Urban Link Travel Time Estimation Using Large-scale Taxi Data with Partial Information

Xianyuan Zhan*

Satish V. Ukkusuri*

*Civil Engineering, Purdue University

24/04/2014

PURDUE UNIVERSITY

PURDUE ENGINEERING

Outline

- Introduction
- Study Region
- Link Travel Time Estimation Model
  - Base Model
  - Probabilistic Model
- Numerical Results
- Conclusion
- Questions/Comments

PURDUE UNIVERSITY

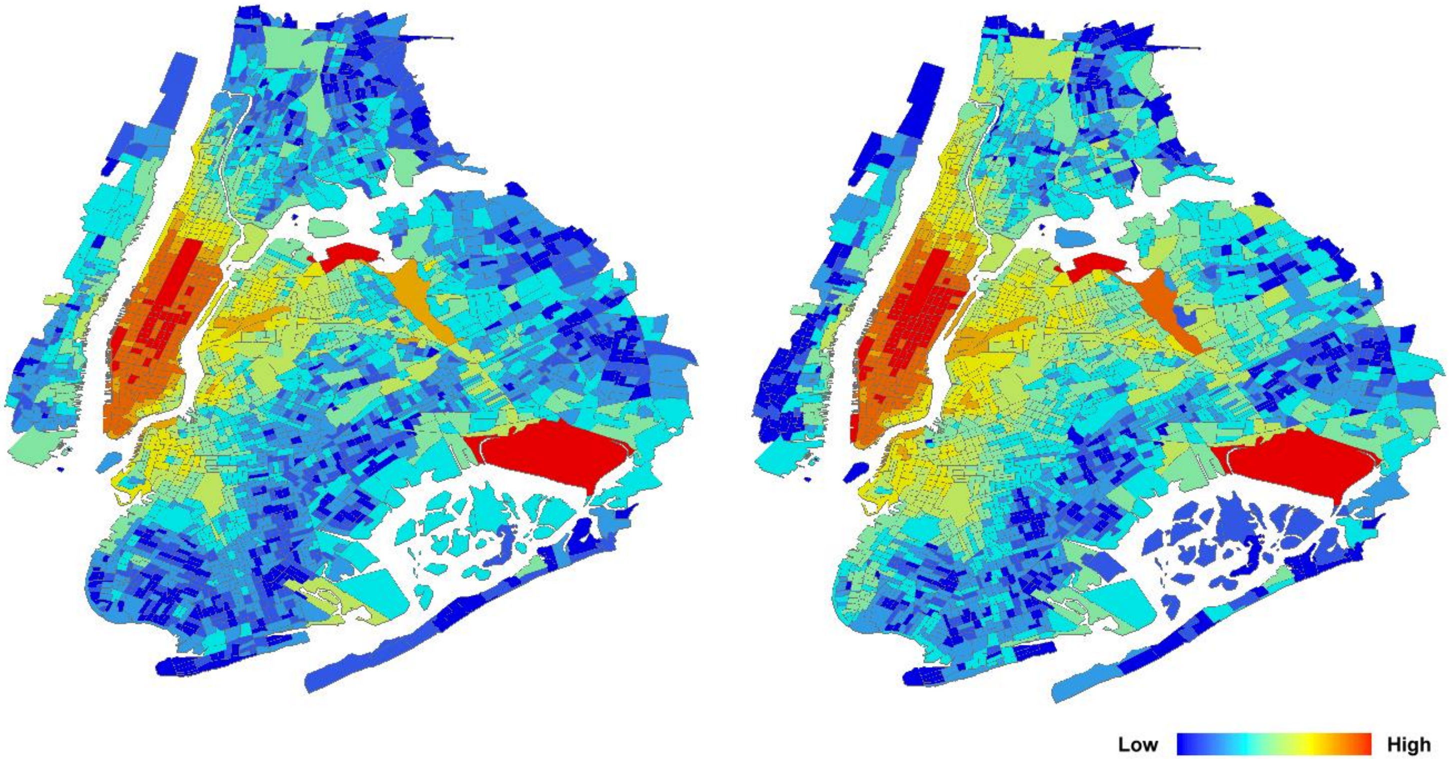
Introduction

- New York City has the largest market for taxis in North America:
  - **12,779** yellow medallion (2006)
  - Industrial revenue **$1.82 billion** (2005)
  - Serving **240 million** passengers per year
  - **71%** of all Manhattan residents' trips
- GPS devices are installed in each taxicab
- Taxi data recorded by New York Taxi and Limousine Commission (NYTLC)
- **Massive amount of data!**
  - **450,000** to **550,000** daily trip records
  - More than **180 million** taxi trips a year
  - Providing a lot of opportunities!

PURDUE UNIVERSITY

Introduction

Taxi trips in NYC



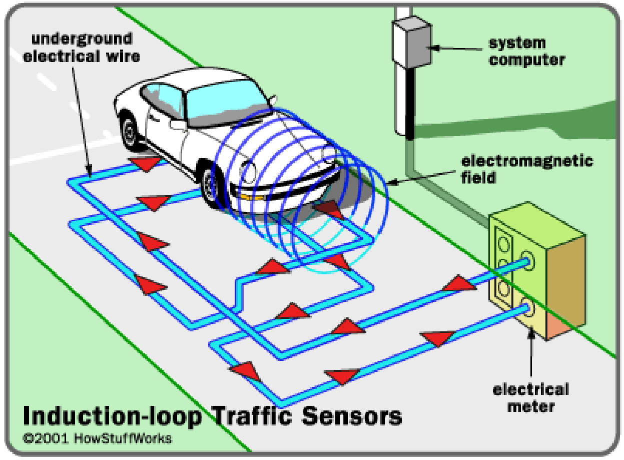
Trip Origin

Trip Destination

Introduction

Estimating urban link travel times

- Traditional approaches:
  - Loop detector data
  - Automatic Vehicle Identification tags
  - Video camera data
  - Remote microwave traffic sensors
- Why taxicab data?
  - Novel large-scale data sources
  - Ideal probes monitoring traffic condition
  - Large coverage
  - Do not need fixed sensors
  - Cheap!



PURDUE UNIVERSITY

Introduction

The data

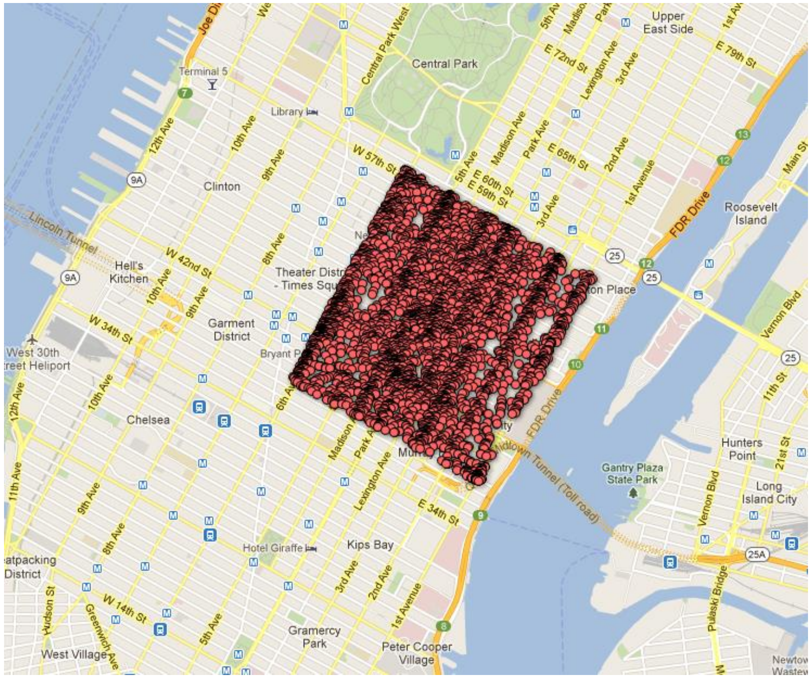
- NYTLC records taxi GPS trajectory data, but not public
- Only trip basis data available
  - Contains only OD coordinate, trip travel time and distance, etc.
  - Path information not available
  - Large-scale data with partial information

The problem

- Given large-scale taxi OD trip data, estimate urban link travel times
- Sub-problems to solve:
  - Map data to the network
  - Path inference
  - Estimate link travel time based on OD data

Study Region

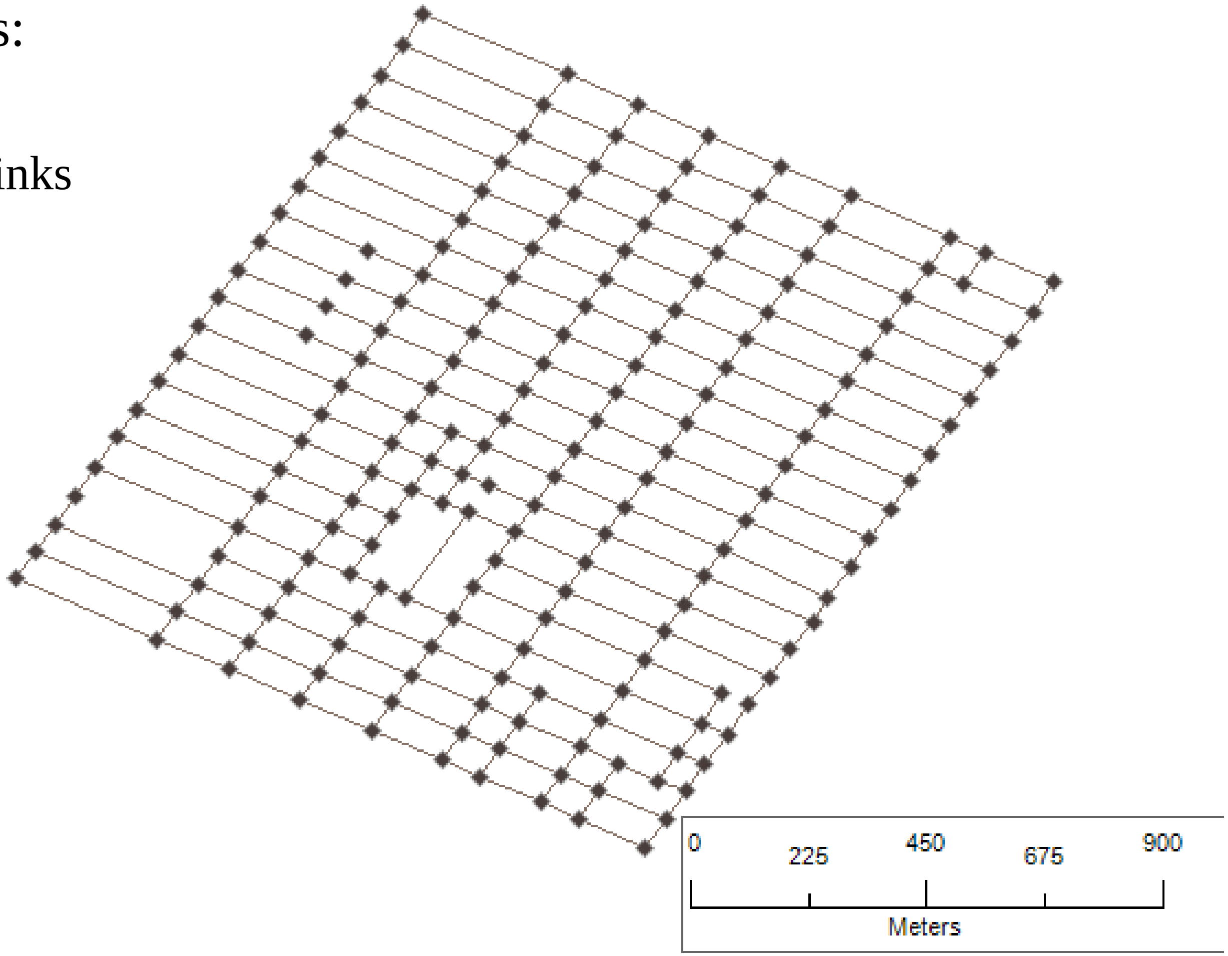
- 1370×1600m rectangle area in Midtown Manhattan
- Data records fall within the region are subtracted



Study Region

Test network

- Network contains:
  - 193 nodes
  - 381 directed links

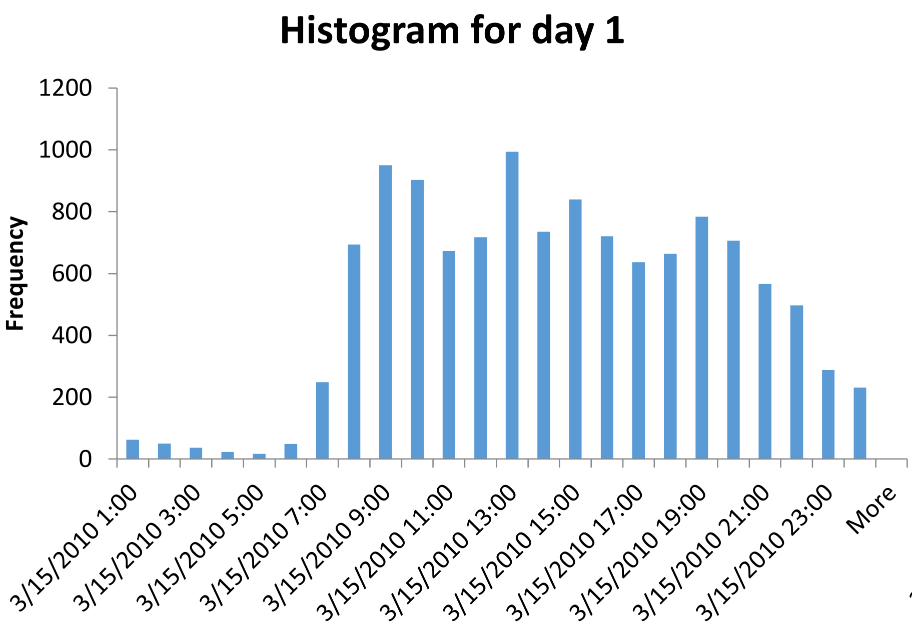


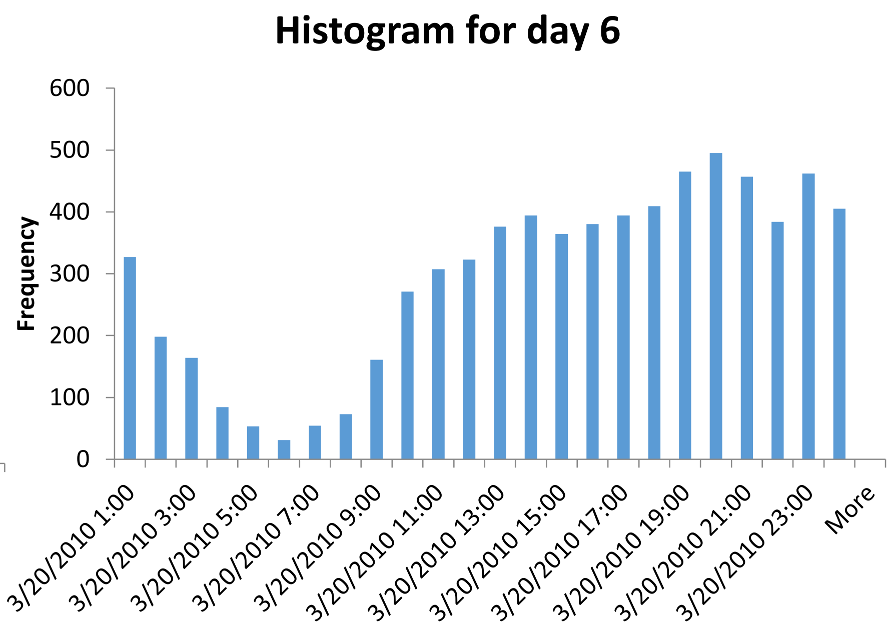
PURDUE UNIVERSITY

Study Region

Number of observations in the study region

- Day 1: Weekday (2010/03/15, Monday)
- Day 2: Weekend (2010/03/20, Saturday)





Base Model

Base link travel time estimation model*

- Hourly average link travel time estimations
- Direct optimization approach
- Overall framework: four phases

Data Mapping

Map origin-destination records on corresonding links

Construct Reasonable Path Set

Shrink algorithm search space into a manageable scale

Route Choice Model

MNL model based route choice considering trip time, distance

Link Travel Time Estimation

Estimate link travel times by solving an optimization problem

PURDUE UNIVERSITY

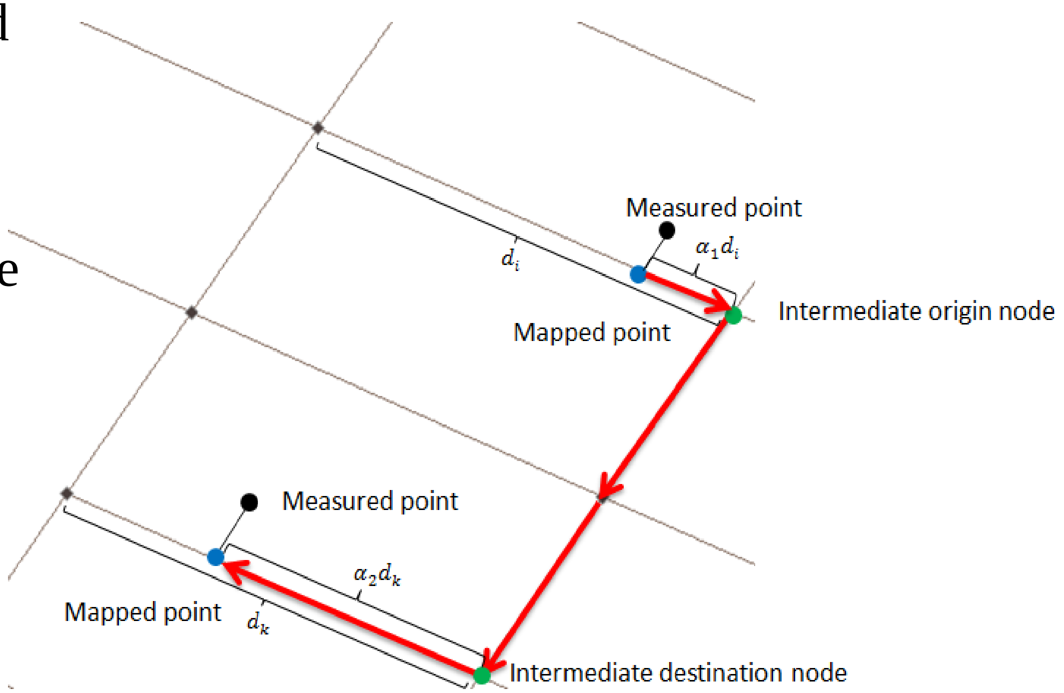
* Zhan, X., Hasan, S., Ukkusuri, S. V., & Kamga, C. (2013). Urban link travel time estimation using large-scale taxi data with partial information. *Transportation Research Part C: Emerging Technologies*, *33*, 37-49.

Base Model

Data mapping

- Mapping points to nearest links in the network

- Mapped point (blue) are used
- Identify intermediate origin/ destination nodes
- α_1, α_2 are defined as distance proportions from mapped points to the intermediate origin/destination node



Base Model

☐ Construct reasonable path sets

- Number of possible paths could be huge!
- Need to shrink the size of possible path set
- Use trip distance to eliminate unreasonable paths
- K-shortest path algorithm* (k=20) is used to generate initial path sets
- Filter out unreasonable paths (threshold: weekday 15%~25%, weekend 50%)

PURDUE UNIVERSITY

* Y. Yen, Finding the K shortest loopless paths in a network, Management Science 17:712–716, 1971.

Base Model

Route choice model

- Assumption:
  - Each driver wants to minimize both trip time and distance to make more trips thus make more revenue
- A MNL model based on utility maximization scheme

$$P_m(\vec{t}, d, \theta) = \frac{e^{-\theta C_m(\vec{t}, d_m)}}{\sum_{j \in R_i} e^{-\theta C_j(\vec{t}, d_j)}}$$

- Path cost measured as a function of trip travel time and distance

$$C_m(\vec{t}, d_m) = \beta_1 \cdot g_m(\vec{t}) + \beta_2 \cdot d_m$$

$$g_m(\vec{t}) = \alpha_1 t_O + \alpha_2 t_D + \sum_{l \in L} \delta_{ml} t_l$$

Base Model

Link travel time estimation

- Minimizing the squared difference between expected ($E(Y_i|R_i)$) and observed (Y_i) path travel times

$$E(Y_i|R_i) = \sum_{m\in R_i} g_m(\vec{t})P_m(\vec{t}, d, \theta)$$

$$\vec{t} = \arg\min_{\vec{t}} \sum_{i\in D} \left(y_i - E(Y_i|R_i)\right)^2$$

- Solve using Levenberg-Marquardt (LM) method
- Parallelized codes developed to estimate the model
- Entire optimization solved within 10 minutes on an intel i7 laptop
- Numerical results show in later section

Probabilistic Model

Limitations of the base model

- Point estimate of hourly average travel time
- Not incorporating variability of link travel times
- Not utilizing historical data
- Problems of compensation effect
- Less robust

Solution: Adopt a probabilistic framework

- Accounting for variability in link travel times
- More robust
- Historical information can be incorporated as priors

PURDUE UNIVERSITY

Probabilistic Model

Assumptions:

1. Link travel time: $x_l \sim \mathcal{N}(\mu_l, \sigma_l^2)$

2. Path travel time is the summation of a set of link travel times

$$P(y_i|k,\boldsymbol{x}) = P(y_i|k,\boldsymbol{\mu},\boldsymbol{\Sigma}) = N\left(\alpha_1\mu_O + \alpha_2\mu_D + \sum_{l\in k}\mu_l\,, (\alpha_1\sigma_O)^2 + (\alpha_2\sigma_D)^2 + \sum_{l\in k}\sigma_l^2\right)$$

3. Route choice based on the perceived mean link travel times and distance

$$\pi_k^i(\boldsymbol{\mu},\boldsymbol{\beta},d_i) = \frac{\exp\left[-C_k^i(\boldsymbol{\mu},\boldsymbol{\beta},d_i)\right]}{\sum_{s\in R^i}\exp\left[-C_s^i(\boldsymbol{\mu},\boldsymbol{\beta},d_i)\right]}$$

- where $\boldsymbol{x},\boldsymbol{\mu},\boldsymbol{\Sigma}$ are the vector of link travel times, their mean and variance

PURDUE UNIVERSITY

Probabilistic Model

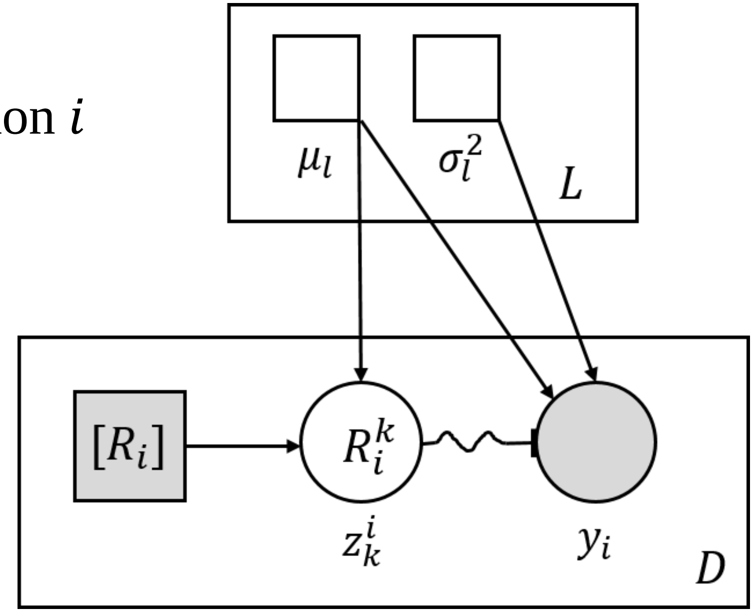
Mixture model

- A Mixture model is developed to model the posterior probability of the observed taxi trip travel times given link travel time parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$

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

- Introducing z_k^i as the latent variable indicating if path k is used by observation i

Plate notation



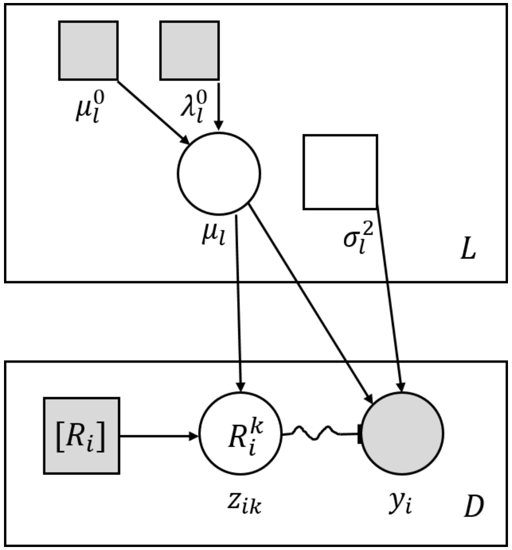
PURDUE UNIVERSITY

Probabilistic Model

Bayesian Mixture model

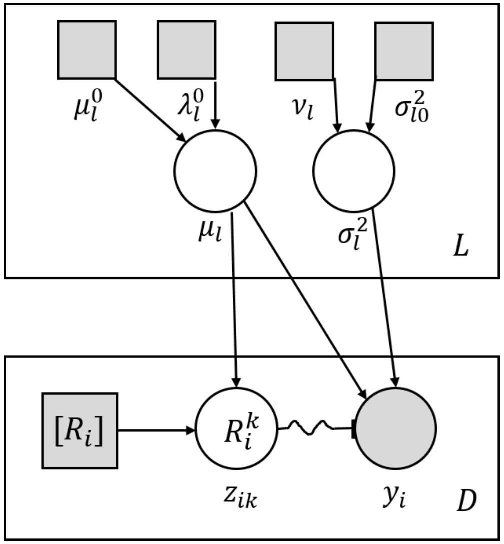
- Incorporating historical information:
  - Prior on $\boldsymbol{\mu}$:

$$H(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{D}) = \prod_{i=1}^{n}\sum_{k\in R^i}\pi_k^i(\boldsymbol{\mu},\boldsymbol{\beta},d_i)P(y_i|k,\boldsymbol{\mu},\boldsymbol{\Sigma})\cdot\prod_{j\in L}p(\mu_j)$$



  - Priors on $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$

$$H(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{D}) = \prod_{i=1}^{n}\sum_{k\in R^i}\pi_k^i(\boldsymbol{\mu},\boldsymbol{\beta},d_i)P(y_i|k,\boldsymbol{\mu},\boldsymbol{\Sigma})\cdot\prod_{j\in L}p(\mu_j)\,p(\sigma_j^2)$$



PURDUE UNIVERSITY

Probabilistic Model

Solution approach

- An EM algorithm is proposed for estimation
- A iterative procedure of two steps:
  - E-step:

$$\mathbb{E}\left(z_k^i\right) = \frac{\sum_{z_k^i} z_k^i \left[\pi_k^i(\boldsymbol{\mu}, \boldsymbol{\beta}, d_i) P(y_i | k, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right]^{z_k^i}}{\sum_{z_k^i} \sum_{s \in R^i} \left[\pi_s^i(\boldsymbol{\mu}, \boldsymbol{\beta}, d_i) P(y_i | s, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right]^{z_s^i}} = \gamma\left(z_k^i\right)$$

  - M-step: Let $\tau_l = \sigma_l^2, \boldsymbol{\tau} = \boldsymbol{\Sigma}$,

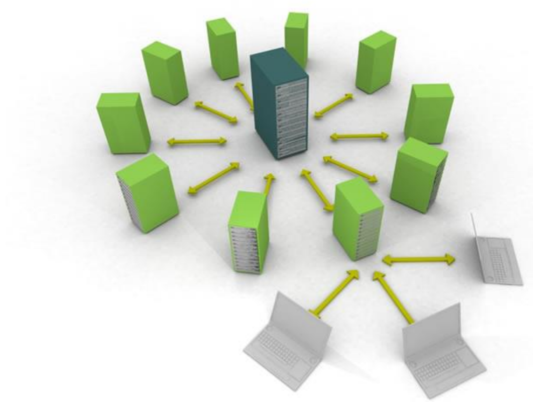
$$Q(\boldsymbol{\mu}, \boldsymbol{\tau}) = \mathbb{E}_{\boldsymbol{z}}[\ln P(\boldsymbol{y}, \boldsymbol{z} | \boldsymbol{\mu}, \boldsymbol{\tau})] = \sum_{i=1}^{n} \sum_{k \in R^i} \gamma\left(z_k^i\right)\left[\ln \pi_k^i(\boldsymbol{\mu}, \boldsymbol{\beta}, d_i) + \ln P(y_i | k, \boldsymbol{\mu}, \boldsymbol{\tau})\right]$$

$$(\boldsymbol{\mu}^{new}, \boldsymbol{\tau}^{new}) = \underset{\boldsymbol{\mu}, \boldsymbol{\tau}}{arg\,max}\, Q(\boldsymbol{\mu}, \boldsymbol{\tau})$$

Probabilistic Model

Solving for large-scale data and large networks

- The M-step involves a large-scale optimization problem
- Our goal:
  - Solve for large-scale data input
  - Solve for large network
  - Short term link travel time estimation (say 15min)
- **Solution:** parallelize the computation!
  - Alternating Direction Method of Multiplier (ADMM) to decouple the problem into smaller sub-problems
  - Solve decomposed sub-problems in parallel
  - Deals with large size of network and data
  - Faster model estimation



PURDUE UNIVERSITY

Numerical Results

Model results for base model

- Validation metrics
- Root mean square error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left({T_i}^{Pr} - {T_i}^{Ob}\right)^2}$$

- Mean absolute percentage error

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{{T_i}^{Pr} - {T_i}^{Ob}}{{T_i}^{Ob}}\right| \times 100\%$$

Numerical Results

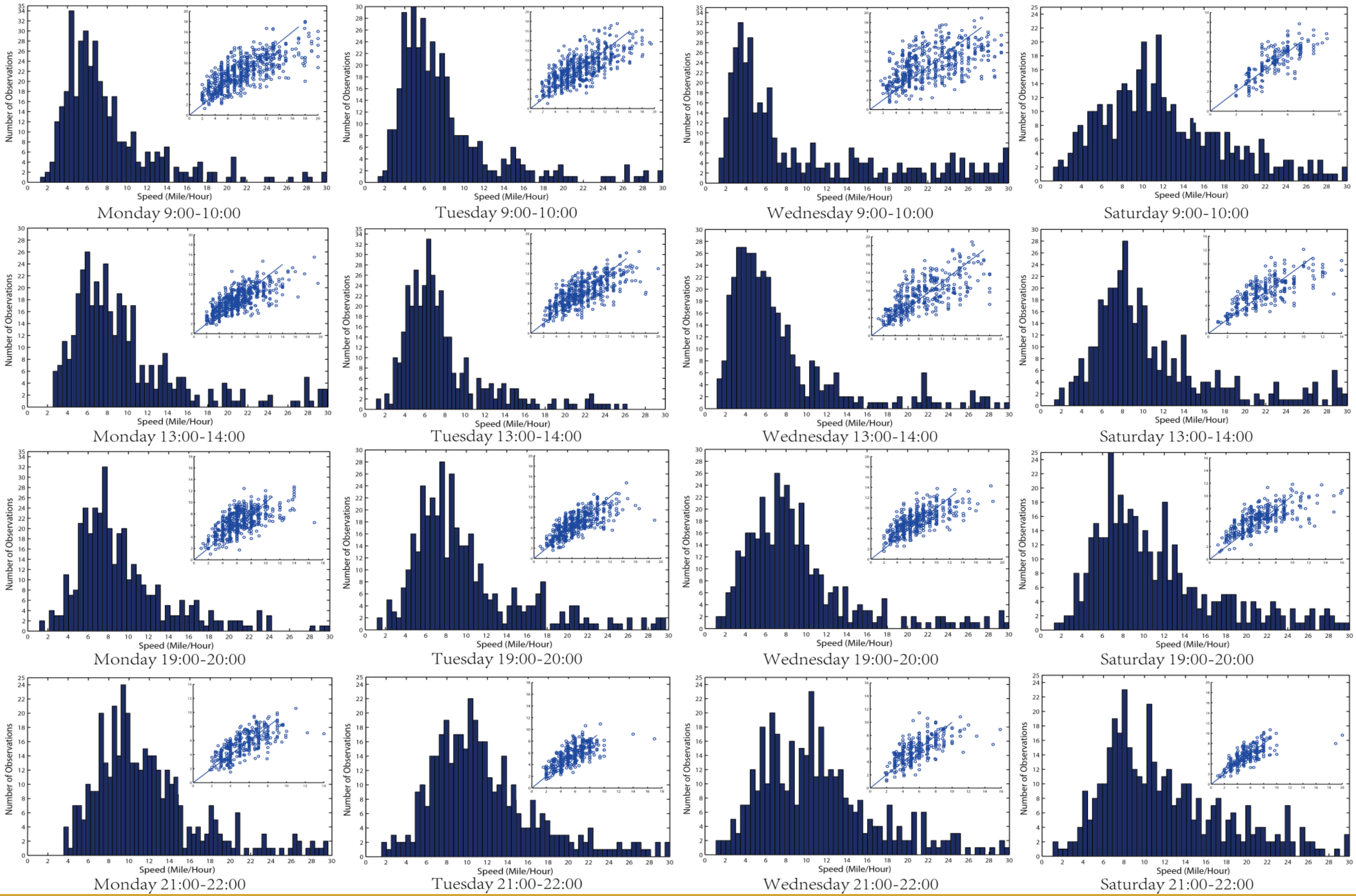
Model results for base model

- Test data: 3/15/2010 ~ 3/21/2010

Day	Error	Time Period			
		9:00-10:00	13:00-14:00	19:00-20:00	21:00-22:00
Monday	RMSE (min)	2.614	1.981	1.937	1.372
	MAPE	29.51%	24.22%	26.27%	21.87%
Tuesday	RMSE (min)	2.461	2.302	1.827	1.437
	MAPE	29.63%	25.59%	23.33%	22.20%
Wednesday	RMSE (min)	3.827*	3.216*	2.18	1.691
	MAPE	41.32%*	34.97%*	28.73%	24.40%
Thursday	RMSE (min)	2.468	2.699	2.49	1.382
	MAPE	27.28%	27.92%	28.54%	21.05%
Friday	RMSE (min)	2.26	2.179	1.692	1.334
	MAPE	27.76%	27.04%	25.17%	22.26%
Saturday	RMSE (min)	1.034	1.69	1.839	1.584
	MAPE	16.84%	24.58%	27.14%	21.61%
Sunday	RMSE (min)	2.041	1.518	1.395	1.16
	MAPE	25.44%	23.70%	22.72%	19.87%

* Traffic disturbance caused by Patrick's Day Parade.

Numerical Results



Conclusion

- Two new models are proposed to estimate urban link travel times
- Utilizing data with only partial information
- Efficiently estimation using base model with reasonable accuracy
- Mixture models are proposed to get more robust and accurate estimations
- Applicable to trajectory data, can provide more accurate estimations

Future work

- Test the mixture models for larger network
- Efficient implementation using distributed computing technique
- Result validation

Q&A

Thank you!

Questions / Comments ?

PURDUE
UNIVERSITY