Traffic Pattern Exploration Using New York Yellow Taxi Dataset

Final Report for CMSC 12300 - Fast & Furious

1. Basic Information about Dataset
2. Hypotheses
3. Generate Index

To extract key differences between trips, we focus on three aspects – weather, location and pickup time. Using a sample subset generated from 12-month dataset from 2015/07 to 2016/06, we transform continuous variable into categorical values, calculate the distribution of weather, location and pickup time and normalize these variables in range (0, 1). The normalized values are considered as “index” in the following report and served as the basis of single trip and matching pair analysis.

1. Weather Index

2. Location Index

3. Pick-up Time Index

The code and result of this part can be found via ../code/time\_index .

Although there are both pick-up time and drop-off time in the dataset, only pickup-time is utilized in generating index for simplicity. The pick-up time index can be further divided into three sub-indices – month index, weekday index and hour index, which represent the month, the day of the week and the hour a trip took place respectively. Sub-indices are calculated by splitting the corresponding time variable from original data strings.

1. Single Trip
2. Matching Pair
3. Most Similar Trips
4. Causality

The code of this part can be found via ../code/matching pair/causality.

* 1. Algorithm

In this section, we explore the relationship between one single index given control of other five index. Based on the most similar trips we found, we establish a simple linear regression model for *X: differences in index* and *y: differences in travelling time* for all the trips using the formula as following:

The regression coefficient reflects the impact of index on travelling time, and as we have normalized index in the previous sections, the result of regression coefficient can be compared directly between different indices.

* 1. Method

MapReduce method with two MRstep is utilized. The first step is to extract index and travelling time of the compared trip and its counterpart, calculate differences of index and travelling time, means, sums of squared difference and sums of produce differences. The second step yields regression coefficients based on the parameters from the first step.

* 1. Result

Due to the capacity of clusters in Google Clouds, we use a subsample of 5000 trips to be matched, and another subsample of 50000 trips as matching pool. Using 11 cores to run the jobs, it took about 40 min to finish the fixed\_effect task. Regression coefficients are calculated inside Virtual Machine.

The simple regression result for each index are close: they have nearly identical coefficients, which doesn’t surprise us as we normalize all the index within range(0,1). The absolute values of regression coefficient are small, indicating that when other five indices are controlled, the difference of a single index can hardly influence the difference of unit travelling time.

1. Prediction
2. Passenger Privacy
3. Challenge
4. Conclusion