at the end). All of these variables, except for wait time, were collected and analyzed in this study. The surge multiplier, which is the most nebulous variable, is calculated based on the ratio of riders to drivers in a region. It is generated by an opaque algorithm in that refreshes multiple times each minute and is visualized by a 2D field in the Uber app's map with occasional hotspots indicating where the multiplier is elevated. Less than 9% of all trips had a surge multiplier (i.e., Surge > 1.0). The Uber app also employs a minimum driver earning of $3 for each trip, which tended to cause model predictions errors on short trips. Several of these intricacies within the Uber app calculations complicated predictions in this study.

The results discussed above indicate that model prediction and geostatistical analysis can do a fair job of estimating Uber trip total earnings despite missing data (wait time) and two largely random components (surge multiplier and tips). All six regression models demonstrated some skill, particularly when estimating values based on log-earnings. While the optimum linear model performed nearly as well, the random forest approach appears to be the most powerful due its ability to handle irregularly distributed data, provide an estimate of confidence, and identify the most important variables. The single regression tree method was too erratic and relied heavily on being pruned to the right complexity level. Similar to the random forests, the kriging simulations are potentially useful to drivers because they provide probabilistic information over a region.

In this section, it’s also worth returning to the PCA results that were used in the principal component regression model. The “prcomp” function in R was only fed 4 variables (distance, duration, surge, and tip) and created 4 principal components (PC) shown below:



```
> uberstart.pca$rotation
                PC1        PC2         PC3          PC4
Duration  0.6830721 -0.1734439 -0.07216536 -0.705777543
Distance  0.6856116 -0.1495531 -0.07740723  0.708222211
Surge    -0.1244408 -0.7993233  0.58765958  0.015907279
Tip       0.2187779  0.5555479  0.80215741 -0.006805692
```

Generally speaking, PCA is used to help reduce the dimensionality of data containing many variables. However, in this case, it was examined to see if the “new” variables it created provided any additional value compared to the linear models discussed earlier. The four variables that were input into the model were chosen because they are the 4 variables that Uber uses to calculate total earnings (other than wait time). Of the resulting components, PC1 places high emphasis on duration and distance (long trips), which one would expect to correlate well with a driver's total earning. The real-world meaning of the second and third PCs is less clear, though the second may relate to tipping. PC4, on the other hand may be useful because it places a positive emphasis on trip distance and a negative emphasis on trip duration. This means that trips covering a large distance in a short period of time would score high in PC4. These are also the exact types of trips that are most profitable and efficient for Uber drivers. Figure 24 shows the start points plotted on a grid with their PC4 scores and it's clear that the high-scoring points are in relatively close proximity to highways. This is important because points with high PC4 scores have good accessibility to roads and offer drivers quick and efficient earnings. The small cluster of points scoring high in PC4 along the east side, north of I-80 also appears to correlate well with other Getis-Ord (G*) hot spots (Figure 16) and kriging results (Figures 19 and 20). Unfortunately, a much larger sample size is necessary to determine if this region of high-earning Uber start points truly exists, or if it's due to coincidence in a small data set. Some caution should also be taken in knowing that PC4 is a very minor component, explaining less than 5% of the variance in the data.

A number of limitations existed in this study. The small data set of 117 Uber trips is not large enough to draw strong conclusions. Furthermore, data set used for model verification (7 points) was also very small and included points outside of Salt Lake County. Both of these factors limit how broadly the results can be interpreted. The exponential and irregular distribution of the data also presented challenges that may have affected results. This was partially accounted for in some methods (regression modeling, autocorrelation tests), but ignored in other methods (PCA, kriging). Additionally, model verification was handled in a very simplistic manner and sophisticated cross-validation, bootstrapping, and significance testing was not explored.

Finally, this study has identified several areas of future work, where results could be expanded upon or improved. First, the methods should be extended to larger data sets that would be capable of producing more meaningful and significant results. This should include a larger range of times and dates, so diurnal and seasonal variations could also be explored. Second, many of the analysis techniques could be applied in more detail and to additional variables (e.g., corrected earnings, earnings above the minimum fare). Third, future analysis should employ greater use of resampling and advanced verification methods. Lastly, it's recommended that future studies more explicitly account for the observed data distributions within the analysis process.

Conclusions

This study has examined Uber data using a wide range of statistical methods to predict driver earnings based on known variables and estimate earnings in 2D space. Despite challenges related to sample size, frequency distributions, and verification data, these techniques still proved useful in identifying overall trends and high-earning locations that may be useful to Uber drivers in Salt Lake City. Confidence was also quantified by applying probabilistic methods, where appropriate. Some conclusions, including the statement that trips ending at the airport have longer distances and higher earnings and tips, proved to be statistically significant. Others however, such as differences in distance or earnings based on rider gender, were not significant. Additional work could improve on the results found here by analyzing larger data sets, using more sophisticated verification and resampling techniques, and examining a greater range of dependent variables.

Bibliography

Balk, B. and K. Elder, 2000: Combining binary decision tree and geostatistical methods to estimate snow distribution in a mountain watershed. *Water Resources Research*, **36**, 1, 13-26.

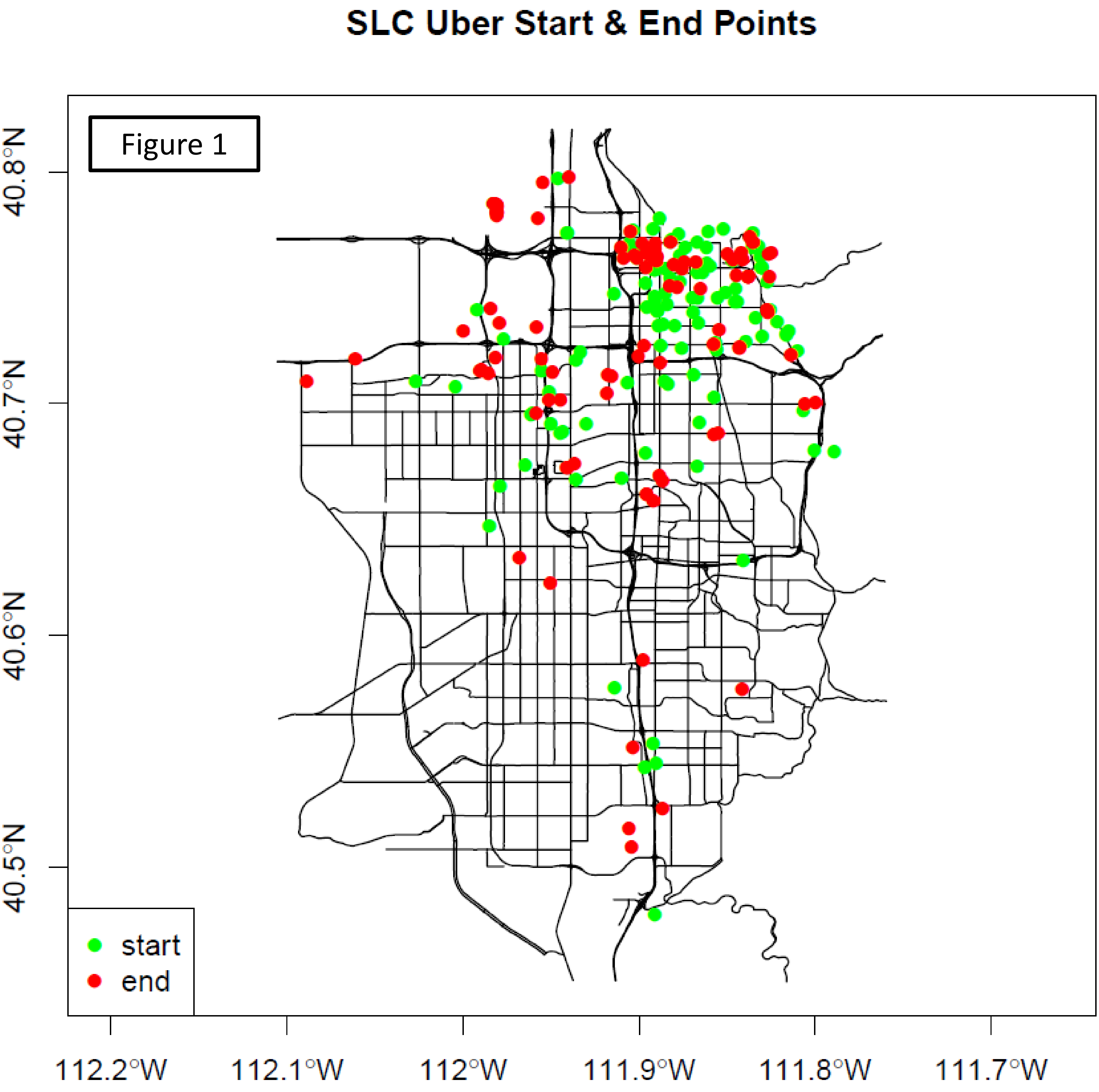
Dogtiev, A., 2017: Uber revenue and usage statistics. *Business of Apps*. Retrieved December 12, 2017, from http://www.businessofapps.com/data/uber-statistics/.

Smith, K., and C. Strong, 2015: Connectivity between historical Great Basin precipitation and Pacific Ocean variability: A CMIP5 model evaluation. *Journal of Climate*, **28**, 6096-6112, DOI: 10.1175/JCLI-D-14-00488.1.

Strong, C., and R. E. Davis, 2008: Variability in the position and strength of winter jet stream cores related to Northern Hemisphere teleconnections. *Journal of Climate*, **21**, 584-592, DOI: 10.1175/2007JCLI1723.1.

Williams, J. K., et al., 2008: A machine learning approach to finding weather regimes and skillful predictor combinations for short-term storm forecasting. *ResearchGate*. Retrieved 10 December, 2017, from https://www.researchgate.net/publication/259869325.

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Machine Learning*, **95**, 51-70, DOI: 10.1007/s10994-013-5346-7.


SLC Uber Start & End Points
Figure 1
40.8°N
40.7°N
40.6°N
40.5°N
112.2°W
112.1°W
112°W
111.9°W
111.8°W
111.7°W
● start
● end

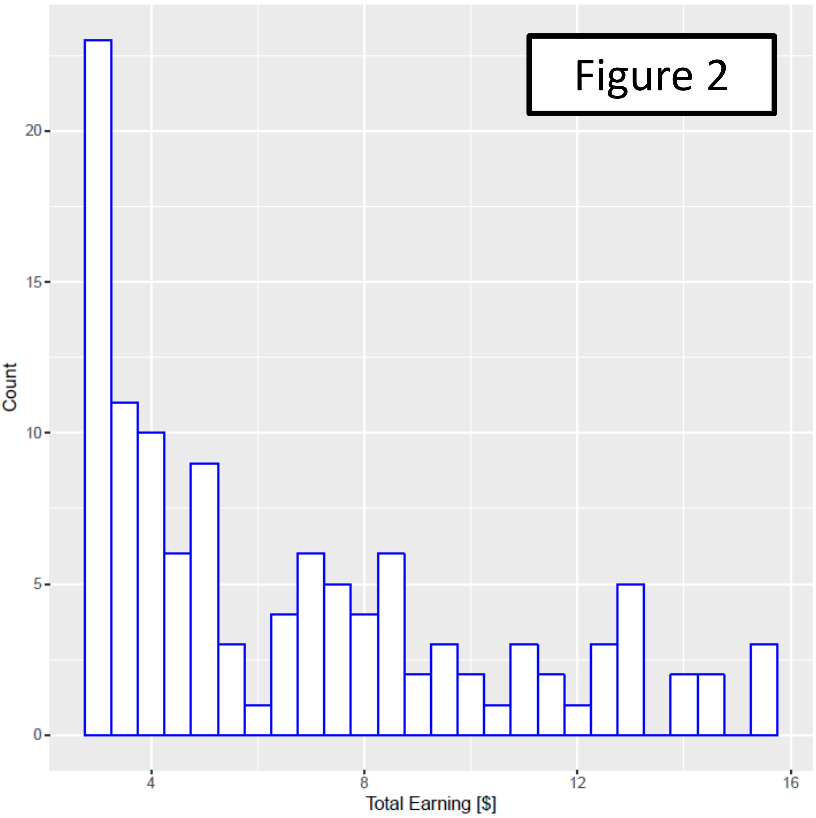


Figure 2

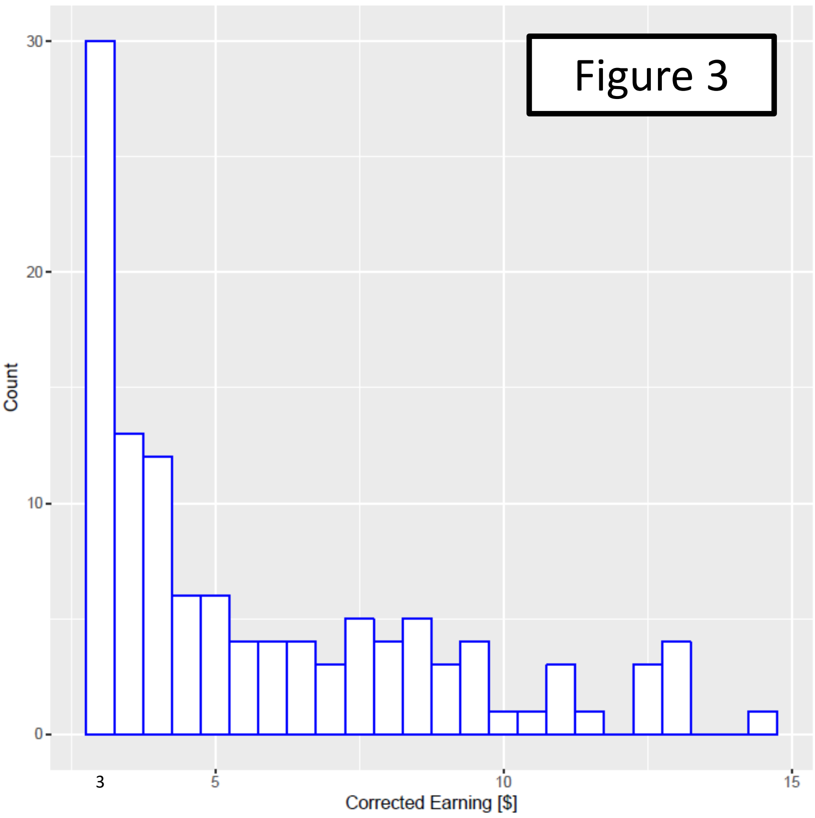


Figure 3

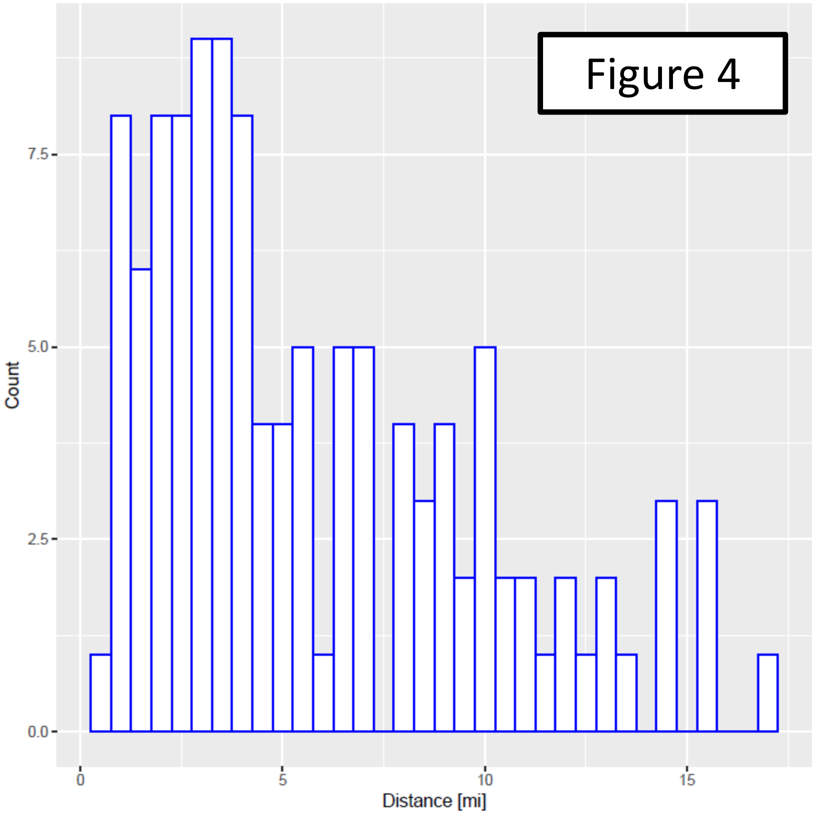


Figure 4

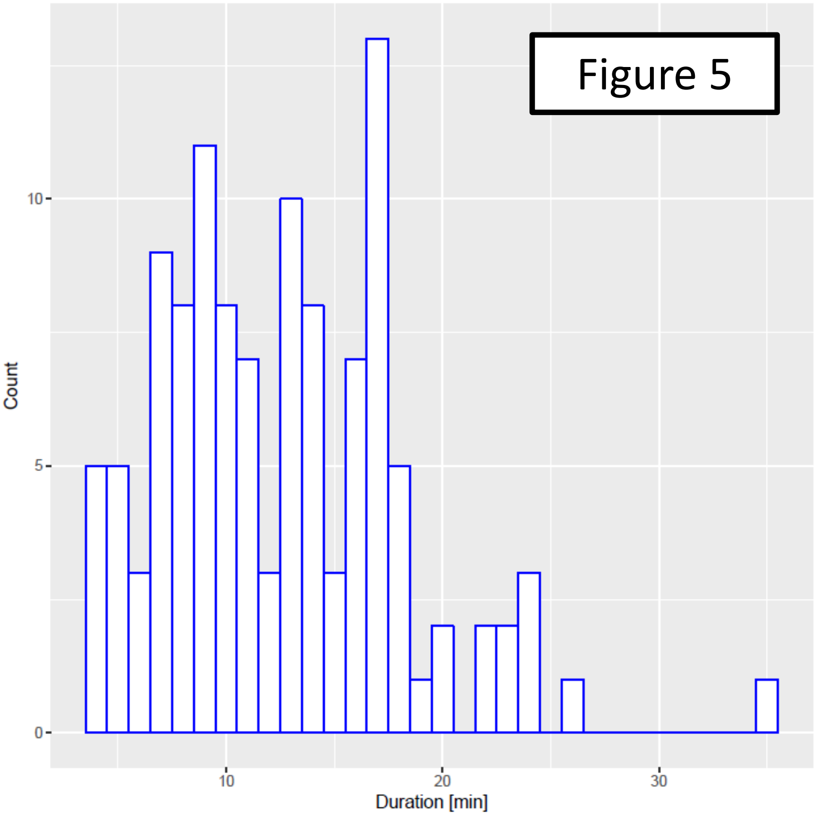


Figure 5

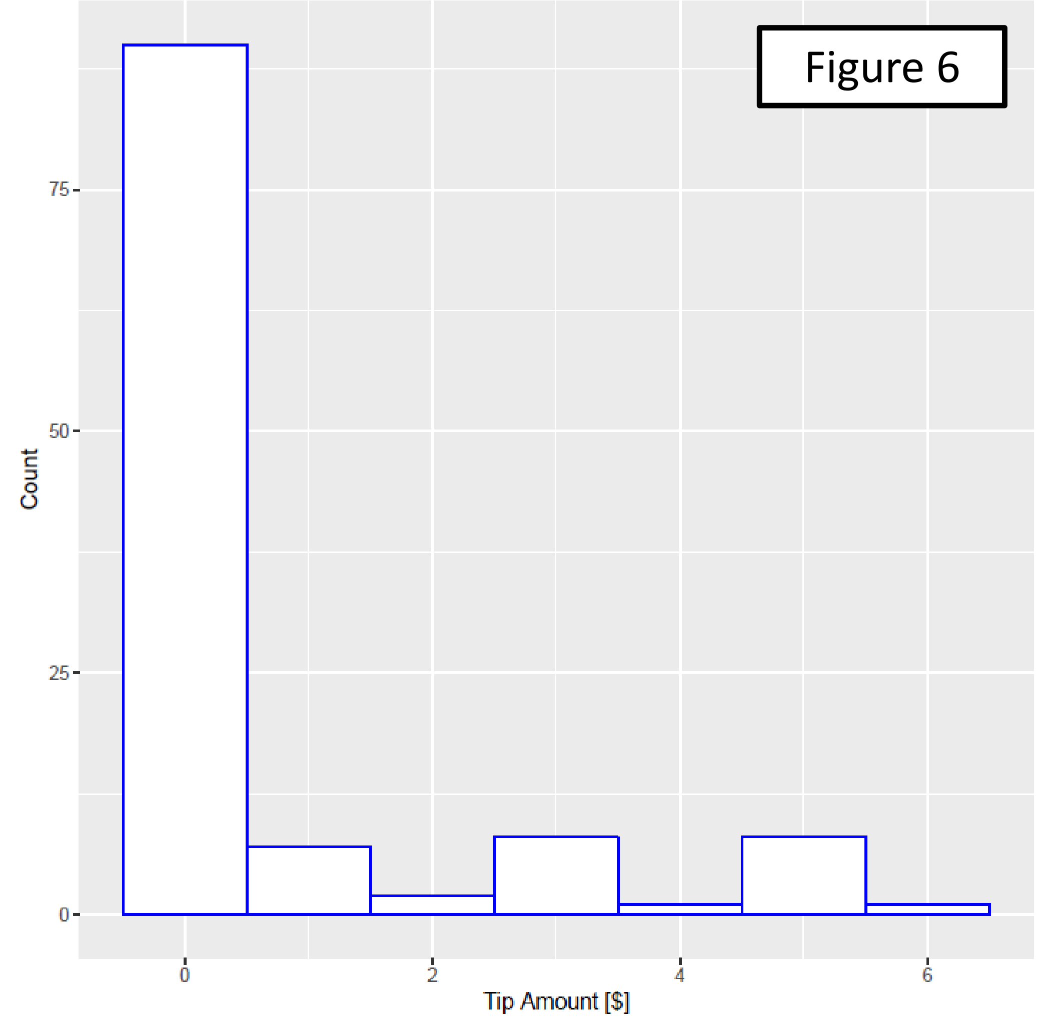
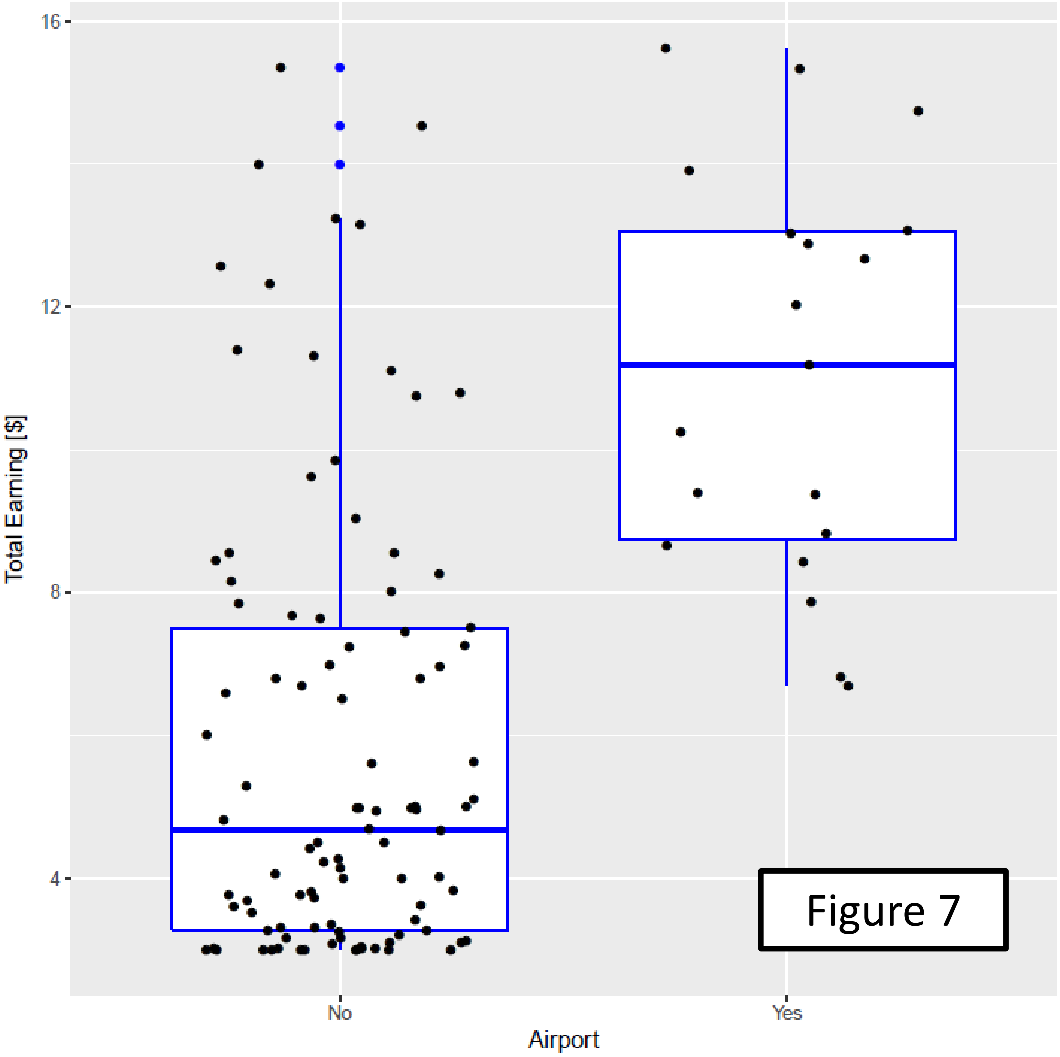


Figure 6
75
50
25
0
Count
0
2
4
6
Tip Amount [$]



16
12
8
4
Total Earning [$]
No
Yes
Airport
Figure 7

Figure 8

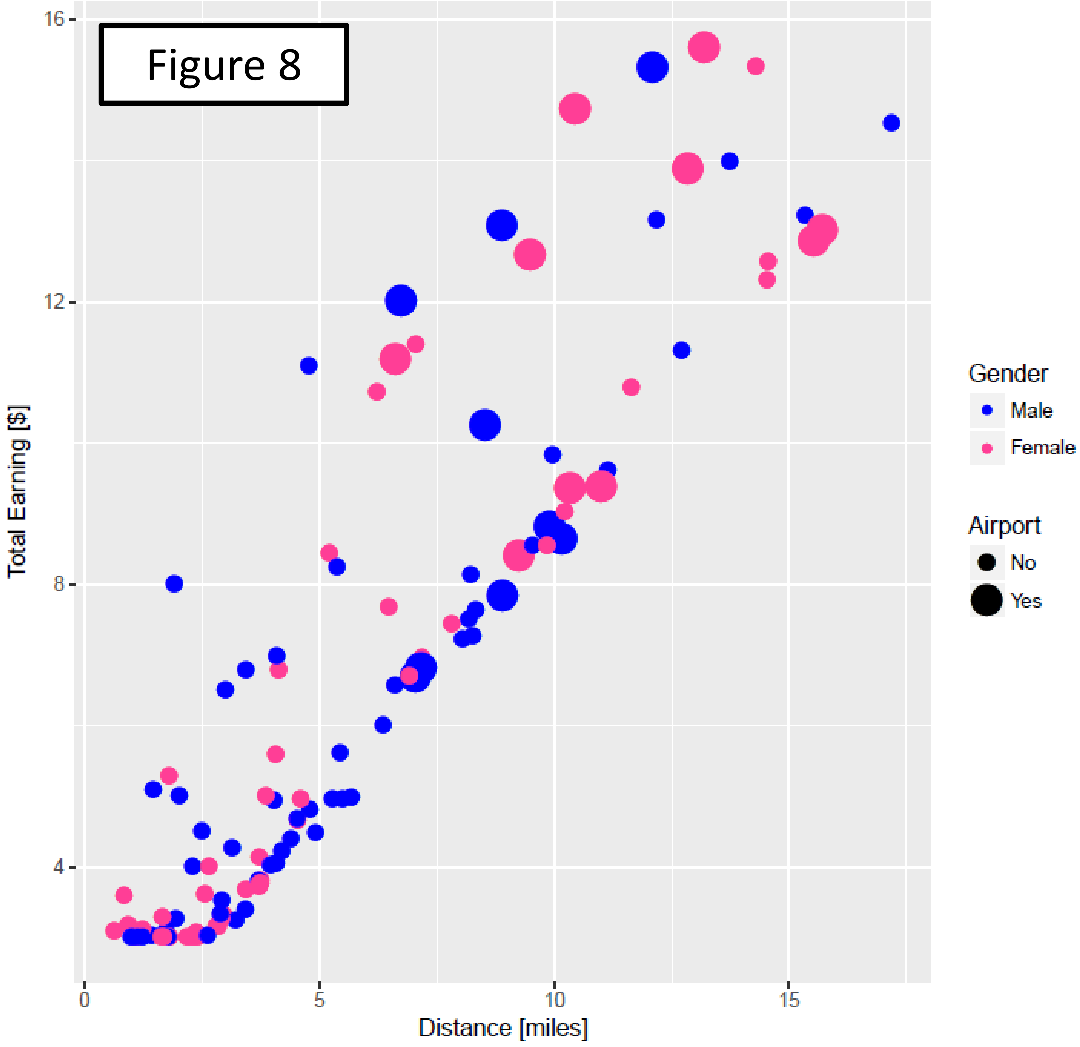
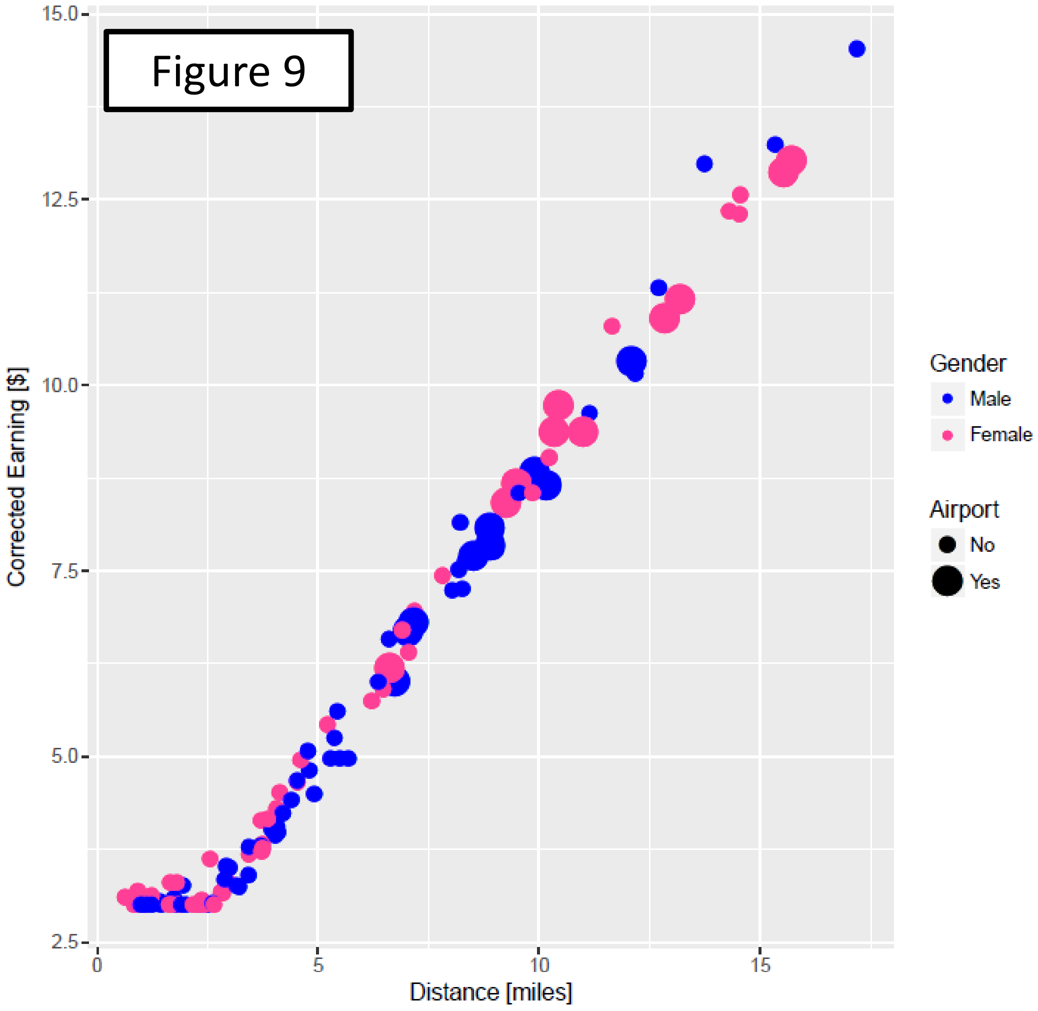


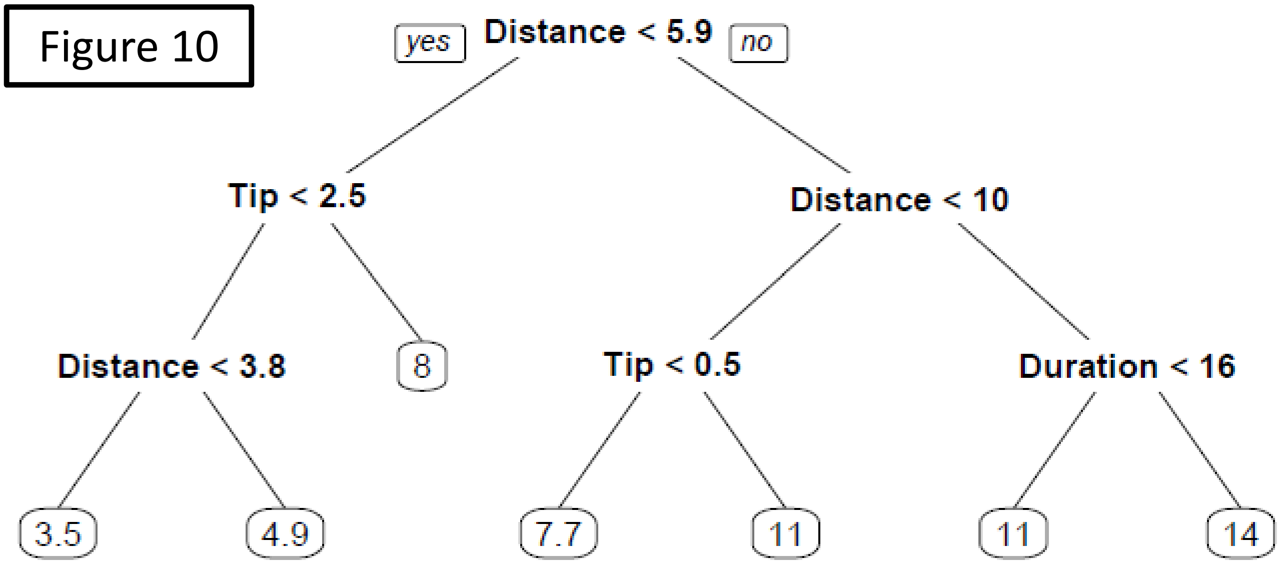
Figure 9

Corrected Ride Earnings



Regression Tree Used for Prediction

Figure 10



Random Forest Variable Importance Plot

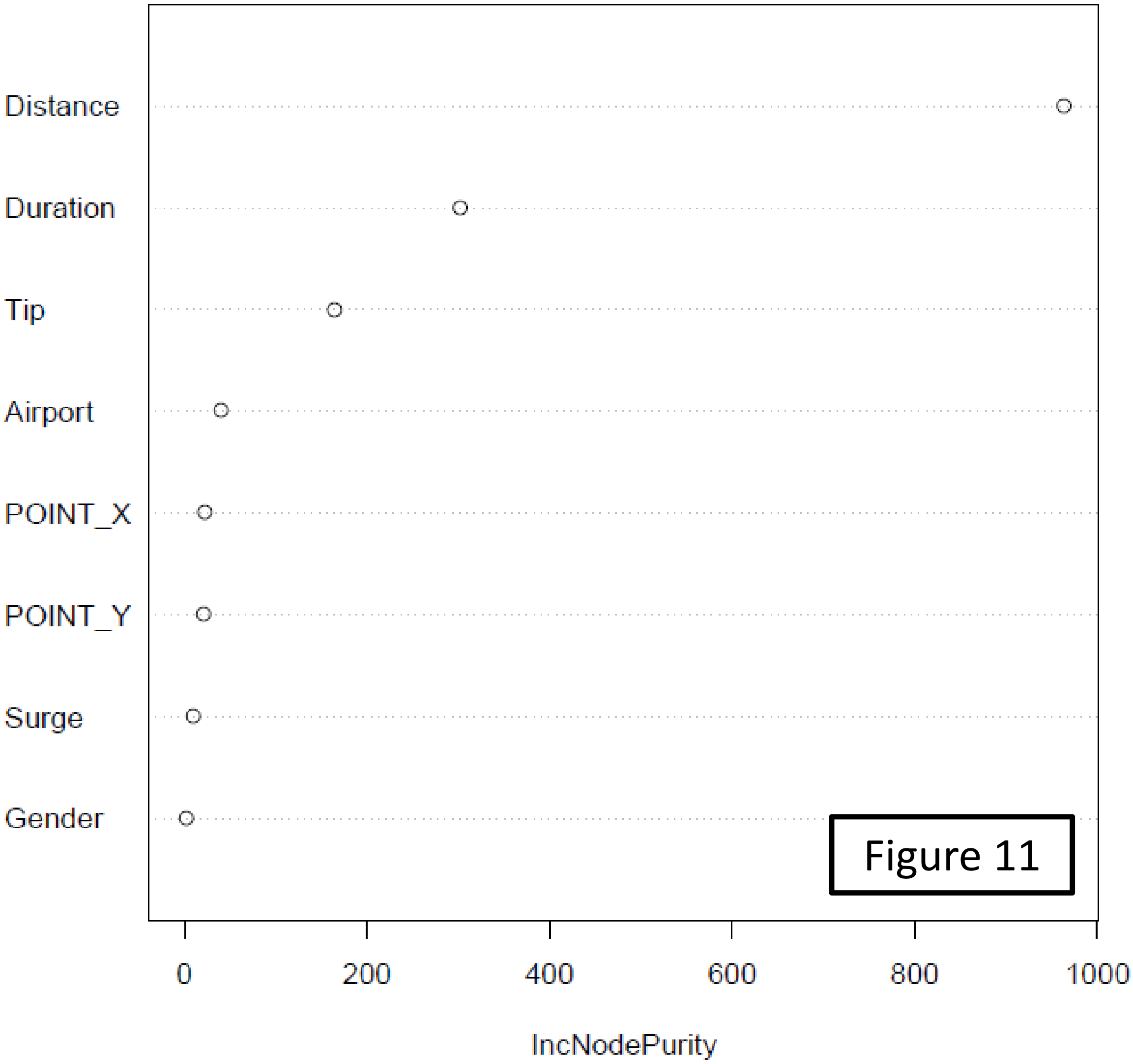


Figure 11

Figure 12

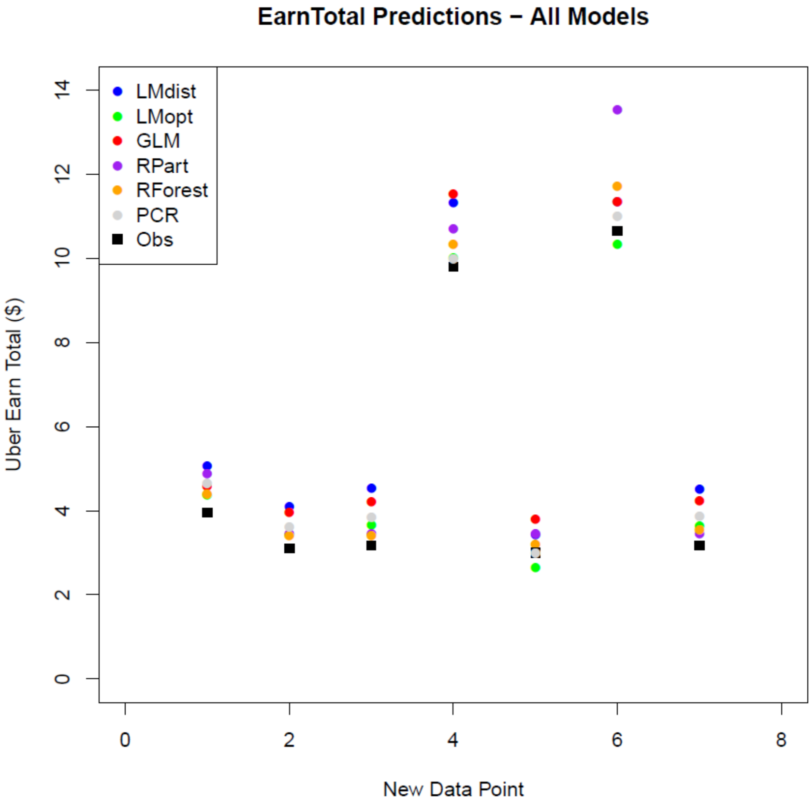


Figure 13

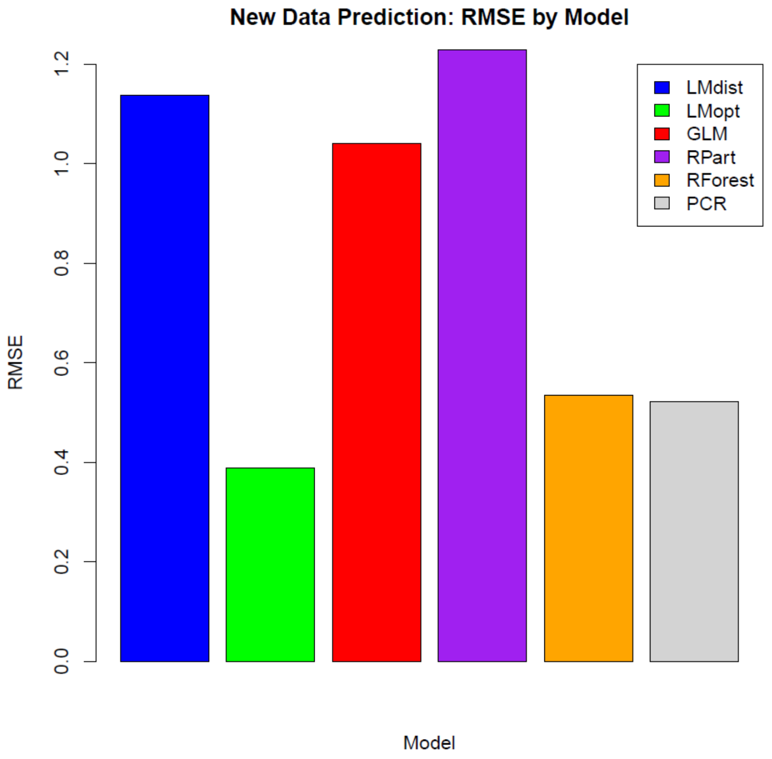


Figure 14

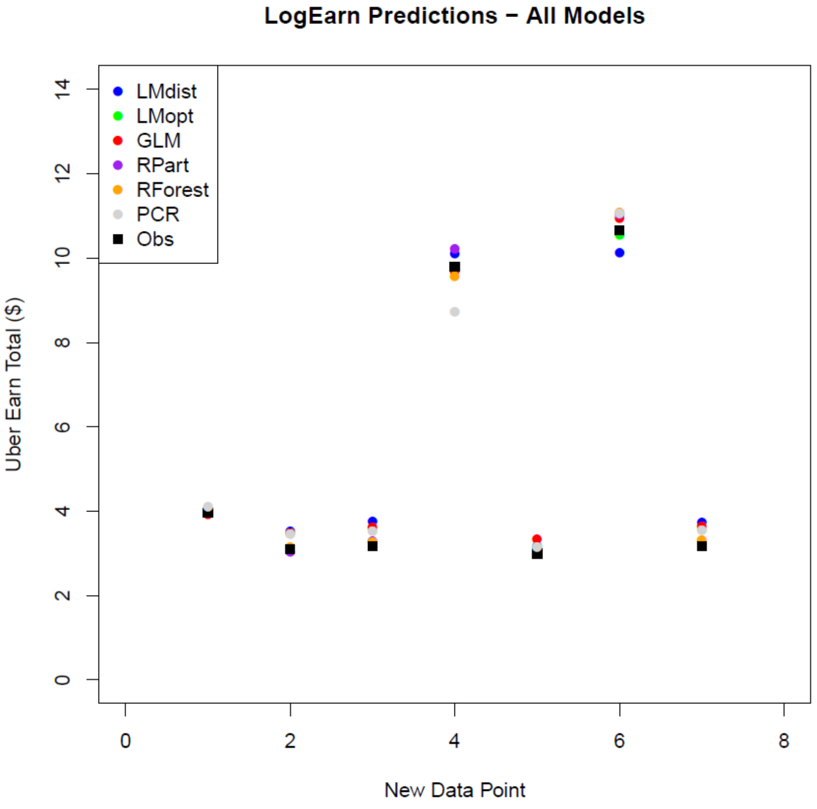


Figure 15

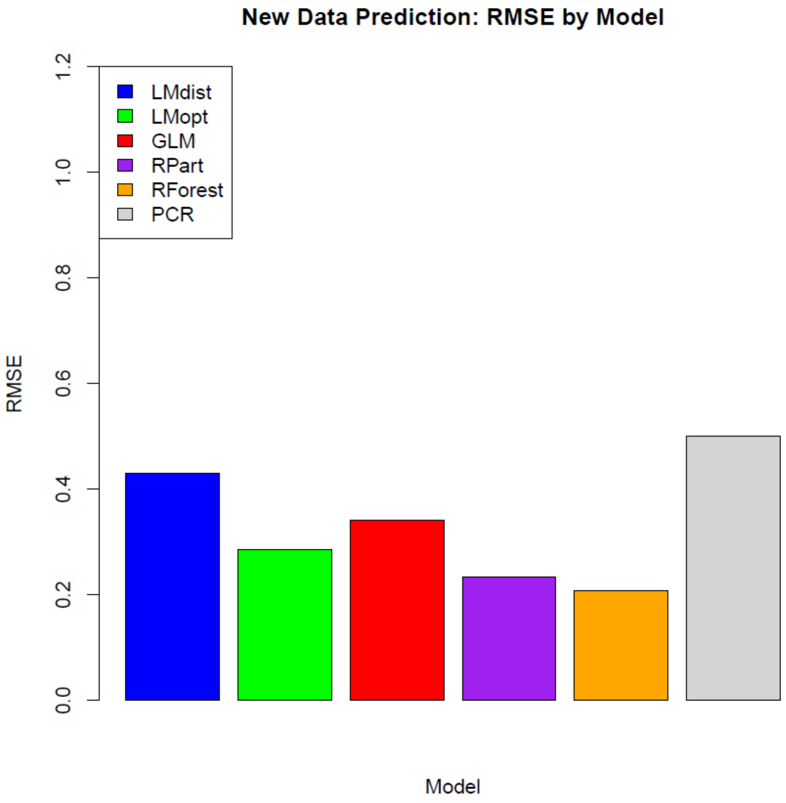
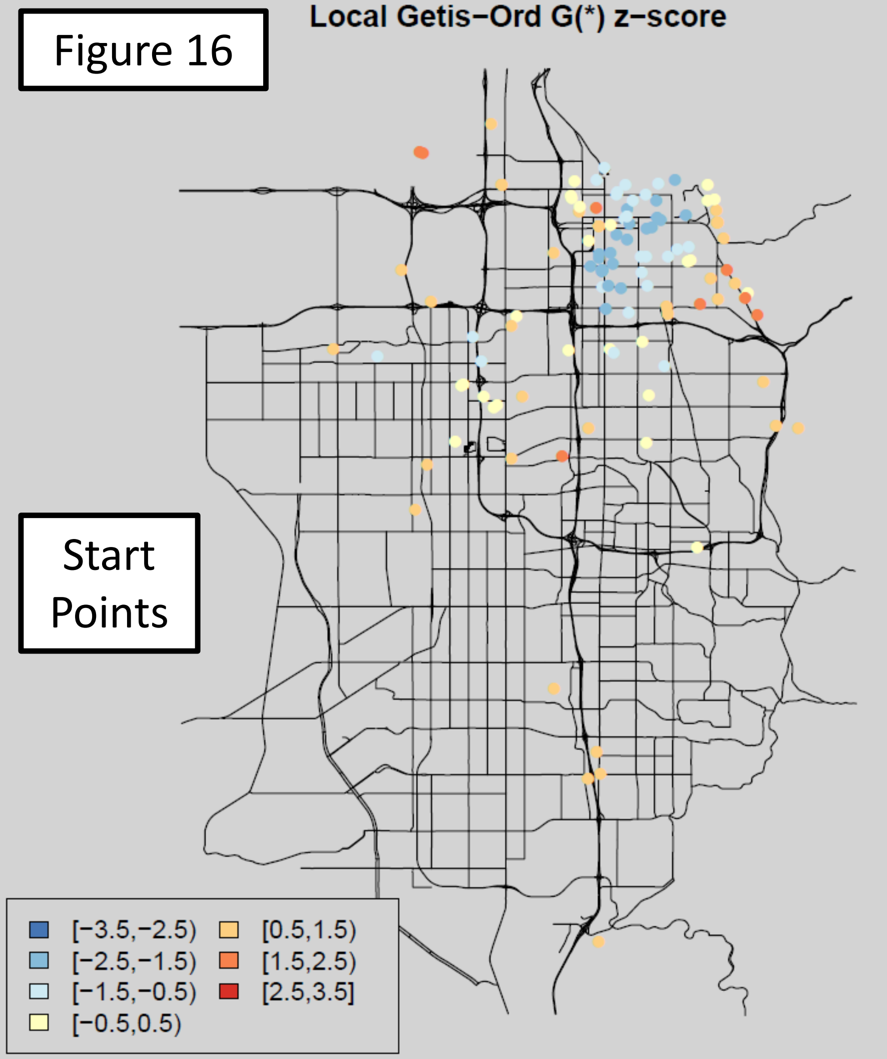
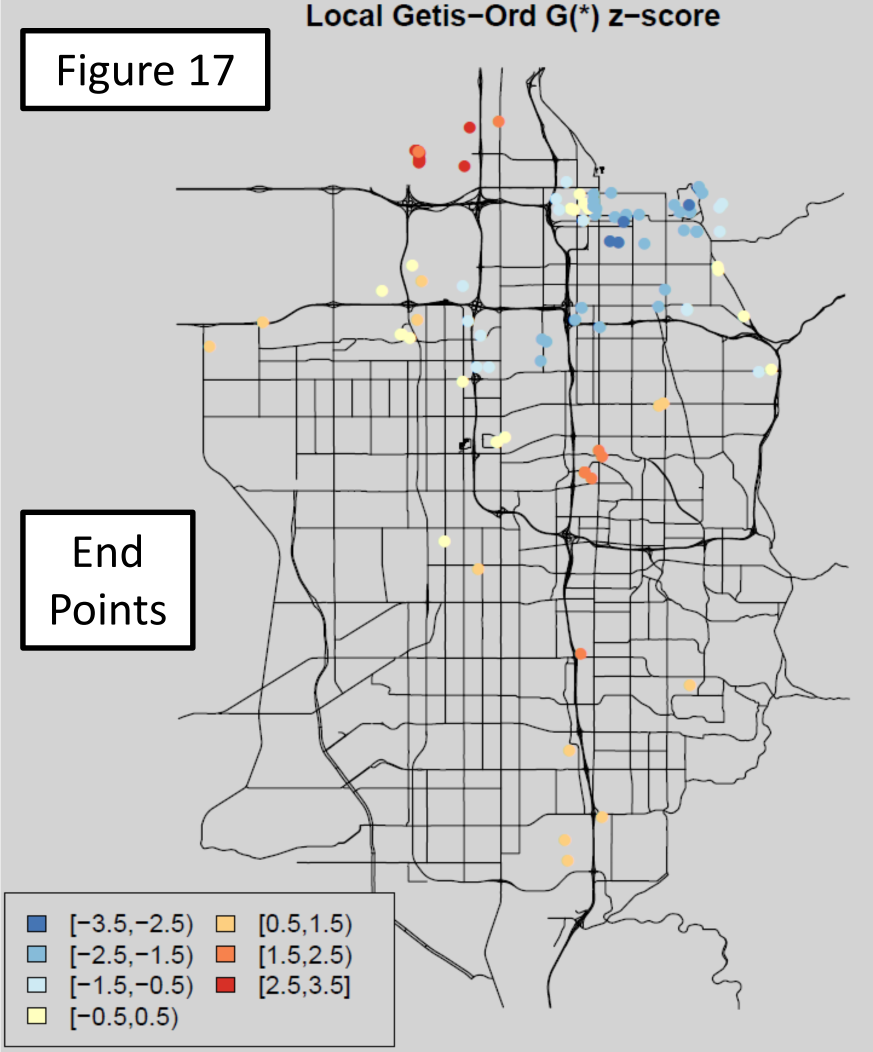


Figure 16

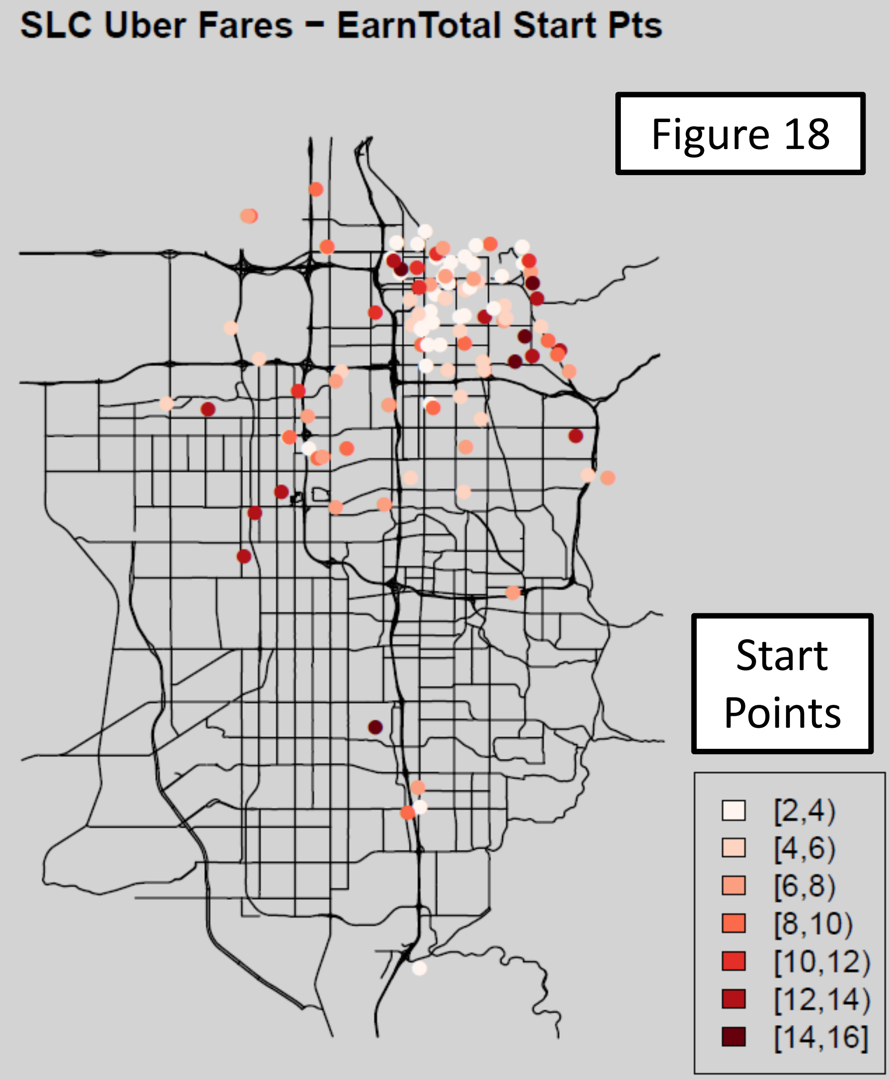


Start Points

Figure 17



End Points



Uber EarnTotal (Ordinary Kriging)

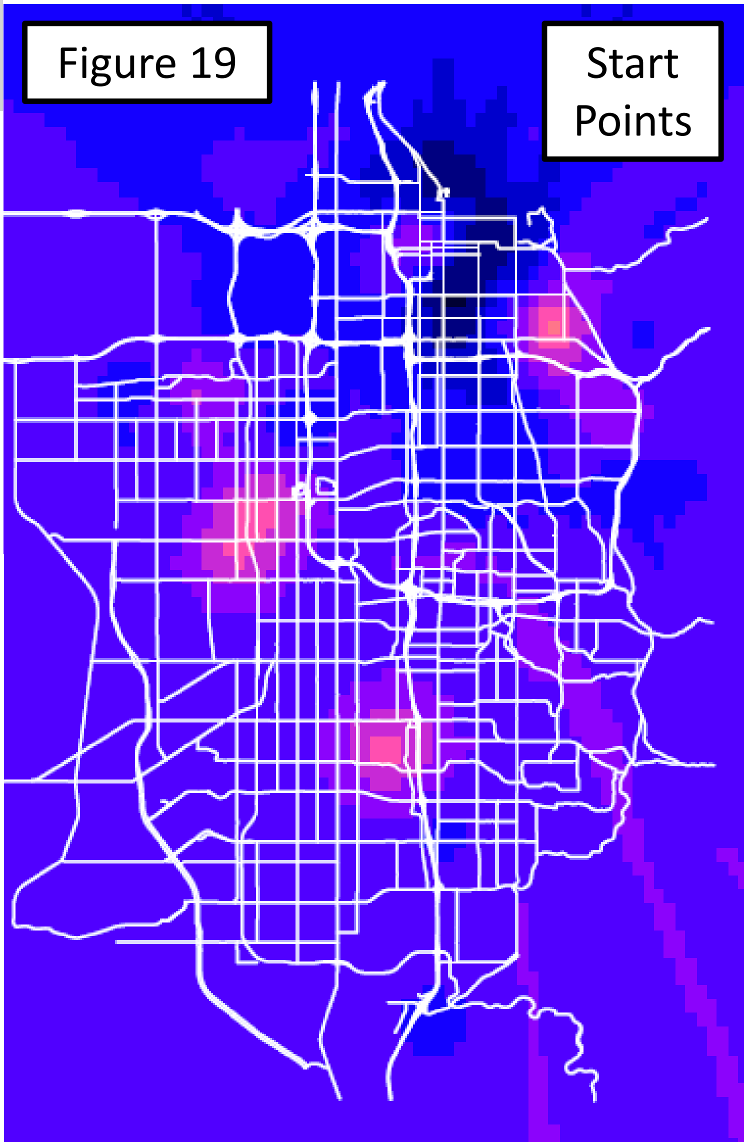
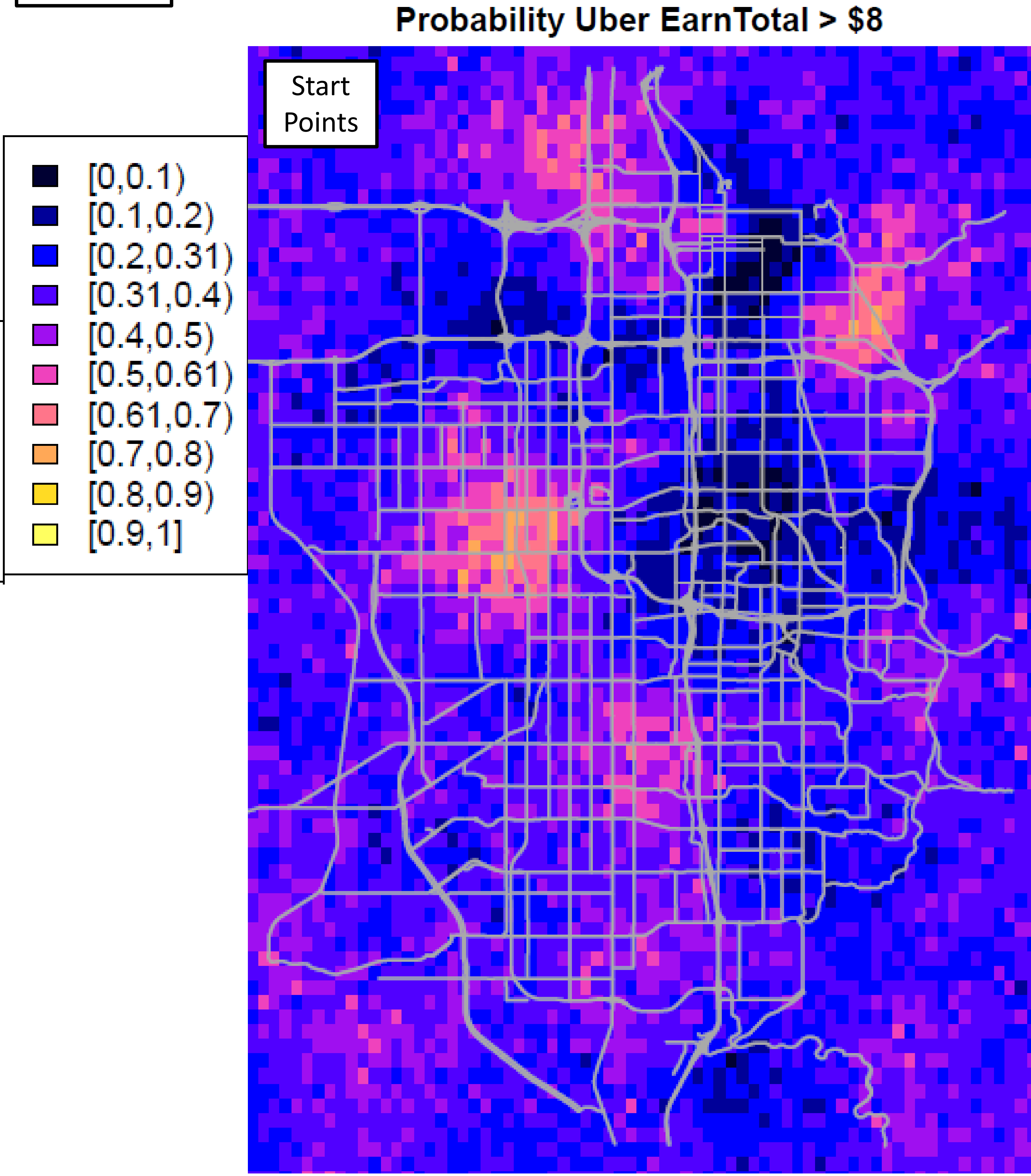


Figure 20


Probability Uber EarnTotal > $8
Start Points
[0,0.1)
[0.1,0.2)
[0.2,0.31)
[0.31,0.4)
[0.4,0.5)
[0.5,0.61)
[0.61,0.7)
[0.7,0.8)
[0.8,0.9)
[0.9,1]

SLC Uber Fares – Earn Total End Points

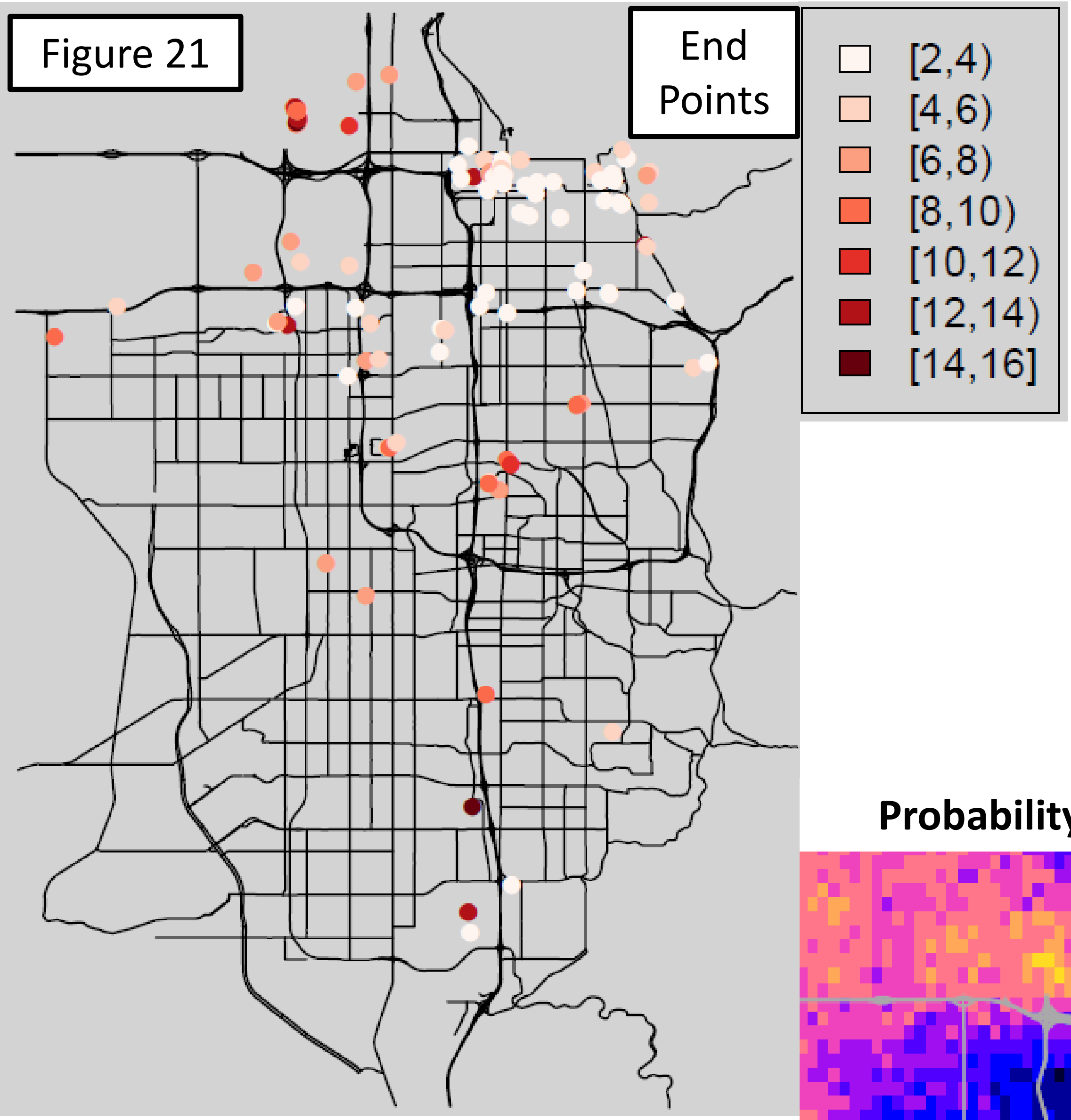


Figure 21

Probability Uber EarnTotal > $8

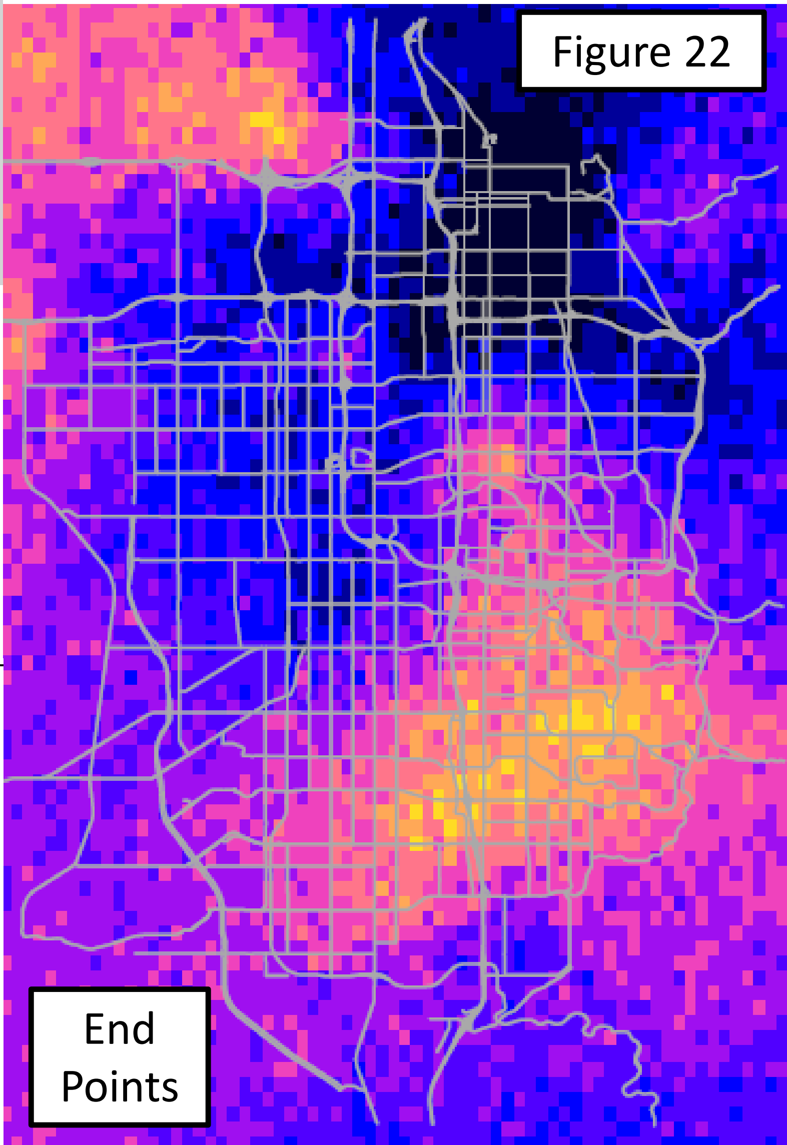


Figure 22

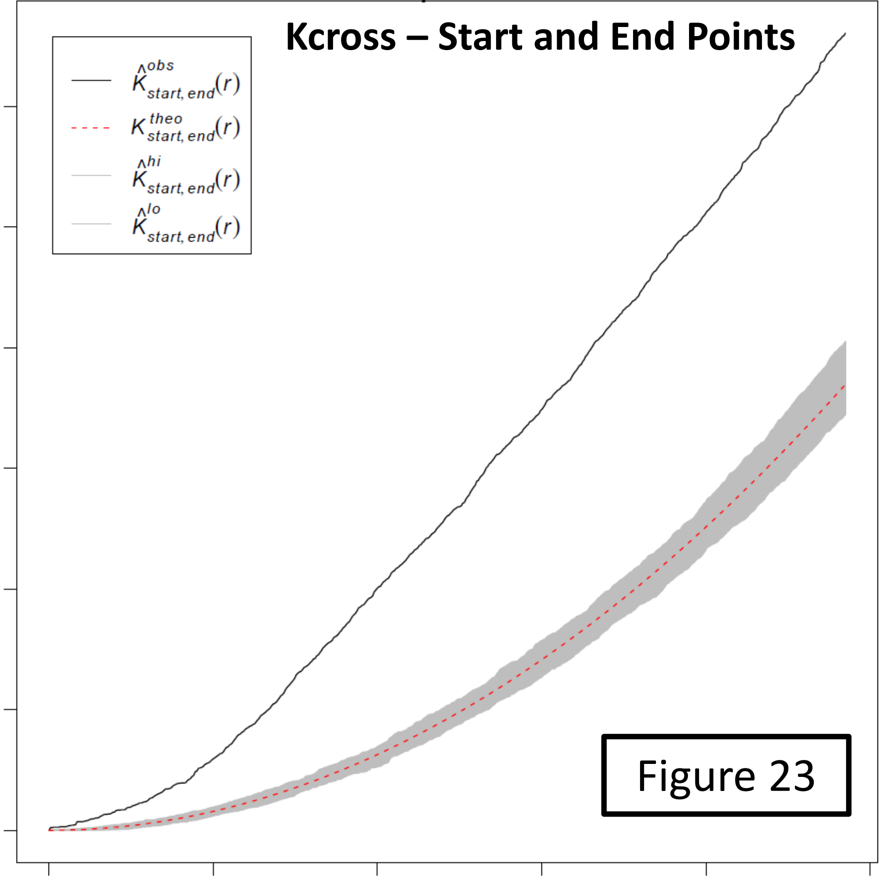


Figure 23

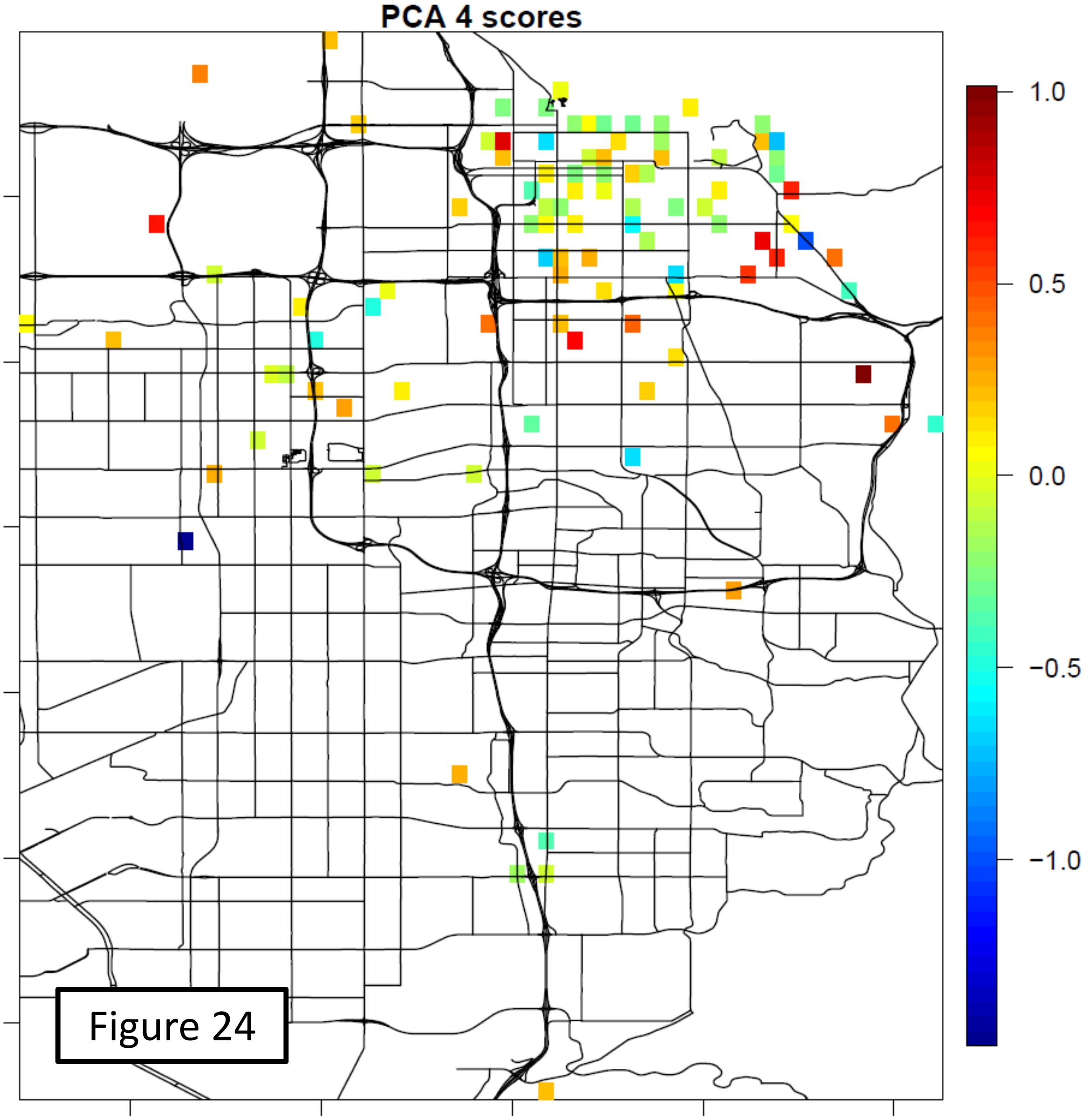


Figure 24