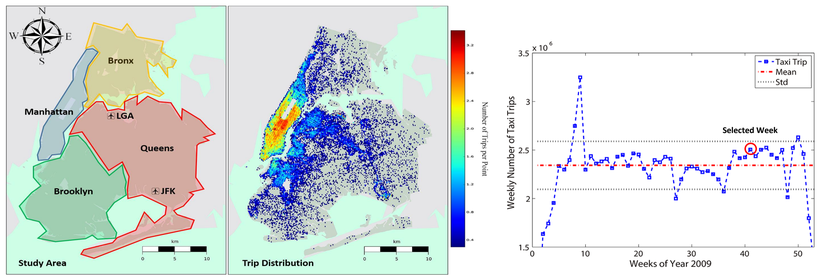
Using Large Scale Taxicab Data to Estimate Link Travel Time, Predict Demand and Measure System Efficiency

Xianyuan Zhan, Research Assistant, Purdue University, **Xinwu Qian** Research Assistant, Purdue University
Satish V. Ukkusuri, Professor, Purdue University

Introduction

- **The era of big data**
  - Advance in sensing technologies
  - Development of large scale pervasive computing infrastructure
- **Big data and transportation engineering**
  - Reconsider traditional research problems
  - Make infeasible problems feasible
- **In this work**
  - Using large scale taxi data from NYC
  - Taxi Ridership analysis
  - Link travel time estimation
  - Taxi system efficiency

Taxi Data



Study area and trip distribution

Data Selected

- Source: NYCTLC
- Information: geo-coordinate and timestamps for pick-up and drop-off locations, trip distance, trip fare and the number of passengers
- Data extracted: October 5th to October 11th, 2009
- Around 500,000 daily trips

Spatial Variation of Taxi Ridership

Motivation

- Statistical analysis of taxi ridership
- Trips are varying spatially
- The effects of determinants is *nonhomogeneous*

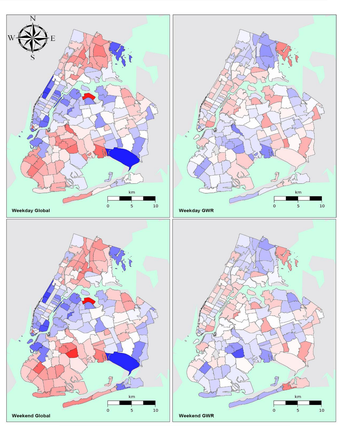
Methodology

- ***Geographically weighted regression***

$$y_i = a_{i0} + \sum_{k=1}^{n} a_{ik} x_{ik} + \epsilon_i$$

$$w_{ij} = \begin{cases} \exp\left[-0.5\left(\frac{d_{ij}}{b}\right)^2\right], d_{ij} < b \\ 0, otherwise \end{cases}$$

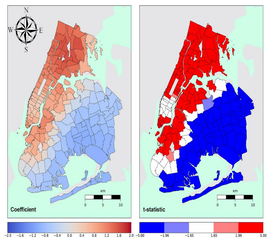
- Dependent variable: Taxi ridership
- Independent variables: commuting time, population, land use, median income, road density, subway accessibility



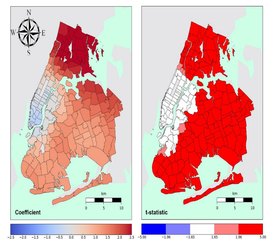
Model residuals: OLS(left) vs GWR(right)

Key Findings

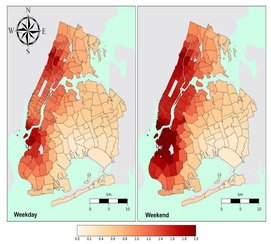
- Urban form has significant impact on ridership
- GWR explains up to 90% of the variance and achieves good prediction
- Both coefficients and t-stats of determinants vary over space
- Failing to consider spatial variation will result in erroneous estimations of determinants



Median Income



Commercial Floor Area



Subway Accessibility

Link Travel Time Estimation

Motivation

- Accurate estimation of urban link travel time is essential for various applications in urban traffic operations and management
- Traditional approaches using fixed sensors: expensive and limited coverage
- Can we estimate link travel time using only partial information from taxi trips?

Methodology

- **Finite mixture distribution**

$$P(y^i|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{D}) = \sum_{k \in R^i} \pi_k^i(\boldsymbol{\mu}, \boldsymbol{\beta}, d_k) P(y^i|k, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$H(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{D}) = \prod_{i=1}^{n} \sum_{k \in R^i} \pi_k^i(\boldsymbol{\mu}, \boldsymbol{\beta}, d_k) P(y^i|k, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

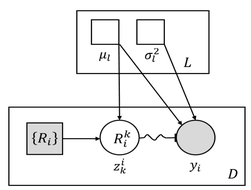
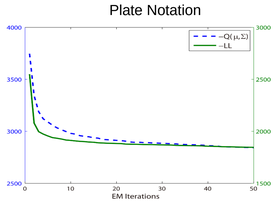


Plate Notation

- Solution approach: EM algorithm
- Test network: A 1175x1780cm rectangle area in Midtown Manhattan, with 136 nodes and 254 directed links



Algorithm Convergence

Key Findings

- Algorithm converges rapidly and entire estimation takes less than 15 minutes
- Robust estimation results: MAPE controlled under 30%
- Can be extended easily as a Bayesian mixture model by making use of historical data

Example Estimation Results for Wednesday

Day	Criteria	Time Period 9:00 - 9:30	13:00 - 13:30	19:00 - 19:30	21:00 - 21:30
Monday	MAPE	31.39%	29.41%	23.95%	23.03%
	$P_{95\%}$	87.16%	84.38%	74.25%	69.31%
	$P_{99\%}$	90.94%	90.07%	81.63%	74.59%
Tuesday	MAPE	26.34%	25.91%	29.17%	20.52%
	$P_{95\%}$	81.70%	84.03%	81.18%	74.25%
	$P_{99\%}$	87.37%	88.78%	87.06%	82.42%
Wednesday	MAPE	26.49%	27.56%	26.10%	23.63%
	$P_{95\%}$	81.79%	85.71%	80.36%	73.82%
	$P_{99\%}$	88.08%	92.69%	88.04%	80.91%
Thursday	MAPE	28.45%	28.62%	24.48%	25.15%
	$P_{95\%}$	84.88%	83.02%	80.04%	77.26%
	$P_{99\%}$	91.15%	88.81%	87.03%	83.27%
Friday	MAPE	25.12%	27.40%	26.61%	24.73%
	$P_{95\%}$	86.96%	82.14%	85.64%	76.54%
	$P_{99\%}$	92.36%	87.38%	90.91%	83.24%
Saturday	MAPE	20.78%	27.01%	26.86%	24.22%
	$P_{95\%}$	76.30%	77.17%	76.29%	72.69%
	$P_{99\%}$	80.52%	83.26%	81.43%	76.31%
Sunday	MAPE	22.98%	26.36%	24.57%	25.61%
	$P_{95\%}$	68.83%	78.93%	75.00%	72.64%
	$P_{99\%}$	75.76%	85.85%	79.49%	78.07%

Estimation Results Validation

Efficiency of Taxi Service System

Motivation

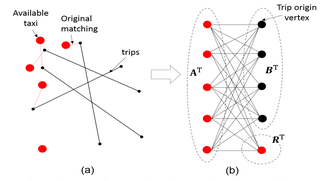
- Vacant taxi trips lead to unnecessary externalities
- How to quantify the efficiency of the system performance and how far is the current system from the theoretically optimal one?

Notions

- Optimal Matching: finding the optimal matching strategy between each pair of taxi driver and passenger
- Unweighted trip integration: results in the minimum number of taxis required to satisfy all the trips.
- Weighted trip integration: results in minimum total matching cost while achieving minimum number of taxis satisfying all the trips

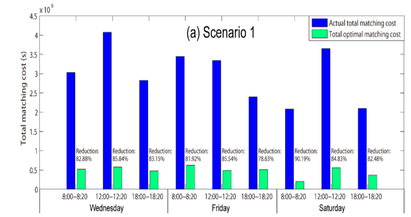
Methodology

- **Optimal matching**: minimum weight perfect bipartite matching using Hungarian method
- **Unweighted trip Integration** : maximum bipartite matching with max-flow algorithm
- **Weighted trip integration**: minimum weight bipartite matching using Hungarian method

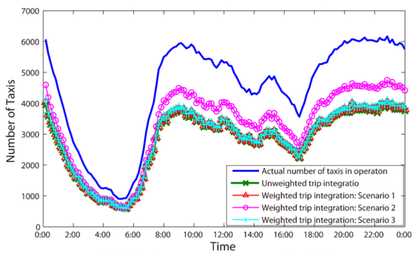


Graph transformation for optimal matching

Graph transformation for weighted trip integration



Example result for optimal matching



(b) Trip integration results for Friday

Example result for weighted trip integration

Key Findings

- Optimal matching can reduce up to 90% of taxi idle time, 87% of vacant trip distance and 82% of revenue loss.
- Using 2/3 of current taxis can serve all observed trips
- Idle time, vacant trip distance and revenue loss can be reduce to half with fewer taxis
- System level information is critical to improve the system efficiency

References

- Qian, X., S.V. Ukkusuri, 2015.Spatial variation of the urban taxi ridership using GPS data. *Applied geography*, 59, 31-42.
- Zhan, X., S.V. Ukkusuri, 2015. Probabilistic Urban Link Travel Time Estimation Model Using Large-Scale Taxi Trip Data. Transportation Research Board 94th Annual Meeting, 15-4054.
- Zhan, X., X. Qian and S.V. Ukkusuri, 2015. A Graph Based Approach to Measure the Efficiency of Urban Taxi Service System. Submitted to *Transportation Research Part B: Methodological.*