Summary of “**Modeling Taxi Demand and Supply in New York City Using Large-Scale Taxi GPS Data**”

(by Ci Yang and Eric J. Gonzales)

This paper aims to show how big data collected from taxis can be used to model taxi demand and supply. It uses 10 months of taxi trip data from New York City in order to come up with a model that can identify locations and times of the day when there is a mismatch between the availability of taxis and the demand for them.

It is suggested that processing and integrating data with a Geographic Information System can provide more insight into transportation in cities and the role of the taxi market within the transportation system.

The data that was used amounts to 147 million taxi trips in NYC, between February 1, 2010 and November 28, 2010. Since the data was too large (40GB) to use tools such as Excel, they used SQL server. They aggregated the location of the pick-ups and drop-offs by NYC census tract and the times were aggregated by hour of the day, which resulted in the number of taxi pick-up counts in each census tract per hour. Population, education, median age, median income per capita, employment by industry sector and transit accessibility are the 6 important variables that they included.

Mehodology:

1. In order to find an appropriate model, they first compared the **Quasi-Poisson Distribution** and the **Negative Binomial Distribution**. For this, they had to compare the mean and the variance of taxi pick-up counts. A separate model was estimated for each hour, and the comparison of mean and variance was considered within each hourly aggregation. The goodness of fit parameter, R^2, was used and it revealed that the quadratic function is better for relating the variance and the mean, which means that the negative binomial distribution is more suitable for the counts of taxi pick-ups.
2. Then, they compared **Poisson Regression** with **Negative Binomial Regression**. This was done in order to compare the fit of the models with the explanatory variables. In this sense, different criteria were used: the Akaike Information Criterion, the Goodness-of-Fit Test, the Sum of Model Deviances and the Likelihood Ratio Test.

Results:

A comparison was made between the results using a conventional Poisson regression model with the results of the negative binomial regression, for which a separate model was estimated for each hour of the day.

The results showed that regardless of model specification, the taxi supply, education, and transit accessibility are always significant determinants of taxi pick-up demand at all times of the day.

Due to the differences between the models, the negative binomial regression seems like the most appropriate model, although not perfect.

In order to identify the locations and times of day when there may be a mismatch between taxi demand and supply, they looked at the Pearson residuals from the models. For a single hour of the day, the residuals for each census tract in the city were be mapped in order to visualize the spatial distribution of the model errors.