A Reliable Travel Time Prediction System with Sparsely Distributed Detectors

Ph.D. Dissertation Defense

Nan Zou

May 10th, 2007
Outline

- Introduction
- Research Objectives
- Framework of the Travel Time Prediction System
- System Components
  - Travel Time Estimation Module
  - Travel Time Prediction Module
  - Missing Data Estimation Module
- Summary
- On-going Works
Introduction

- Travel times (completed and en-route trips) are crucial information for an Advanced Traveler Information System (ATIS).
Introduction (cont’d)

- Travel time prediction is a challenging task due to the impacts of
  - Geometric features
  - Traffic patterns
  - Availability of the detection system
  - Delay and/or missing of the real-time data, etc.
Issues Associated with Existing Models and Systems:

- **High system costs**
  - Densely distributed detectors (i.e., 0.5-mile apart)
  - Accurate speed detection
  - Recurrent measurement on travel times
    - Coifman et al. (2002, 2003), van Lint et al. (2003), Liu et al. (2006)

- **Reliability**
  - Missing or delayed data
  - Nonrecurrent congestions (for example, incidents)
Features of A Cost-efficient and Reliable Travel Time Prediction System

- Required input variables should be obtainable from sparsely distributed traffic detectors.
- Take advantage of some actual travel times from the field, but not rely on a large number of such data.
- Be capable of operating under normal and/or some data-missing scenarios and effectively dealing with related issues during real-time operations.
- Estimate the impact of the missing data and avoid potential large prediction errors.
Research Objectives

- Develop a travel time estimation module
  - Reliable estimates of completed trips
  - Under all types of recurrent traffic patterns
  - With sparsely distributed traffic detectors

- Construct a travel time prediction module
  - For freeway segments
  - Large detector spacing
  - Historical travel times and traffic patterns

- Integrate a missing data estimation module
  - To deal with various missing data and delay patterns
  - Estimate the impact of the missing data
T.T. Estimation vs. Prediction

Space

\[ d_D \]

\[ d_2 \]

\[ d_1 \]

Time

\[ t \]

\[ t + TT \]
T.T. Estimation

Space

$\text{d}_D$

$\text{d}_1$

$\text{d}_2$

Time

$t$

$t + TT$
T.T. Prediction

Space

\[ d_{D} \]

\[ d_{2} \]

\[ d_{1} \]

Time

\[ t \]

\[ t+TT \]
Existing Travel Time Prediction Systems

- Example systems
  - Houston, TX; Atlanta, GA; Chicago, IL; and Seattle, WA, etc.

- Almost all real-world systems use current detected traffic conditions as the prediction of the future
  - Completed trips instead of en-route trips
  - Big difference
Completed Trips vs. En-route Trips

Space

$d_D$

$t-TT1$

$t$

$t+TT2$

Time

$d_1$

$d_2$

Trip 1

Trip 2
Completed Trips vs. En-route Trips
Completed Trips vs. En-route Trips
System Flowchart

Real-Time Detector Data at Time $t$ → Database of Traffic Data

Travel Time Estimation Module

Database of Historical Travel Times

Data Missing?

N

Missing Data Estimation Module

Links with Unreliable Missing Data

Links with Reliable Missing Data Estimation Only

Predicted Travel Time for Time $t$

$\color{red}{t+1}$

Stop Predicting for Impacted Segments

Travel Time Prediction Module
System Flowchart

1. Real-Time Detector Data at Time $t$
2. Database of Traffic Data
3. Travel Time Estimation Module
   - Database of Historical Travel Times
4. Data Missing?
   - Yes (Y)
     - Missing Data Estimation Module
       - Links with Unreliable Missing Data
         - Stop Predicting for Impacted Segments
       - Links with Reliable Missing Data
         - Estimation Only
   - No (N)
     - Travel Time Prediction Module
       - Predicted Travel Time for Time $t$
5. $t = t + 1$
Literature Review

Travel Time Estimation
- Flow-based models
- Vehicle identification approaches
- Trajectory-based models
Limitations of Flow-based Models

- Reliability of detector data
  - Detection errors (volume drifting) vary over time and space

- Traffic patterns
  - Require uniformly distributed traffic across all lanes

- Geometric features
  - Cannot model ramp impact
Limitations of Vehicle Identification Approach

- Traffic patterns
  - Lane-based approach, therefore requires low lane changing rate
  - Requires uniform traffic conditions across lanes
- Geometric features
  - May not fit geometric changes, such as lane drop and lane addition
- System cost
  - High. Require new hardware or high bandwidth
- Reliability
  - Low detection resolution under high speed
  - Reduced accuracy under low light (video-based)
Limitations of Existing Trajectory-based Models

- Requires reliable speed measurement
  - Not available from most traffic detectors
- Assumes constant traffic-propagation speed
- May not perform well on long links
  - currently all studies are based on detectors less than 0.5-mile apart
A Hybrid Travel Time Estimation Model with Sparsely Distributed Detectors

- A Clustered Linear Regression Model as the main model
  - For traffic scenarios that have sufficient field observations
- An Enhanced Trajectory-based Model as the supplemental model
  - For other scenarios
Clustered Linear Regression Model

- Travel times may be constrained in a range under one identified traffic scenario
  - For example, the travel time cannot be free-flow travel time when congestion is being observed at one detector
- Assume a linear relation between the travel time under one traffic scenario with traffic variables from pre-determined critical lanes
Critical Lanes

- Those lanes that directly contribute to estimate the average travel speed of through traffic
- May includes both mainline lanes and ramp lanes
- From both upstream and downstream detector locations
Model Formulation of the Clustered Linear Regression Model

\[
\tau_d(t) = \sum_{la \in \text{CLT}_{d,d+1}^T} b_{d,la}^{T,p} \frac{o_{d,la}(t, \gamma_p^d \tau_d^E(p))}{v_{d,la}(t, \gamma_p^d \tau_d^E(p))} + \sum_{la \in \text{CLR}_{d,d+1}^R} b_{d,la}^{R,p} \frac{o_{d,la}(t, \gamma_p^d \tau_d^E(p))}{v_{d,la}(t, \gamma_p^d \tau_d^E(p))}
\]

\[
+ \sum_{la \in \text{CLT}_{d,d+1}^{d+1}} b_{d+1,la}^{T,p} \frac{o_{d+1,la}(t + \gamma_p^d \tau_d^E(p), (1 - \gamma_p^d) \tau_d^E(p))}{v_{d,la}(t + \gamma_p^d \tau_d^E(p), (1 - \gamma_p^d) \tau_d^E(p))}
\]

\[
+ \sum_{la \in \text{CLR}_{d,d+1}^{d+1}} b_{d+1,la}^{R,p} \frac{o_{d+1,la}(t + \gamma_p^d \tau_d^E(p), (1 - \gamma_p^d) \tau_d^E(p))}{v_{d,la}(t + \gamma_p^d \tau_d^E(p), (1 - \gamma_p^d) \tau_d^E(p))}
\]

\[+ b_{d}^{0,p}\]
An Enhanced Trajectory-based Model

- Combines two types of trajectory estimation:
  - Traffic propagation relations when the vehicle is close to one detector
  - An enhanced piecewise linear-speed-based model when vehicle is far from both detector

- Does not require speed in input variables
  - Estimate the occupancy first, then use occupancy-flow-speed relation to estimate the vehicle’s speed
Trajectory-based Method

- Estimated
- Actual

Space

Time

d_2

t

t + \Delta T_1

d_1
An Enhanced Trajectory-based Method
An Enhanced Trajectory-based Method

![Diagram of trajectory-based method]
Model Formulation

\[
O(x, t) = \begin{cases} 
    o_d \left( t + \frac{x - x_d}{u_c^{\max}}, t + \frac{x - x_d}{u_c^{\min}} \right) & \text{if } x - x_d < \hat{x} \\
    o_{d+1} \left( t - \frac{x_{d+1} - x}{u_c^{\min}}, t - \frac{x_{d+1} - x}{u_c^{\max}} \right) & \text{if } x_{d+1} - x < \hat{x} \\
    o_d \left( t + \frac{\hat{x} - x_d}{u_c^{\max}}, t + \frac{\hat{x} - x_d}{u_c^{\min}} \right) & \text{otherwise} \\
    + \frac{(x - x_d - \hat{x})}{\hat{x}} \\
    \times (o_{d+1} \left( t - \frac{x - (x_{d+1} - \hat{x})}{u_c^{\min}}, t - \frac{x - (x_{d+1} - \hat{x})}{u_c^{\max}} \right)) \\
    - o_d \left( t + \frac{\hat{x}}{u_c^{\max}}, t + \frac{\hat{x}}{u_c^{\min}} \right) 
\end{cases}
\]

\[
\hat{x} = \begin{cases} 
    \min \left( \frac{l_d}{3}, \frac{1}{3} \text{mi} \right) & \text{when } l_d \geq 1 \text{ mile} \\
    \frac{l_d}{3} & \text{otherwise} 
\end{cases}
\]

\[
x_d \leq x \leq x_{d+1}
\]

\[
u_c^{\min} \text{ and } u_c^{\max} \text{ are the minimum and the maximum traffic propagation speeds.}
\]
Model Formation (cont’d)

\[ u(x, t) = \begin{cases} 
  u_{\text{free}} & , o(x, t) \leq o_{\text{free}} \\
  u_{\text{cong}} + (u_{\text{free}} - u_{\text{cong}})(1 - \frac{o(x, t) - o_{\text{free}}}{o_{\text{cong}} - o_{\text{free}}})^m & , o_{\text{free}} < o(x, t) \leq o_{\text{cong}} \\
  u_{\text{min}} + (u_{\text{cong}} - u_{\text{min}})(1 - \frac{o(x, t) - o_{\text{cong}}}{o_{\text{max}} - o_{\text{cong}}})^n & , o_{\text{cong}} < o(x, t) \leq o_{\text{max}} \\
  u_{\text{min}} & , \text{otherwise} 
\end{cases} \]
Numeric Examples

- I-70 eastbound from MD27 to I-695
- 10 detectors on a 25-mile stretch
- Flow count and occupancy data
- 30-second intervals
Methods for Comparison

- Proposed hybrid model
  - Clustered Linear Regression (CLR) model
  - Enhanced Trajectory-based (ETB) model
- Flow-based method (Nam and Drew, 1996)
- Piecewise Linear Speed-based (PLSB) method (Van Lint and van der Zijpp, 2003)
## Volume Drifting Issue

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Daily Volume at</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detector 4</td>
<td>37040</td>
<td>39121</td>
<td>41595</td>
<td>42707</td>
<td>35190</td>
<td>29891</td>
</tr>
<tr>
<td><strong>Daily Volume at</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detector 6</td>
<td>37903</td>
<td>39695</td>
<td>42373</td>
<td>43410</td>
<td>35117</td>
<td>29741</td>
</tr>
<tr>
<td><strong>Difference</strong></td>
<td>863</td>
<td>574</td>
<td>778</td>
<td>703</td>
<td>-73</td>
<td>-150</td>
</tr>
<tr>
<td><strong>Relative</strong></td>
<td>2.33%</td>
<td>1.47%</td>
<td>1.87%</td>
<td>1.65%</td>
<td>-0.21%</td>
<td>-0.50%</td>
</tr>
<tr>
<td><strong>Daily Volume at</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detector 8</td>
<td>45332</td>
<td>49022</td>
<td>50160</td>
<td>50670</td>
<td>39469</td>
<td>34806</td>
</tr>
<tr>
<td><strong>Daily Volume at</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detector 9</td>
<td>44979</td>
<td>48945</td>
<td>49796</td>
<td>50449</td>
<td>39314</td>
<td>34784</td>
</tr>
<tr>
<td><strong>Difference</strong></td>
<td>-353</td>
<td>-77</td>
<td>-364</td>
<td>-221</td>
<td>-155</td>
<td>-22</td>
</tr>
<tr>
<td><strong>Relative</strong></td>
<td>-0.78%</td>
<td>-0.16%</td>
<td>-0.73%</td>
<td>-0.44%</td>
<td>-0.39%</td>
<td>-0.06%</td>
</tr>
</tbody>
</table>
Flow-based Model

(481 actual travel time samples, January 19th, 2007)
## Travel Time Surveys

<table>
<thead>
<tr>
<th>Date and Time</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-2</td>
</tr>
<tr>
<td>12/1/2005 AM</td>
<td>Y</td>
</tr>
<tr>
<td>1/19/2006 AM</td>
<td></td>
</tr>
<tr>
<td>1/20/2006 AM</td>
<td></td>
</tr>
<tr>
<td>1/20/2006 PM</td>
<td></td>
</tr>
<tr>
<td>2/1/2006 AM</td>
<td></td>
</tr>
<tr>
<td>2/2/2006 AM</td>
<td></td>
</tr>
<tr>
<td>2/7/2006 PM</td>
<td></td>
</tr>
<tr>
<td>2/28/2006 AM</td>
<td>Y</td>
</tr>
<tr>
<td>3/1/2006 PM</td>
<td></td>
</tr>
<tr>
<td>3/7/2006 AM</td>
<td></td>
</tr>
<tr>
<td>3/9/2006 PM</td>
<td></td>
</tr>
<tr>
<td>4/6/2006 AM</td>
<td></td>
</tr>
<tr>
<td>4/20/2006 AM</td>
<td></td>
</tr>
<tr>
<td>6/13/2006 AM</td>
<td>Y</td>
</tr>
<tr>
<td>6/15/2006 PM</td>
<td>Y</td>
</tr>
</tbody>
</table>
Performance on Individual Links - Link (5,6)

- 446 samples on January 19\textsuperscript{th}, 2006
- 411 samples on February 28\textsuperscript{th}, 2006
- 4 identified scenarios

<table>
<thead>
<tr>
<th>ID</th>
<th>Description of the Scenario</th>
<th>Detector 5 Occ. in Ln. 1</th>
<th>Detector 5 Occ. in Ln. 2</th>
<th>Detector 6 Occ. in Ln. 1</th>
<th>Detector 6 Occ. in Ln. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No congestion on the link</td>
<td>≤12</td>
<td>≤10</td>
<td>≤10</td>
<td>≤10</td>
</tr>
<tr>
<td>2</td>
<td>Congestion at Detector 5; no congestion at Detector 6</td>
<td>&gt;12</td>
<td>&gt;10</td>
<td>≤10</td>
<td>≤10</td>
</tr>
<tr>
<td>3</td>
<td>Congestion at both Detectors 5 and 6</td>
<td>&gt;12</td>
<td>&gt;10</td>
<td>&gt;10</td>
<td>&gt;10</td>
</tr>
<tr>
<td>4</td>
<td>Other</td>
<td></td>
<td></td>
<td>Other combinations</td>
<td></td>
</tr>
</tbody>
</table>
### Performance on Individual Links - Link (5,6) (cont’d)

<table>
<thead>
<tr>
<th>Scenario 2</th>
<th>All Samples (35 Observations)</th>
<th>Travel Times ≤ 95 sec. (20 Observations)</th>
<th>Travel Times &gt; 95 sec. (15 Observations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AAE (Sec.)</td>
<td>AARE (%)</td>
<td>AAE (Sec.)</td>
</tr>
<tr>
<td>CLR</td>
<td>5.63</td>
<td>6.57</td>
<td>6.16</td>
</tr>
<tr>
<td>ETB</td>
<td>5.14</td>
<td>5.43</td>
<td>3.71</td>
</tr>
<tr>
<td>PLSB</td>
<td>6.17</td>
<td>6.49</td>
<td>5.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Scn. 3</th>
<th>All Samples (33 Observation)</th>
<th>Scenario 1 (60 Observations)</th>
<th>Scenario 4 (151 Observations)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AAE (Sec.)</td>
<td>AARE (%)</td>
<td>AAE (Sec.)</td>
</tr>
<tr>
<td>CLR</td>
<td>6.60</td>
<td>4.79</td>
<td></td>
</tr>
<tr>
<td>ETB</td>
<td>19.48</td>
<td>13.35</td>
<td>2.67</td>
</tr>
<tr>
<td>PLSB</td>
<td>26.33</td>
<td>17.65</td>
<td>2.92</td>
</tr>
</tbody>
</table>
Performance on Multiple Links

- Subsegment (3, 10)
- About 10 miles
- 71 samples on April 6\textsuperscript{th}, 2006
- 114 samples on April 20\textsuperscript{th}, 2006
## Performance on Multiple Links (cont’d)

<table>
<thead>
<tr>
<th></th>
<th>Travel Time Range (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>520 to 800</td>
</tr>
<tr>
<td><strong>Sample Size</strong></td>
<td>23</td>
</tr>
<tr>
<td><strong>Maximum Travel Time (sec)</strong></td>
<td>796</td>
</tr>
<tr>
<td><strong>Average Travel Time (sec)</strong></td>
<td>742.3</td>
</tr>
<tr>
<td><strong>Hybrid Model</strong></td>
<td></td>
</tr>
<tr>
<td>AAE (sec)</td>
<td>59.4</td>
</tr>
<tr>
<td>AARE (%)</td>
<td>8.1%</td>
</tr>
<tr>
<td><strong>PLSB</strong></td>
<td></td>
</tr>
<tr>
<td>AAE (sec)</td>
<td>139.8</td>
</tr>
<tr>
<td>AARE (%)</td>
<td>18.4%</td>
</tr>
</tbody>
</table>
Performance on Multiple Links (cont’d)

![Graph showing travel times over time]

April 6th, 2006
System Flowchart

Database of Traffic Data

Real-Time Detector
Data at Time $t$

Travel Time Estimation Module

Database of Historical Travel Times

Data Missing?

Y

Links with Reliable Missing Data

Links with Unreliable Missing Data

Stop Predicting for Impacted Segments

N

Predicted Travel Time for Time $t$

Travel Time Prediction Module

Missing Data Estimation Module

System Flowchart
Travel Time Prediction

- **Parametric Models**
  - Time series model
  - Linear regression model
  - Kalman Filter model

- **Nonparametric models**
  - Neural Network model
  - Nearest Neighbor model
  - Kernel model and local regression model
Autoregressive Integrated Moving Average (ARIMA)

- Advantages:
  - Ability to predict a time series data set
  - Good for predicting traffic data (volume, speed, or occupancy) at one detector

- Disadvantages:
  - Focus on the mean value, therefore cannot well predict scenarios that less frequently occur
  - It is hard to model multiple sets of time series data together (for example, multiple series of data from detectors)
Linear Regression Models

- One single linear regression model cannot predict well for all traffic scenarios, therefore multi-model structure is often used:
  - Layered/clustered linear regression model
  - Varying coefficient linear regression model
Kalman Filter Model

- Ability to auto-update parameters based on the evaluation of the prediction accuracy of the previous time interval.

- Good performance when the true value can be obtained with a short delay (Chien et al., 2002 and 2003).

- May not work well for a prediction system with long travel times (long travel times = long delay for the update process).
Neural Network Models

- Widely used to predict travel times
- Accurate and robust because of its good ability to recognize patterns
- Multi-layer Perceptron (MLP) and Time Delay Neural Network (TDNN) are mostly seen in the literature
- A large amount of training data
MLP

TDNN
**k-Nearest Neighbor Model**

- Looks for $k$ most similar cases as the current condition from the historical database to come out a prediction
- Requires a fairly large historical database

\[
dist_{EUC}(p, q) = \sqrt{\sum_{i=1}^{K} (p_i - q_i)^2}
\]

\[
dist_{NUW}(p, q) = \sqrt{\sum_{i=1}^{K} w_i (p_i - q_i)^2}
\]
Other Nonparametric Models

- Share a common structure
  - A clustering function
  - A kernel function (linear, nonlinear and/or other form) for each cluster

- For example
  - Kernel regression
  - Layered linear regression
  - Time-varying coefficient linear regression
A Hybrid Travel Time Prediction Model

- A Multi-topology Neural Network model
  - A rule-based clustering function
  - Customized topologies for various traffic scenarios
- An Enhanced $k$-Nearest Neighbor Model
  - For cases with sufficient good matches in the historical data
A Multi-topology Neural Network Model

- Categorize congestion patterns, instead of time-of-day, with a rule-based clustering function
- Select only data in critical lanes as input variables
  - Geometric features
  - Traffic patterns
- Various topology to fit different traffic patterns
IF \( t \geq TML_{d}^{wk} \) and \( t \leq TMU_{d}^{wk} \) THEN

IF \( \exists l_{a}, o_{d,l_{a}}(t-j) > OM_{d,l_{a}} \) for all \( l_{a} \in CLM_{d,d+1}^{d} \) and 
\( 0 \leq j \leq THN_{d} \), THEN
\[ p_{d}(t) = 1 \] (morning congestion)
ELSE
IF \( o_{d,l_{a}}(t-j) \leq OM_{d,l_{a}} \) for all \( l_{a} \) and \( j \), where \( l_{a} \in CLM_{d,d+1}^{d} \) and 
\( 0 \leq j \leq THN_{d} \), THEN
\[ p_{d}(t) = 0 \] (off-peak period)
ELSE
\[ p_{d}(t) = p_{d}(t-1) \]
END IF
END IF
ELSE
IF \( t \geq TEL_{d}^{wk} \) and \( t \leq TEU_{d}^{wk} \) THEN

IF \( \exists l_{a}, o_{d,l_{a}}(t-j) > OE_{d,l_{a}} \) for all \( l_{a} \in CLE_{d,d+1}^{d} \) and 
\( 0 \leq j \leq THN_{d} \), THEN
\[ p_{d}(t) = -1 \] (evening congestion)
ELSE
IF \( o_{d,l_{a}}(t-j) \leq OE_{d,l_{a}} \) for all \( l_{a} \) and \( j \), where
\( 0 \leq j \leq THN_{d} \), THEN
\[ p_{d}(t) = 0 \] (off-peak period)
ELSE
\[ p_{d}(t) = p_{d}(t-1) \]
END IF
END IF
ELSE
\[ p_{d}(t) = 0 \] (off-peak period)
END IF

where, \( TML_{d}^{wk} \) and \( TMU_{d}^{wk} \) are the lower and upper time boundaries for morning peak hours in link \((d,d+1)\) on weekday \(wk\) in the historical traffic patterns; \( TEL_{d}^{wk} \) and \( TEU_{d}^{wk} \) are the lower and upper time boundaries for evening peak hours in link \((d,d+1)\) on weekday \(wk\) in the historical traffic patterns; \( 0:00 \leq TML_{d}^{wk} < TMU_{d}^{wk} \leq TEL_{d}^{wk} < TEU_{d}^{wk} < 24:00 \); 
\( d^{*} = d \) or \( d+1 \);
\( OM_{d,l_{a}} \) is the occupancy threshold at lane \( l_{a} \) at detector \( d \) in the morning;
\( OE_{d,l_{a}} \) is the occupancy threshold at lane \( l_{a} \) at detector \( d \) in the evening;
\( CLM_{d,d+1}^{d} \) and \( CLE_{d,d+1}^{d} \) are sets of critical lanes at detector \( d^{*} \) in link \((d,d+1)\) in the morning and in the evening respectively; and
\( THN_{d} \) is the required duration for the traffic condition to maintain congested or uncongested stably;
Enhanced Topology

- Combines time-series and non-time-series data
**k-Nearest Neighbor Model for Travel Time Prediction**

- An updated distance function
  - Based on three types of traffic state

- Geometric features
  - Take traffic data from critical lanes only
  - The time range of input data increases with the distance to the origin

- Daily and weekly traffic patterns
  - Varying search window based on historical traffic patterns
Modified Definition of the Distance

\[ mdis = \sqrt{\sum_{i=1}^{k} w_i (p_i^* - q_i^*)^2} \]

\[ p_i^* = \begin{cases} 
  p_i & \text{when } TC_d^{l_a} (t, t + \Delta t) = 0 \\
  OC_d^{l_a} & \text{when } TC_d^{l_a} (t, t + \Delta t) = 1 \\
  OF_d^{l_a} & \text{when } TC_d^{l_a} (t, t + \Delta t) = -1 
\end{cases} \]

\[ q_i^* = \begin{cases} 
  q_i & \text{when } TC_d^{l_a} (t_h, t_h + \Delta t) = 0 \\
  OC_d^{l_a} & \text{when } TC_d^{l_a} (t_h, t_h + \Delta t) = 1 \\
  OF_d^{l_a} & \text{when } TC_d^{l_a} (t_h, t_h + \Delta t) = -1 
\end{cases} \]
Consideration of Traffic Patterns

\[ mdis = \sqrt{\sum_{i=1}^{k} w_i (\hat{p}_i - q_i^*)^2} \]

Where

\[ \hat{p}_i = \begin{cases} M, & \text{if } |t - t_h| > T_{th}(d, t) \\ p_i^* \times \hat{w}, & \text{otherwise} \end{cases} \]

\[ \hat{w} = \begin{cases} 1, & \text{if } \exists s, wk_h \in W_s \text{ and } wk_c \in W_s \ (1 \leq s \leq S) \\ M, & \text{otherwise} \end{cases} \]

\[ \bigcup_{s=1}^{S} W_s = \{ \text{all weekdays} \} \]

\( M \) is a very large number.
\( wk_c \) and \( wk_h \) are weekdays of the current case and the historical case respectively.
Numerical Examples

- Same dataset from I-70 eastbound
- Subsegment (3, 10)
  - About 10 mile
- Comparison 1:
  - Predicted travel times vs. estimated travel times
- Comparison 2:
  - Predicted travel times vs. actual travel times
# Models for Comparison

<table>
<thead>
<tr>
<th>Model Description</th>
<th>4 Weeks of Training Data</th>
<th>10 Weeks of Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hybrid model developed in this study</td>
<td>HM4</td>
<td>HM10</td>
</tr>
<tr>
<td>Neural Network model in the developed hybrid model</td>
<td>NN4</td>
<td>NN10</td>
</tr>
<tr>
<td>k-Nearest Neighbors model in the developed hybrid model</td>
<td>kNN4</td>
<td>kNN10</td>
</tr>
<tr>
<td>Constant current speed-based model</td>
<td></td>
<td>CCSB</td>
</tr>
<tr>
<td>Time-varying coefficient model</td>
<td>TVC4</td>
<td>TVC10</td>
</tr>
</tbody>
</table>
Predicted vs. Estimated

- 6:00 to 10:30 and 15:00 to 19:30
- AM: May 16th, 2006 to May 19th, 2006
- PM: May 16th, 2006 and May 17th, 2006
## All Sample Days

<table>
<thead>
<tr>
<th>Model</th>
<th>Average Absolute Error (second)</th>
<th>Average Absolute Relative Error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCSB</td>
<td>77.92</td>
<td>10.89</td>
</tr>
<tr>
<td>TVC4</td>
<td>173.99</td>
<td>28.10</td>
</tr>
<tr>
<td>TVC10</td>
<td>65.64</td>
<td>9.44</td>
</tr>
<tr>
<td>kNN4</td>
<td>64.38</td>
<td>9.04</td>
</tr>
<tr>
<td>kNN10</td>
<td>60.86</td>
<td>8.56</td>
</tr>
<tr>
<td>NN4</td>
<td>53.88</td>
<td>7.81</td>
</tr>
<tr>
<td>NN10</td>
<td>48.68</td>
<td>7.07</td>
</tr>
<tr>
<td>HM4</td>
<td>48.84</td>
<td>6.92</td>
</tr>
<tr>
<td>HM10</td>
<td>45.69</td>
<td>6.53</td>
</tr>
</tbody>
</table>
# Each Peak Period

<table>
<thead>
<tr>
<th>Average Absolute Error (seconds)</th>
<th>5/16 AM</th>
<th>5/16 PM</th>
<th>5/17 AM</th>
<th>5/17 PM</th>
<th>5/18 AM</th>
<th>5/19 AM</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCSB</td>
<td>56.37</td>
<td>73.93</td>
<td>106.62</td>
<td>106.84</td>
<td>63.95</td>
<td>71.97</td>
</tr>
<tr>
<td>TVC4</td>
<td>186.04</td>
<td>127.27</td>
<td>232.82</td>
<td>128.47</td>
<td>166.45</td>
<td>168.96</td>
</tr>
<tr>
<td>TVC10</td>
<td>39.64</td>
<td>105.05</td>
<td>83.84</td>
<td>121.86</td>
<td>41.38</td>
<td>36.99</td>
</tr>
<tr>
<td>kNN4</td>
<td>34.42</td>
<td>84.09</td>
<td>81.48</td>
<td>126.45</td>
<td>34.10</td>
<td>58.85</td>
</tr>
<tr>
<td>kNN10</td>
<td>31.71</td>
<td>71.08</td>
<td>79.46</td>
<td>127.66</td>
<td>33.68</td>
<td>54.25</td>
</tr>
<tr>
<td>NN4</td>
<td>31.81</td>
<td>68.18</td>
<td>64.77</td>
<td>93.47</td>
<td>32.80</td>
<td>53.39</td>
</tr>
<tr>
<td>NN10</td>
<td>30.70</td>
<td>65.38</td>
<td>55.64</td>
<td>75.96</td>
<td>36.83</td>
<td>43.92</td>
</tr>
<tr>
<td>HM4</td>
<td>29.44</td>
<td>54.00</td>
<td>58.37</td>
<td>87.17</td>
<td>28.95</td>
<td>49.82</td>
</tr>
<tr>
<td>HM10</td>
<td>29.09</td>
<td>52.10</td>
<td>53.75</td>
<td>75.96</td>
<td>35.26</td>
<td>36.69</td>
</tr>
</tbody>
</table>
### Predicted vs. Actual

- 70 actual travel times collected by a third party company
- Same sample peak periods

<table>
<thead>
<tr>
<th></th>
<th>Average Travel Time (seconds)</th>
<th>HM4 AAE (seconds)</th>
<th>HM10 AAE (seconds)</th>
<th>Number of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>All samples</td>
<td>655.67</td>
<td>56.58</td>
<td>51.69</td>
<td>70</td>
</tr>
<tr>
<td>TT≤580</td>
<td>532.58</td>
<td>15.74</td>
<td>15.11</td>
<td>24</td>
</tr>
<tr>
<td>580&lt;TT≤900</td>
<td>703.86</td>
<td>80.45</td>
<td>72.02</td>
<td>36</td>
</tr>
<tr>
<td>TT&gt;900</td>
<td>949.67</td>
<td>113.43</td>
<td>95.29</td>
<td>10</td>
</tr>
</tbody>
</table>
System Flowchart

Real-Time Detector Data at Time $t$ → Travel Time Estimation Module → Database of Traffic Data

- Database of Historical Travel Times
- Stop Predicting for Impacted Segments

Data Missing?

- Missing Data Estimation Module
  - Links with Unreliable Missing Data
    - Links with Reliable Missing Data Estimation Only
  - Predicted Travel Time for Time $t$ → Travel Time Prediction Module

$t = t + 1$
Missing Data Estimation

- Data discard
- Single imputation
- Multiple imputation
Multiple Imputation Technique

- Estimate the distribution of the missing values
- Randomly draw missing values until the distributions converge
- Repeat the imputation for $m$ times
Proposed Models

- Model M-1:
  - An integrated missing data imputation and travel time prediction model
  - Rely on data of the entire target segment

- Model M-2:
  - Multiple imputation model for missing values
  - Rely on data from predefined subsegment
Numerical Examples

- Same dataset on I-70
- Four weekdays
  - June 20th (Tuesday)
  - 21st (Wednesday)
  - 22nd (Thursday)
  - 26th, 2006 (Monday)

- Comparison focuses on the impacts of:
  - The missing rate
  - The imputation models
    - Mean substitution (MS), Bayesian Forecast (BS)
  - The number of imputation
Performance Comparison (Travel Time)

- One most critical detector has missing data
Different Congestion Levels

<table>
<thead>
<tr>
<th>TT ≤ 700</th>
<th>MS</th>
<th>BF</th>
<th>M-2-50</th>
<th>M-1-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>3.10%</td>
<td>2.78%</td>
<td>3.11%</td>
<td>2.54%</td>
</tr>
<tr>
<td>40%</td>
<td>4.10%</td>
<td>3.63%</td>
<td>3.26%</td>
<td>2.80%</td>
</tr>
<tr>
<td>60%</td>
<td>5.05%</td>
<td>4.47%</td>
<td>3.73%</td>
<td>3.07%</td>
</tr>
<tr>
<td>100%</td>
<td>8.53%</td>
<td>7.47%</td>
<td>5.42%</td>
<td>6.37%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>700 &lt; TT ≤ 900</th>
<th>MS</th>
<th>BF</th>
<th>M-2-50</th>
<th>M-1-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>8.23%</td>
<td>7.35%</td>
<td>7.43%</td>
<td>6.65%</td>
</tr>
<tr>
<td>40%</td>
<td>8.76%</td>
<td>8.56%</td>
<td>7.29%</td>
<td>6.43%</td>
</tr>
<tr>
<td>60%</td>
<td>9.33%</td>
<td>8.64%</td>
<td>7.71%</td>
<td>6.95%</td>
</tr>
<tr>
<td>100%</td>
<td>10.48%</td>
<td>10.26%</td>
<td>8.58%</td>
<td>8.66%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TT &gt; 900</th>
<th>MS</th>
<th>BF</th>
<th>M-2-50</th>
<th>M-1-50</th>
</tr>
</thead>
<tbody>
<tr>
<td>20%</td>
<td>13.46%</td>
<td>12.80%</td>
<td>12.36%</td>
<td>10.96%</td>
</tr>
<tr>
<td>40%</td>
<td>13.82%</td>
<td>15.05%</td>
<td>12.57%</td>
<td>11.86%</td>
</tr>
<tr>
<td>60%</td>
<td>14.29%</td>
<td>15.28%</td>
<td>12.99%</td>
<td>12.76%</td>
</tr>
<tr>
<td>100%</td>
<td>16.12%</td>
<td>15.86%</td>
<td>13.55%</td>
<td>14.07%</td>
</tr>
</tbody>
</table>
June 20th, 2006
$m=5, 10, 20, 50$

**Missing rate: 20%**

**Missing rate: 40%**

**Missing rate: 60%**

**Missing rate: 100%**
Summary

Contributions

- Perform an in-depth review of literature associated with travel time prediction
- Develop a modeling framework for a travel time prediction system with widely spaced detectors on the freeway
  - Propose a hybrid model for estimating travel times
  - Develop a hybrid model for travel time prediction
  - Construct an integrated missing data estimation model for contending missing data issue
Future Research

- Determining *optimal detector locations* for better prediction performance
- **Detecting incidents** and other special events to minimize the potential prediction errors
- Monitoring *change in traffic patterns* and estimating potential impacts
Thank you!

Any questions?