Where Can I Find My Next Passenger?

Kenneth Ezirim [1] Ted Brown [2] Shweta Jain [3]

[1]Graduate Center, City University of New York

[2]Queens College, City University of New York

[3]York College, City University of New York

November 19th, 2014

Outline

Introduction

- Taxi are common means of transportation in Manhattan
- Approx. 135k vehicles (New York City TLC 2011 and 2012 Annual report)
- about one half million rides a day
- 2011 and 2012 trip records over 350 million records

Motivation

Understanding the following:

- Distribution of time between trips or **vacant time**

What is Vacant time?

It is the time between a drop-off and a pickup. In other words, the period when a driver has no passenger on board.

- Variation of vacant time across different locations
- Advantages of using optimized vacant time

Data Features

Trip record features:

- Unique identifiers - medallion, shift number and trip number
- Geospatial Information of Pickup and Dropoff locations
- No information on trip route
- Information on fare, tolls and tips

Anomalies:

- Zero latitude and longitude values
- GPS distance $>$ Trip distance
- Discrepancy in total fare amount etc.

Feature Selection

- Selected only trips with pickup and dropoff locations in Manhattan
- Trip records selected devoid of anomalies
- Random sampling due to size of dataset
- Trip features include medallion, shift number, trip number, pickup and dropoff latlng, pickup and dropoff datetime, fare amount
- Dataset housed on Hive DB with full support for map/reduce

Definitions

Important definitions:

- **Trip**: movement of passenger from pickup location to drop-off location
- **Shift**: sequence of taxi trips, with intermediate wait times.
- **Sector**: a geographically mapped area with boundaries represented with a polygon. i-th sector is denoted as s_i

Given two sector s_i and s_j, trip time is denoted as $t(i, j)$ and wait time is denoted as $\tau(i, j)$.
Transition from a dropoff sector s_i to a pickup sector s_j is denoted as $tr(i, j)$.

Area Partitioning using Taxi Data

- We used dropoff latitude and longitude coordinates to partition Manhattan into sectors.
- There are many ways of achieving this, one of which is to:
  - Select an arbitrary number of cluster centroids, for instance, 75.
  - Then find cluster centroids for sectors using K-Means clustering algorithm.
  - Use Voronoi (Tessellation) algorithm to generate polygon for each sector.

Area Partitioning using Geospatial information

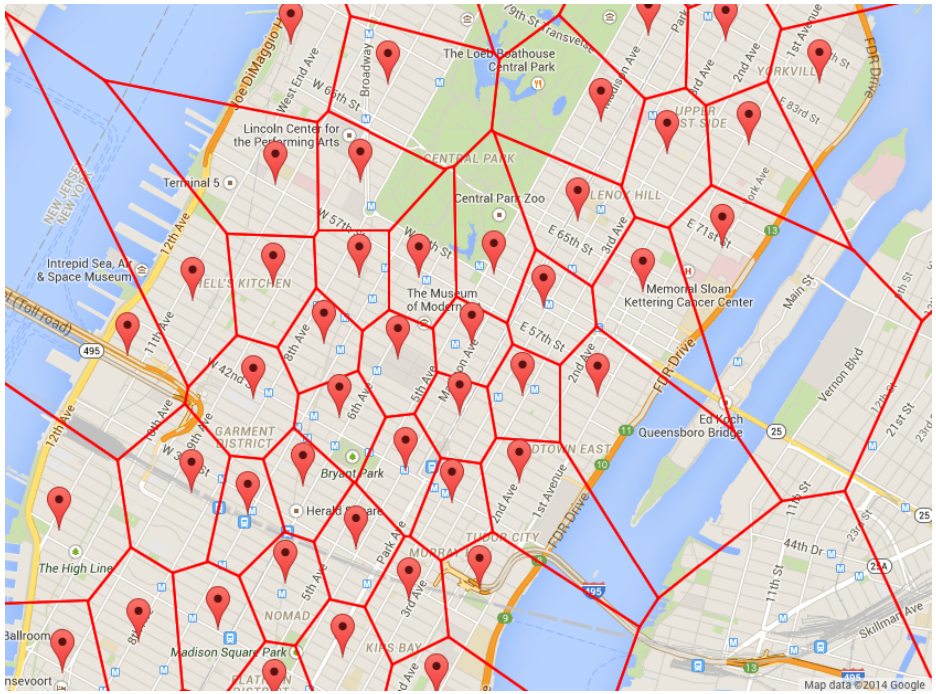


Figure 1 : Snapshot of partitioning using Geospatial information from NYC TLC Taxi Data 2011. Data used in creating the partitions were trip records for the month April 2011.

Vacant Time Distribution

Follows Lévy Distribution, with probability density distribution approximated by this function

$$f(\tau;\mu,c) = \sqrt{\frac{c}{2\pi}}\frac{e^{-\frac{c}{2(\tau-\mu)}}}{(\tau-\mu)^{3/2}} \tag{1}$$

over the domain $\tau > \mu$ with $c = 2.6$ and $\mu = 0.9$

Vacant Time Distribution

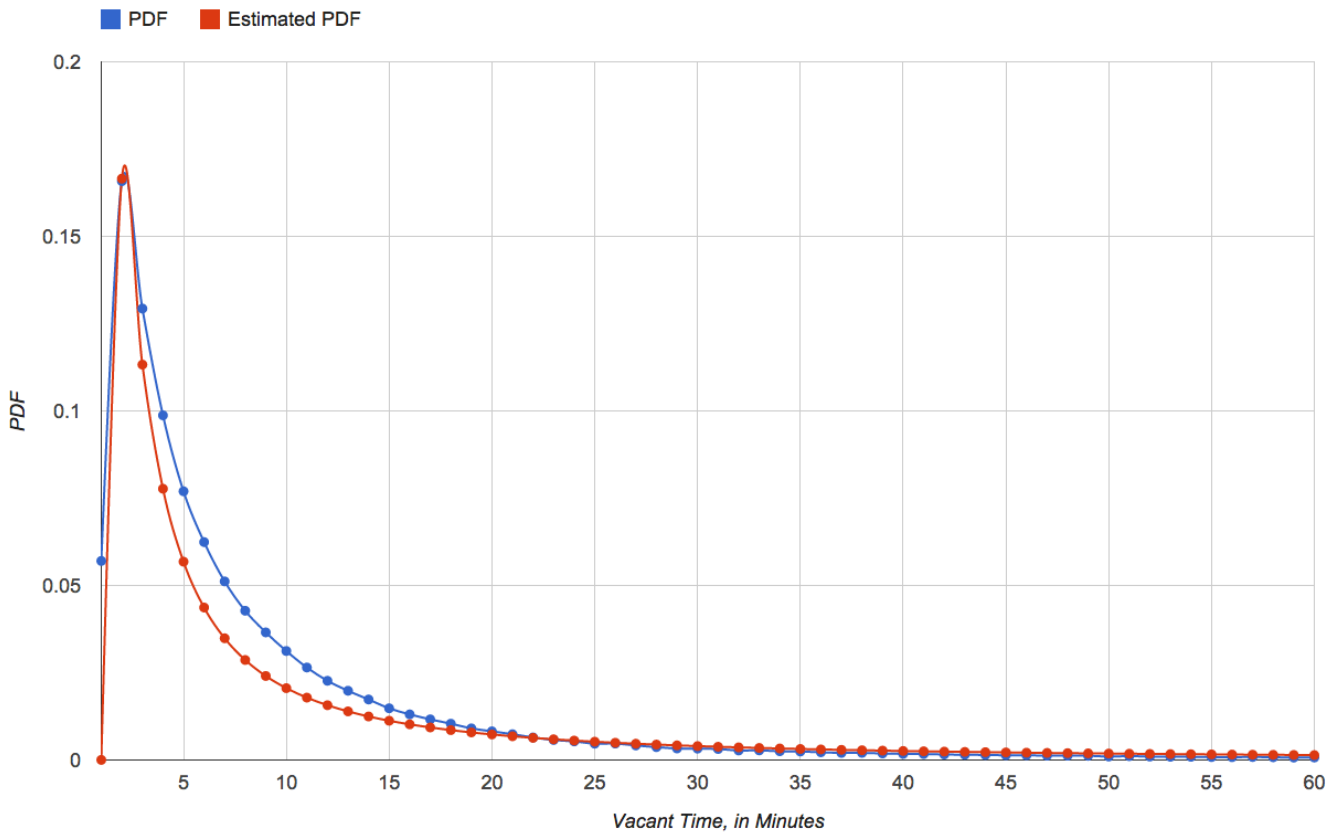


Figure 2 : Vacant Time is distributed according to Lévy Distribution.

Visualization

Sample Study

Date: April 4, 2011

26,784 unique Taxi shifts

304,833 trips

- Dropoff Distribution and Dropoff-Pickup Transition Distribution ▸ Link
- Vacant time distribution ▸ Link
- Trip time distribution ▸ Link
- Fare Amount Distribution ▸ Link

Shift Optimization

Taxi shifts can be optimized to:

1. Increase the number of trips made during a shift
2. Increase the average revenue generated by taxi cabs
3. Reduce the amount of time spent by taxi cabs hunting for passengers.

Shift Optimization

Optimization based on vacant time τ

$$j = \underset{k \in [1,n]}{\arg\min}\{\tau(i,k)\} \tag{2}$$

Driver chooses sector s_j that has the shortest vacant time.

Shift Optimization

Optimization based on possibility of a pickup in a sector.
Given a large dataset, the probability of a pickup in s_j with the last dropoff sector in s_i, is

$$p_{ij} = tr(i,j) / \sum_{k=1}^{n} tr(i,k) \tag{3}$$

where n is the number of sectors and $0 \leq p_{ij} < 1$.

Shift Optimization

Assume a driver after dropoff in s_i and wants to go to s_j because

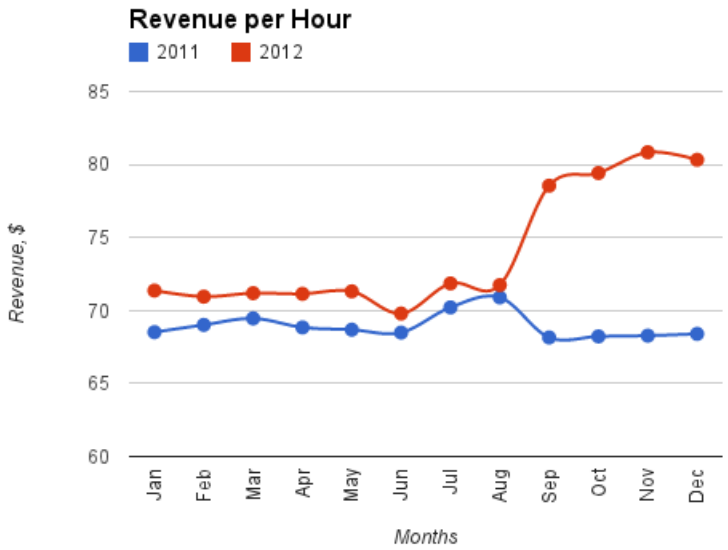
$$j = \arg\max_{k \in [1,n]} \{p_{ik}\} \tag{4}$$

Finding another trip another trip with trip time $t(k, l)$ on the way to s_j is good for the driver, that is

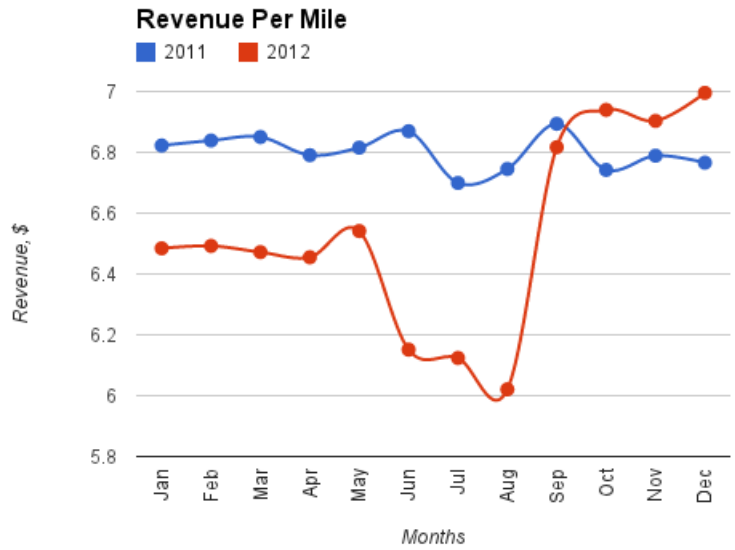
$$\tau(i, k) + t(k, l) + \tau(l, j) \leq \tau(i, j) \tag{5}$$

But $t(k, l)$ is difficult to predict because it depends on the passenger.

Analytical Results



(a) Revenue per Hour



(b) Revenue per Mile

Figure 3 : Plots showing the variation of revenue per hour and revenue per mile for the year 2011 and 2012.

Analytical Results

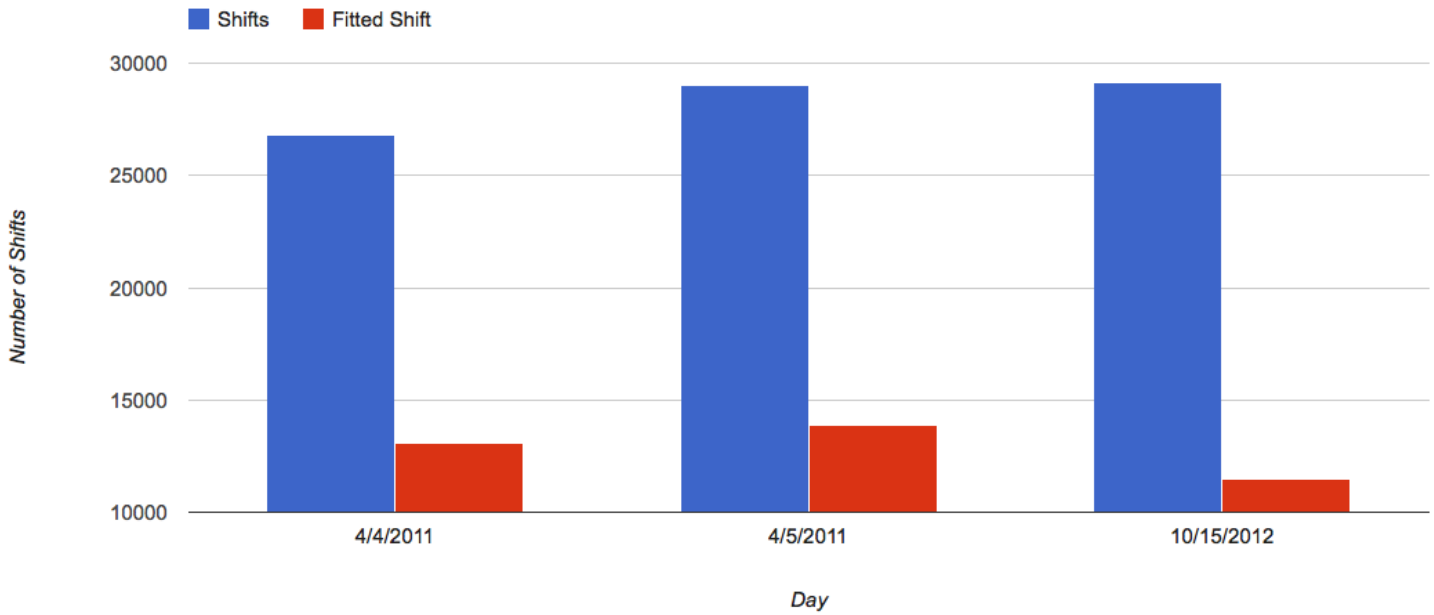


Figure 4 : Illustration of the number of shifts that qualified for optimization given that only a single trip can be fitted in any acceptable vacant time period.

Analytical Results

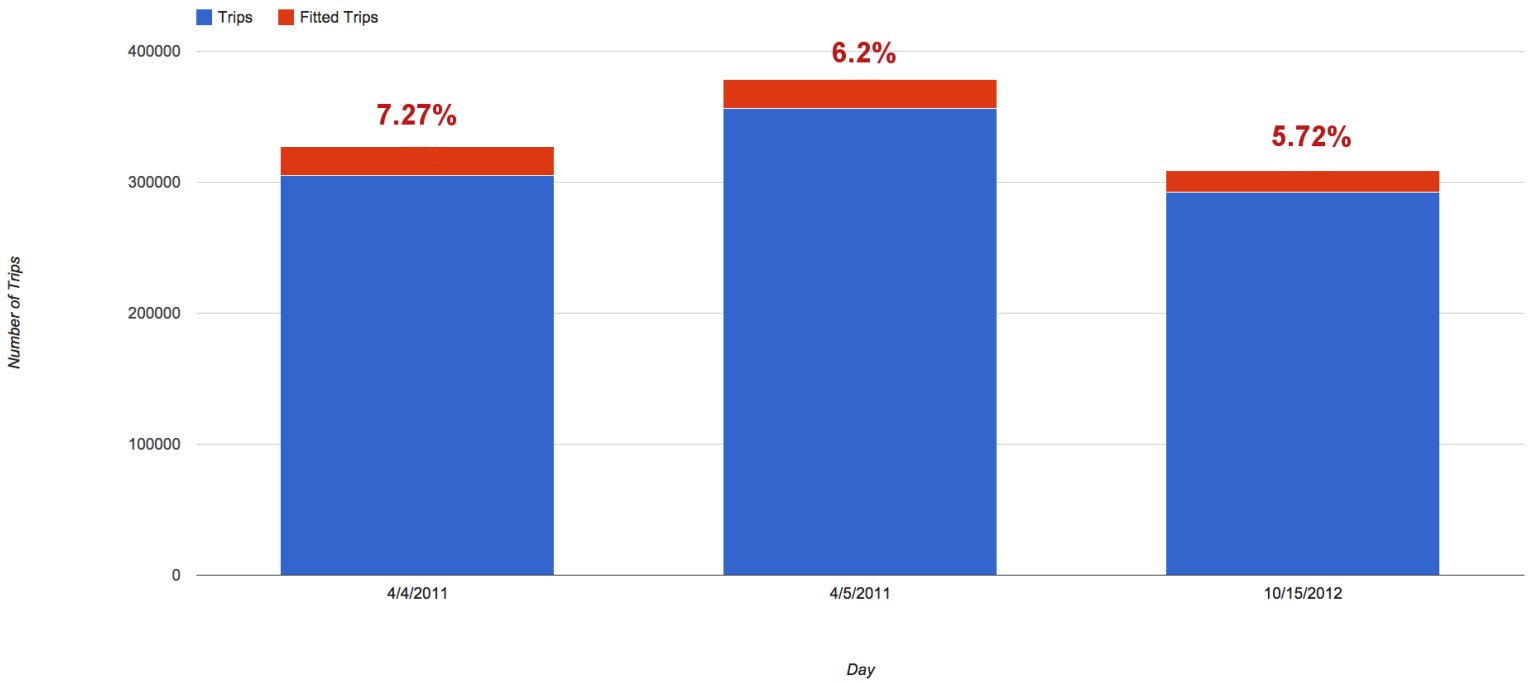


Figure 5 : Illustration of the number of trips recovered via shift optimization given that only a single trip can be fitted in any acceptable vacant time period.

Analytical Results

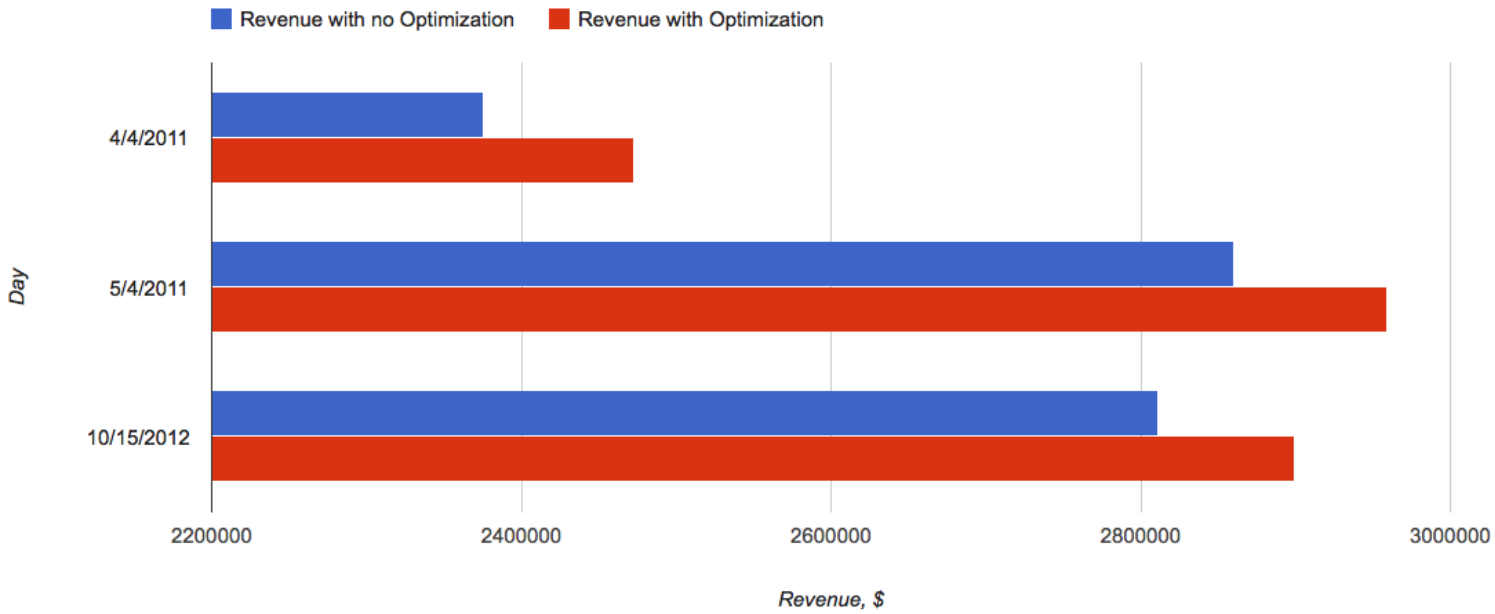


Figure 6 : Illustration showing the impact of shift optimization on the revenue generated by taxi cabs.

Advantages

Vacant time information can help select best sector to find a passenger.
This will lead to:

- Increase in revenue generated
- Reduction of latency in finding passengers
- Reduction of greenhouse effect

Conclusions

- Vacant time exhibits Levy distribution
- Another approach of partitioning using Taxi data
- Benefits of vacant time information

Future works:

- A prediction model based on time and location
- A cloud-computing assisted smart phone app to assist drivers in locating passenger