Machine Learning Approach for Dynamic Bus Arrival Time Prediction

Sean X. He
Department of Civil and Environmental Engineering
Rensselaer Polytechnic Institute

2023 ITS-NY 30th Annual Meeting
June 15th, 2023
Accurate bus arrival time predictions

- Enhanced passenger experience
- Time management and productivity
- Improved accessibility
- Increased ridership
- Efficient resource allocation
- Enhancing traffic management and urban planning
- Intelligent decision making
Bus Service Performance Metrics

Source: NEW YORK CITY TRANSIT & BUS KEY PERFORMANCE METRICS, MTA (July 2022)
Bus On-Time Arrival Statistics

- CDTA Data (09/2022 – 03/2023)
  - 630K stop arrival data
  - Workdays
  - Route 1 – Central Ave

![Bus arrival delay histogram](image)
Bus Arrival Time Information

- **Travelers’ convenience**
  - plan their trips and minimize waiting times

- **Operational efficiency**
  - optimize schedules, reduce operational costs

- **Equity**
  - disproportionately affect low-income and marginalized communities

- **Transit ridership**
  - environmental benefits
### Challenges in Accurate Bus Arrival Time Prediction

#### Bus Operations
- Route specific factors
- Passenger demand
- Accommodation for special needs
- Dwell time

#### Environment
- Traffic variability
- Weather conditions
- Dynamic events
- Traffic management decisions

- Influencing factors are uncertain and highly dynamic
- Accurate predictions rely on massive high-quality and reliable data
- Requires sophisticated algorithms and techniques
**Markov property**: In the evolution of a Markov process, the current state depends only on the previous state and does not depend on the past.

<table>
<thead>
<tr>
<th>Markov Process</th>
<th>Bus Arrival Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Current state</td>
<td>Arrival time at current stop</td>
</tr>
<tr>
<td>Previous state</td>
<td>Arrival time at previous stop</td>
</tr>
<tr>
<td>Uncertainties in the</td>
<td>Uncertainties in:</td>
</tr>
<tr>
<td>environment</td>
<td>a. Traffic condition;</td>
</tr>
<tr>
<td></td>
<td>b. Bus travel demand(dwelling time)</td>
</tr>
</tbody>
</table>

Pattern to be learned by ML.

Learn the **pattern**!
Bus Arrivals as a Markov Process

State:

1 2 3 N

Transition matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_2^d$</th>
<th>$S_2^e$</th>
<th>$S_2^m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1^a$</td>
<td>0.1</td>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td>$S_1^b$</td>
<td>0.4</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>$S_1^c$</td>
<td>0.5</td>
<td>0.2</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Given state $S_1^a$ at node 1, the probability of $S_2^m$ happening at node 2.

Bus stop:

1 2 3 N

Given arrival time $T_1^a$ at stop 1, the probability of arriving at time $T_2^m$ at stop 2.
Illustrative Example

Matrix size and discrete time interval are flexible to be adjusted.

If the bus arrives at stop 1 at 10:09, the probability of arriving at stop 2 at 10:12 is 0.1.
Uncertainties to be Learned

Bus stop: 1 ➔ 2 ➔ 3 ➔ N

Transition matrix will be learned from historical bus arrival data

- Identifications of the trip, stop, schedule
- Arrival time at previous stop

Using historical bus arrival data

Input: Machine learning model
Output: The probability distribution of arrival times at the stop (transition matrix)
Machine Learning Approach

• Supervised machine learning problem
  • Label: the frequency of different arrival time intervals

• XGBoost is used in the case study of this research
  • Extreme Gradient Boosting
  • A scalable, distributed gradient-boosted decision tree (GBDT)

• Loss function: Mean Square Error (MSE)

<table>
<thead>
<tr>
<th>Learning input</th>
<th>Learning output</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ Stop ID</td>
<td>▪ The probabilities of arrival times at the stop</td>
</tr>
<tr>
<td>▪ Type of trip</td>
<td>▪ One row in the transition matrix represented the probability mass function</td>
</tr>
<tr>
<td>▪ Scheduled arrival time</td>
<td></td>
</tr>
<tr>
<td>▪ Actual arrival time at previous stop</td>
<td></td>
</tr>
</tbody>
</table>
Prediction Using the Learned Transition Matrices

To predict the arrival time of any stop $N$ from stop $A$, we multiply all transition matrices between stop $A$ and stop $N$. **Note:** The result is a distribution.

Bus stops: $A$ — $B$ — $C$ — $D$ — ... — $M$ — $N$

If the bus arrives at stop $A$ at time $T_A^b$:

Then, only this row will be used.
Framework

Historical bus arrival data → Machine learning model → Transition matrices → Predicted arrival times for all subsequent stops

Bus finishes one trip → Bus arrives at each stop

Feed new data into the pool

update
Directions:
East: Colonie Center to Downtown Albany
West: Downtown Albany to Colonie Center

Data:
09/2022 – 03/2023, workdays
Features: stop, direction, schedule arrival time, actual arrival time, date

Discrete time interval:
1 minute
Settings

• Stops
  • 37(long trip) / 30(short trip) for eastbound
  • 29(long trip) / 28(short trip) for westbound

• Schedules
  • 36(long trip) / 73(short trip) for eastbound
  • 36(long trip) / 75(short trip) for westbound

• Discrete time interval: 1 minute

• Transition matrix
  • Size 27*27
  • The rows and columns represent the arrival times that deviate the scheduled arrival time from -5 to 22 (actual – scheduled, minute)
  • 1296(long trip) / 2117(short trip) transition matrices for eastbound
  • 1008(long trip) / 2025(short trip) transition matrices for westbound

• 630,000 stop arrival data for Route 1: 80% for training; 20% for testing
Results on the Training Set

Percentage in the acceptable range [-1,5]: 81.7%
Results on the Testing Set

Percentage in the acceptable range [-1,5]: 80.2%
Results—Temporal Variation

- Light blue area represents 70% trust interval
- Darker blue line: median
- Large variation during 8:00 – 19:00
Results—Spatial Variation: Eastbound

- Colonie Center to Downtown Albany
- Origin stop removed
- Larger variations for the later stops on the route
- Lowest variations at:
  - Central Ave & Osborne Rd
  - Central Ave & Yardboro Ave
  - 1010 Central Ave
Results — Spatial Variation: Westbound

- Downtown Albany to Colonie Center
- Origin stop removed
- Larger variations for the later stops on the route
- Lowest variations at:
  - Central Ave & Henry Johnson Blvd
Benefits

• Minimal data requirement: Bus arrival data only
  • Easy to transfer

• Uncertainties are well addressed by a machine learning model
  • Anticipate the environment instead of simply reacting to observations in real time

• Flexible prediction information
  • In the case study, expectation of arrival time is used for prediction
  • The maximum likelihood and trust intervals can also be provided

• Flexible modeling of transition matrices as per operational needs
  • Time intervals
  • Could be simplified as transition vectors

• High accuracy
Acknowledgements

Collaborators:

Xiaoyu Ma, Jack Reilly, Calvin Young, Rich Fantozzi

Sponsors:
Thank you!
hex6@rpi.edu