Data-Driven Mesoscopic Simulation of Large-Scale Surface Transit Networks

MASc Candidate: Bo Wen Wen
Supervisor: Prof. Amer Shalaby

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

Presentation Outline

- Introduction
- Modelling Framework
- Data
- Model Estimation
- Data-driven Simulation
- Case Study

Introduction

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

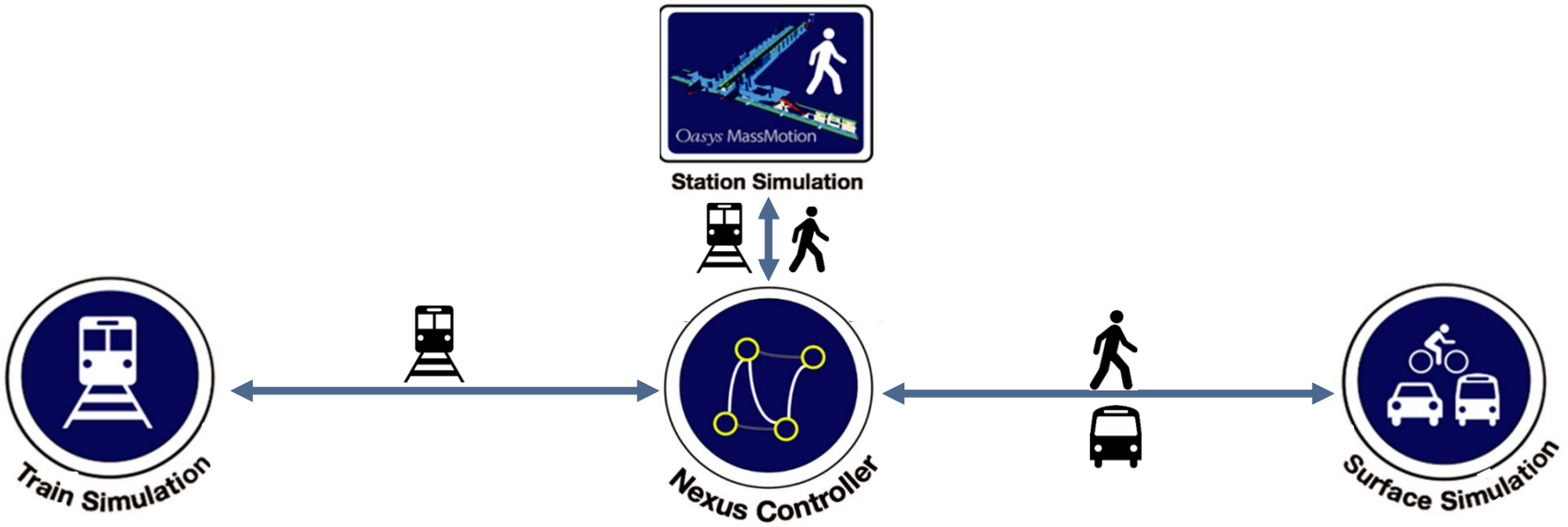
The Nexus Platform [1]

- **Simulation platform**: currently in development, motivated by the need for a high-fidelity multimodal transit network modelling system with capability to:
  - Represents the dynamic behaviour of transit lines and stations
  - Predicts passenger travel behaviour under normal and irregular conditions
- **Scenario analyses**: disruptions, response strategies, and long range planning

1. Srikukenthiran, 2015

The Nexus Platform

- Connects specialized simulators of train operation, pedestrian simulation and surface vehicle movement into a network allowing for modular, multi-modal simulation



Research Motivation

- An accurate and efficient surface transit simulator to connect with Nexus
- Simulation of large-scale transit networks where detailed microsimulation is not needed
- Rapid construction of the transit simulation model with little manual effort

Research Motivation

- Traditional models:
  - Difficult to calibrate and computationally intensive
  - Updated infrequently (out of date)

- Open transit data (AVL, APC, GTFS, AFC, etc.) provides:
  - The potential to capture real world stochasticity
  - Rapidly build models using appropriate methodological tools for big data

- Instead of modelling the kinematics of vehicles, transit vehicle arrivals can be modelled using historical data.

Research Objectives

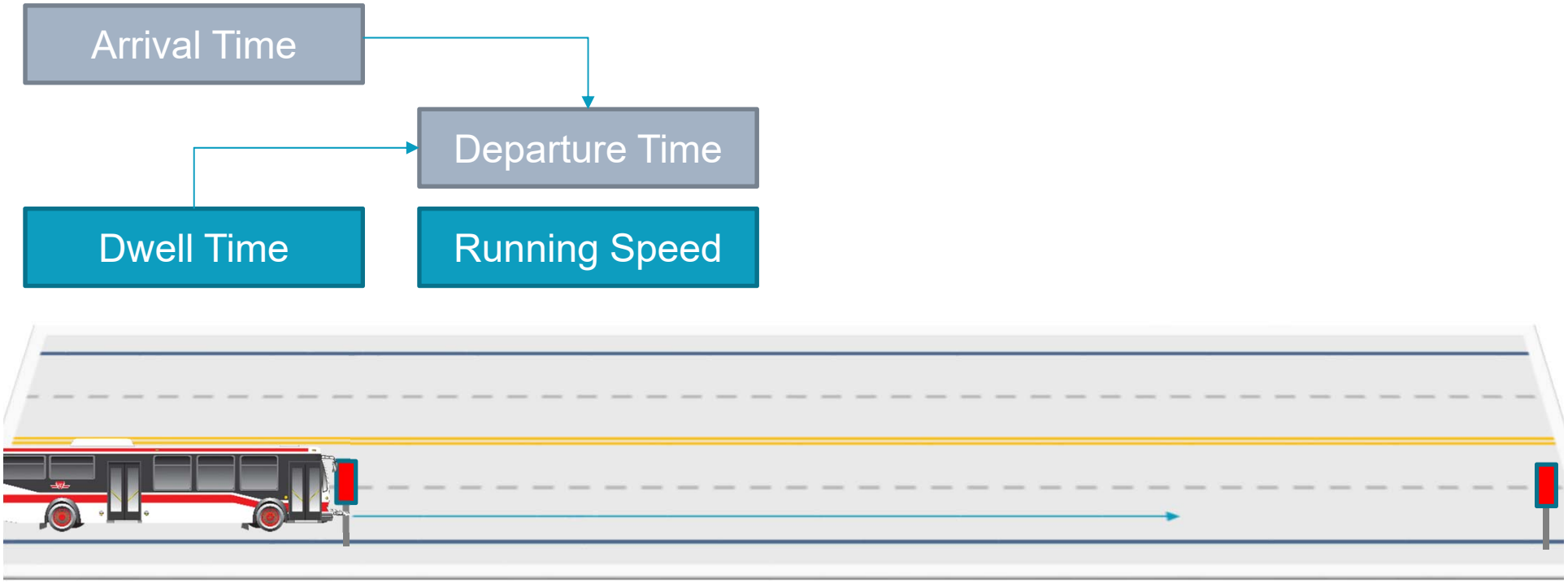
- Develop segment-based (stop-to-stop) transit simulation model based on running speeds and dwell times

- Measures of effectiveness:
  - Accurate network representations
  - Rapid model construction
  - Efficient simulation

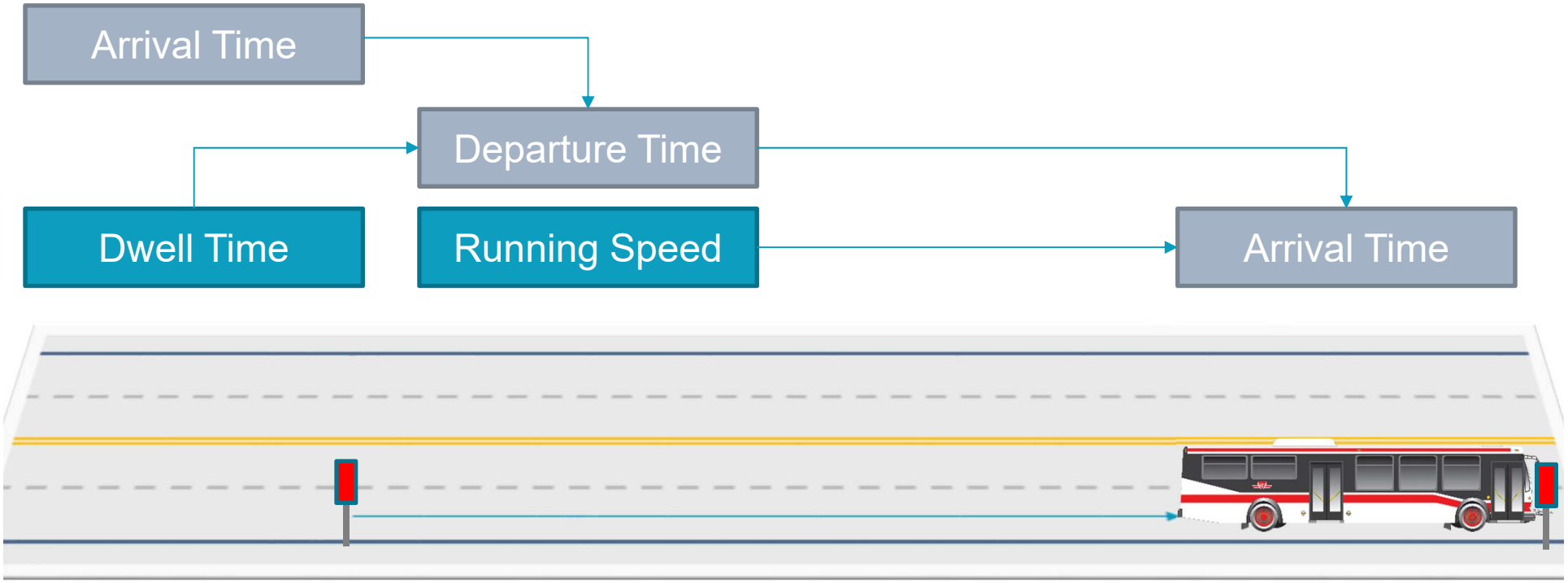
Modelling Framework

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

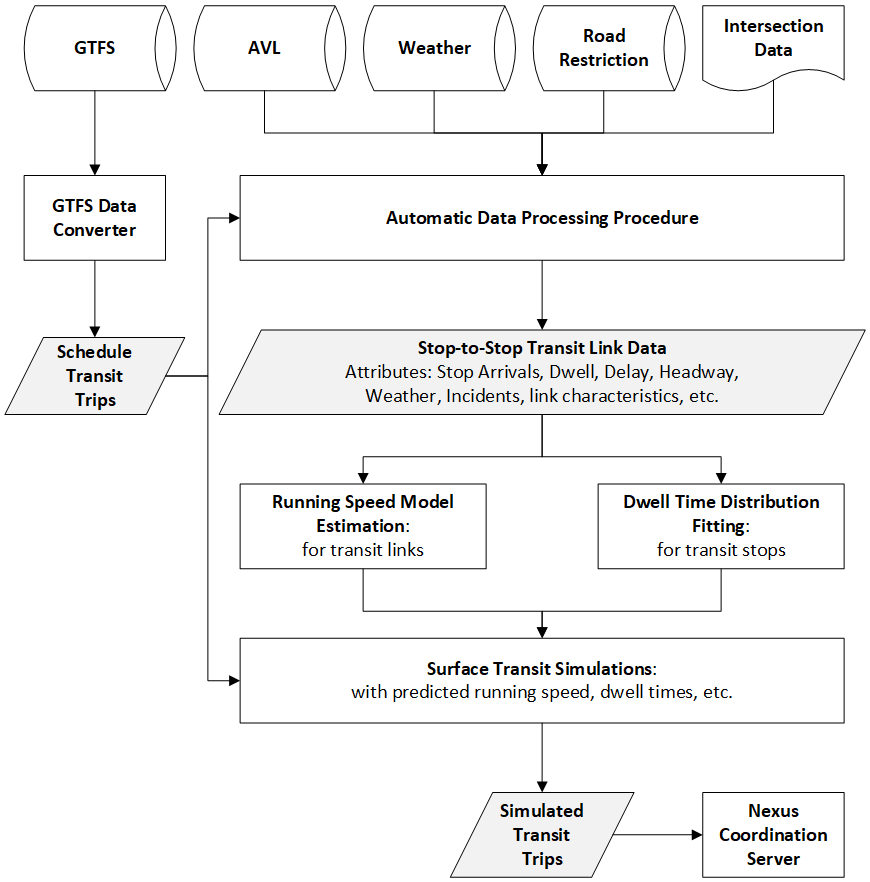
Modelling Framework



Modelling Framework



Modelling Framework



Model Type 1: Basic Analysis

- Route level model

- This type of model accounts for:
  - temporal effects:
    - time of day, and day of the week
  - transit operational characteristics:
    - headway, delay, and previous speeds.
  - basic link characteristics:
    - link distance, link name

Model Type 2: Advanced Analysis

- Network level model

- This type of model accounts for:
  - temporal effects
  - transit operational characteristics
  - expanded link characteristics:
    - stop locations, link distances, link name (link identification), number of signalized intersections, left and right turns made by transit vehicles between stops, traffic and pedestrian volumes
  - route characteristics
    - dedicated right of way, streetcar versus bus route, disruptions, road restrictions or incidents, precipitation.

Data Requirements

- Model 1: Basic Analysis
  - AVL or GPS traces of transit vehicle trips
  - Schedule information about the route

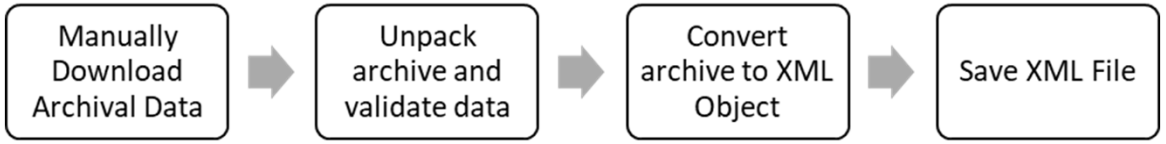
- Model 2: Advanced Analysis
  - AVL data streams for the entire network
  - GTFS transit network schedules
  - Signalized intersection locations
  - Intersection volume data
  - Road restriction data streams
  - Weather data streams

Data

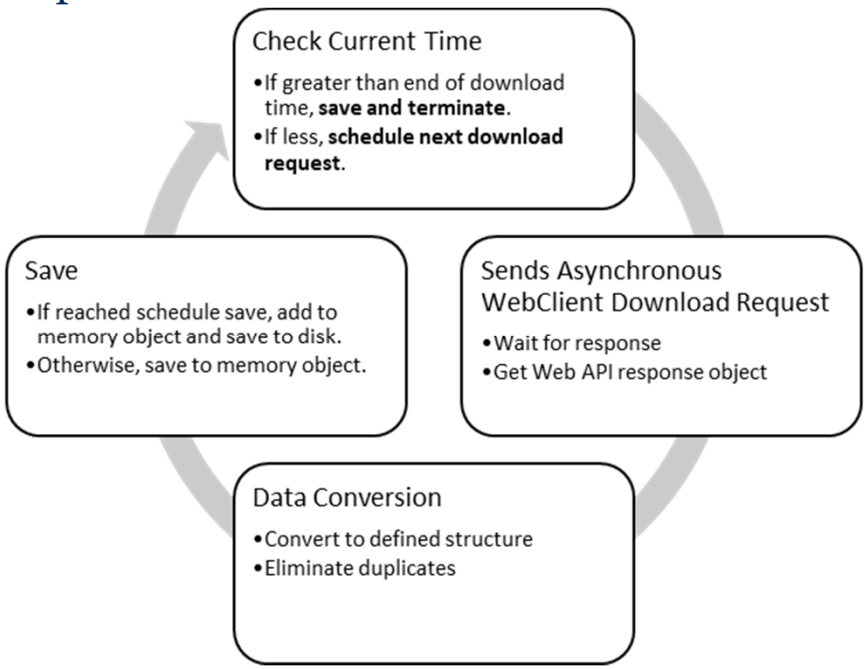
UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

Methods - Automatic Data Collection

- Manual download procedure for archival data



- Automatic download procedure for real-time online API Data



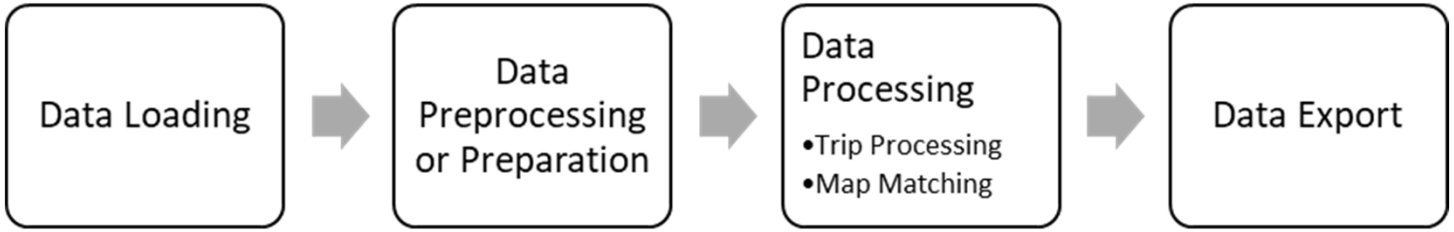
Methods - Automatic Data Collection

- For archival data, retrieval can be performed periodically
  - **General Transit Feed Specification (GTFS)**:
    - Open Data Toronto GTFS data archive
  - **Signalized intersection locations and volume**:
    - Open Data Toronto active archive

- Periodically sends web requests to retrieve real-time data from public APIs throughout the data collection period
  - **Automatic Vehicle Location (AVL)**:
    - Nextbus real-time data streams, 20 seconds resolution
  - **Road restriction**:
    - Open Data Toronto real-time data streams, periodic updates
  - **Weather**:
    - OpenWeatherMap real-time data streams, 3-hour precipitation

Methods - Data Processing

- Program procedure for data processing



- Processes unstructured location and feature data into structured and defined variables

- Preprocesses the data to
  - exclude duplicate points and
  - invalid points (illogical locations)

Methods - Data Processing

- Use AVL and GTFS data to compute various transit operational characteristics.
  - Trip construction
  - Trip matching based on trip geometry
  - Compute trip characteristics:
    - Arrival times,
    - Dwell times,
    - Headway,
    - Delay, etc.

- Spatially and temporally matched additional data to transit trips
  - Signalized intersection location and volume
  - Road restrictions
  - Weather

Methods – Variable Definition

Variable Name	Description	Variable Type	Typ. Range
RunningSpeed	Arrival to arrival speed between two stops, dependent var. for running speed model	Continuous	0 to 120 kph
DwellTime	Dwell time at the start stop, dependent variable for dwell time model	Continuous	0 to 300 secs
RouteCode.f	Route Code of travelling vehicle	Categorical	163 levels
hasIncident.f	If the link segment has road restriction	Categorical	0, 1
prevLinkRunningSpeed	Previous Running Speed upstream of the current link	Continuous	0 to 120 kph
prevTripRunningSpeed	Previous Trip's Running Speed on the current link	Continuous	0 to 120 kph
Day.f	Day of week	Categorical	0 to 6
Time_mins	Time of day in minutes since start of study period	Continuous	0 to 86,400 mins
linkDist	Distance of the current link	Continuous	0 to 11,600 m
Delay	Estimated schedule delay experienced by the vehicle on the link	Continuous	-1000 to 5000 s
Headway Ratio	The Ratio between Scheduled and Estimated headway of the vehicle at a stop	Continuous	0 to 30
totalPptn	Total precipitation reported at the nearest weather station to current link	Continuous	0 to 10 mm
num_VehLtTurns	Number of Left Turns by the transit vehicle on the link	Categorical	0 to 2
num_VehRtTurns	Number of Right Turns by the transit vehicle on the link	Categorical	0 to 3
num_VehThroughs	Number of through movements at intersections made by the transit vehicle	Categorical	0 to 14
num_TSP_equipped	Number of TSP equipped intersections on the link	Categorical	0 to 6
num_PedCross	Number of pedestrian crossings on the link	Categorical	0 to 3
sum_SigIntxnApproach	Total number of signalized approaches of the intersections on the link	Categorical	0 to 49
avgVehVol	Average vehicle volume of the link	Categorical	0 to 20,000
avgPedVol	Average pedestrian volume of the link	Categorical	0 to 10,000
isStartStopNearSided.f	If start stop is near sided	Categorical	0 or 1
isEndStopFarSided.f	If end stop is far sided	Categorical	0 or 1
isStreetcar.f	If the route on the link a streetcar route	Categorical	0 or 1
isSeparatedROW.f	If the link on the route separated right-of-way	Categorical	0 or 1
linkName	The name of the link	Categorical	9267 levels

*Italicized variables are used in network level models

Model Estimation

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

Methods – Model Estimation

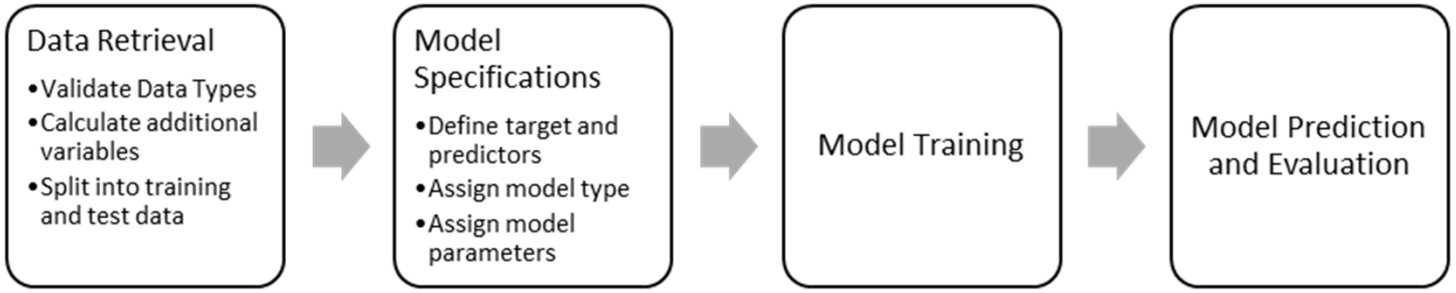
- Running Speed Regression Models
  - Multiple Linear Regression (MLR)
  - Support Vector Machine (SVM)
  - Linear Mixed Effect Model (LME)
  - Regression Tree (RT)
  - Random Forest (RF)

- Dwell Time Model
  - dwell times at transit stops followed the lognormal distribution [1-5]

1. Bellei and Gkoumas, 2010; 2. Li et al., 2012; 3. Meng and Qu, 2013; 4. Rashidi et al., 2014; 5. Zhang Jian and Bai Hai-jian, 2015

Running Speed Model Estimation

- Program procedure for estimating regression models



- Running Speed Model trained in R, using R.Net via C#
  - Efficient data manipulation (with R data.table)
  - Open source machine learning packages
  - Rapid model prototyping

Multiple Linear Regression (MLR)

- Based on ordinary least squares.
- Four fundamental assumptions [1]:
  - Linear relationships
  - Homoscedasticity
  - Normally distributed errors
  - Independency

- The general form of Multiple Linear Regression model[2] :

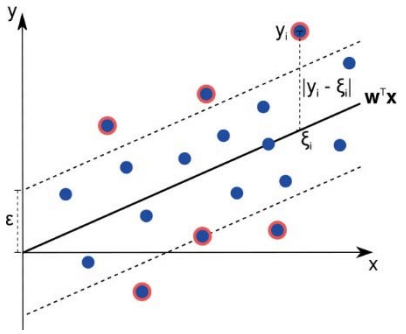
$$\mathbf{Y = a + b_1X_1 + b_2X_2 + \ldots + b_iX_i + \varepsilon}$$

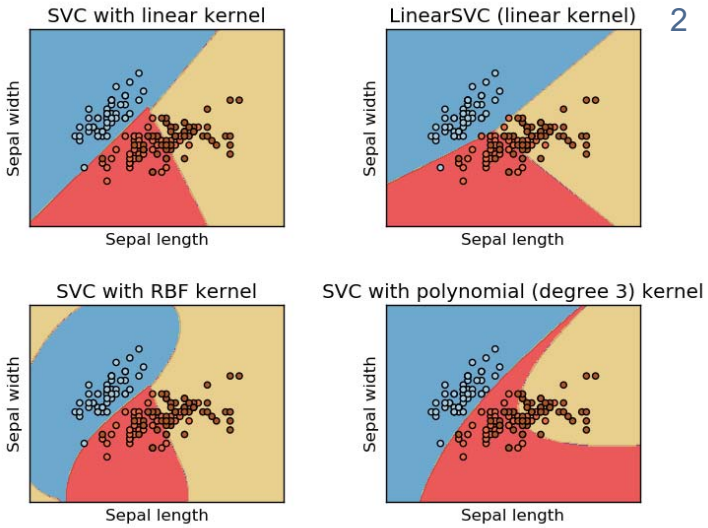
  - **Y**: response variable,
  - $\mathbf{b_i}$: estimated coefficients for predictor variables,
  - $\mathbf{X_i}$: predictor variables,
  - **ε**: residuals

1. Marill, 2004

Support Vector Machine (SVM)

- Based on hyperplane margin optimization [1]
  - Edge training points “supports” the minimum margin vector
- Kernel Functions
  - Linear: fast
  - Polynomial: can become too wavy, and it is very slow.
  - Radial Basis function: commonly used, most flexible, but slower than linear kernel.
- Different loss functions determines how model is trained:
  - ν-SVR: controls number of vectors
  - ε-SVR: penalizes errors
- ε-SVR is most suitable
  - Consistent objective in reducing errors
  - Need to address overfitting with cross-validation





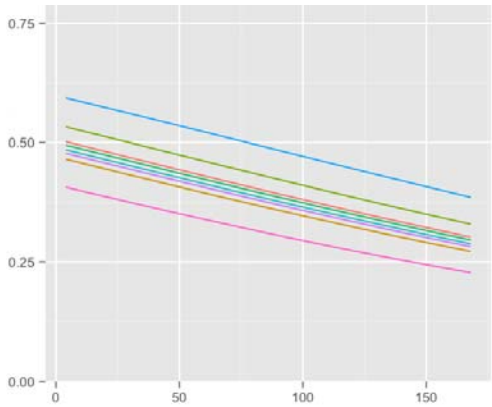
2

1. Chang and Lin, 2011;
2. Scikit-learn developers, 2014

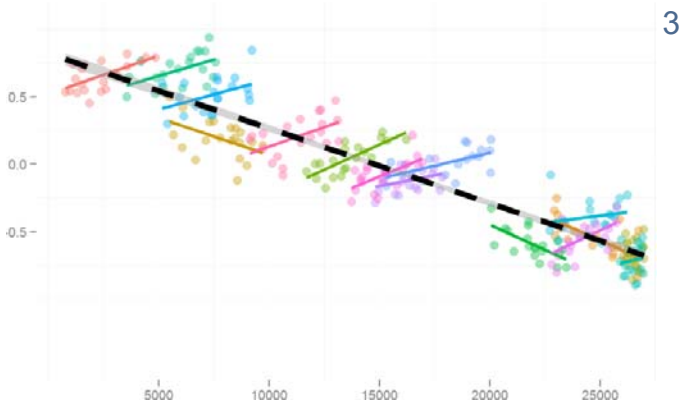
Linear Mixed Effect Model (LME)

- Accounts for the random sampling variations due to repeated measurements. [1]
  - Deals with heteroscedasticity
- Models random effects by:
  - Varying intercepts
  - Varying slopes
- Same assumptions for each level of the random effects as MLR:
  - Linear relationships
  - Normally distributed errors

2



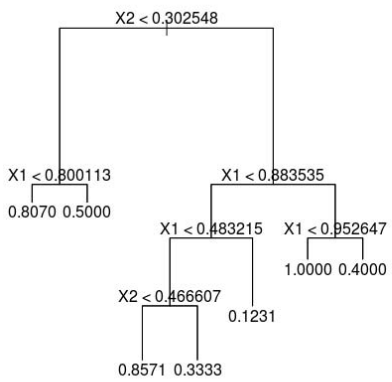
3

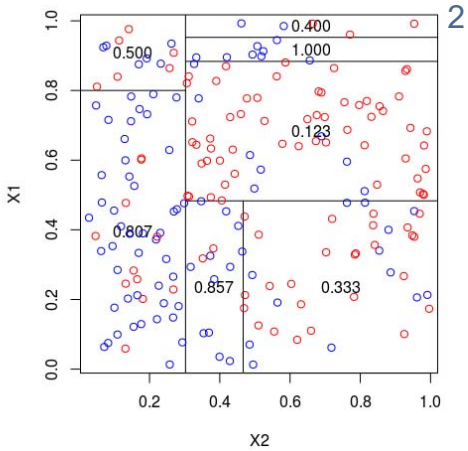


1. Bates et al., 2015;
2. Daniel Von, 2014;
3. Human Language Processing (HLP) lab at the University of Rochester, 2014

Regression Tree (RT) [1]

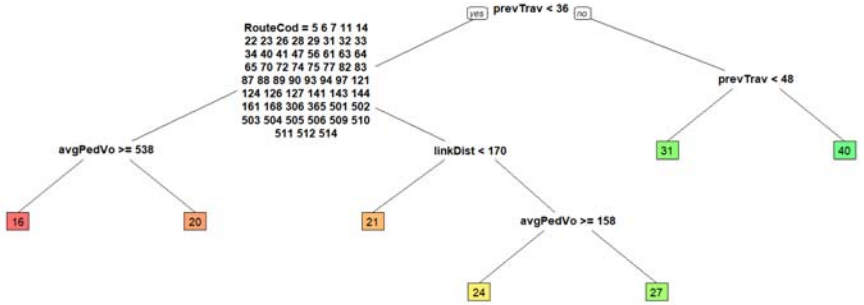
- Partition to determine data clusters.
- Construction of trees are based on splitting criteria.
- Aims to minimize Gini impurity, thus reduce probability of misclassifications.
- Variables that affects the split the most is the most important.





2

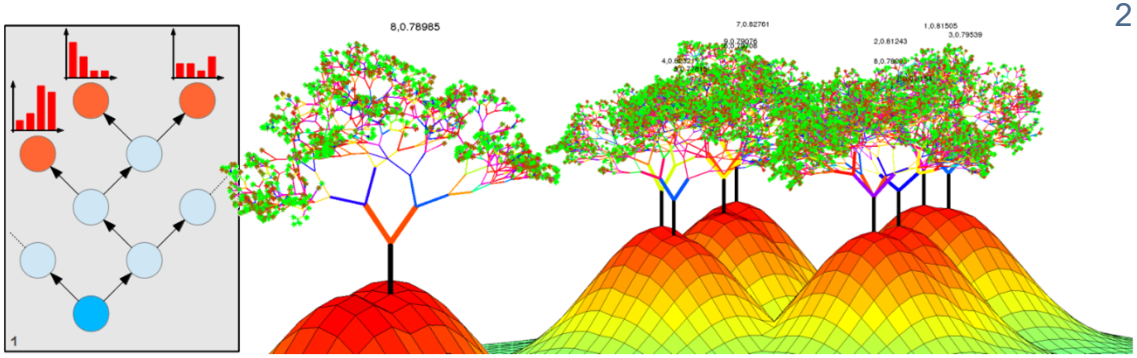
- Complexity and depth of tree are determined by
  - complexity parameter (cp)
  - minimum split criteria
  - prune cp
- Tree pruning with cross-validation can minimize overfitting.



1. Terry M. Therneau and Elizabeth J. Atkinson, 2017;
2. Charpentier, 2013

Random Forest (RF)

- Grow a number of trees based on random draws of the original samples (with replacements) [1]
- An ensemble method:
  - Each tree is a weak learner, but collectively are strong
  - The result from all the trees produces a single prediction
- Works well for clustered data and can replicate complex relationships
- Each draw is independent
- Low correlation needed between residuals and between trees
- Shown not to overfit and reduce bias



2

1. Breiman, 2001;
2. R. Hänsch and O. Hellwich, 2015

Comparisons of Running Speed Models

- Model Fitness

$$R^2 = 1 - \frac{\mathrm{SS_R}/\mathrm{df_e}}{\mathrm{SS_T}/\mathrm{df_t}}, \mathrm{df_e} = n - 1, \mathrm{df_t} = n - p - 1$$

- Mean absolute percentage error:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=0}^{n}\left|\frac{\hat{y}_i - y_i}{y_i}\right| \times 100\%$$

- Mean absolute error:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=0}^{n}|\hat{y}_i - y_i|$$

- Relative absolute error:

$$\mathrm{RAE} = \frac{\sum_{i=0}^{n}|\hat{y}_i - y_i|}{\sum_{i=0}^{n}|y_i - \bar{y}|}$$

Comparisons of Running Speed Models

- Root mean square error:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(\hat{y}_i - y_i)^2}$$

- Root relative square error:

$$\text{RRSE} = \sqrt{\frac{\sum_{i=0}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=0}^{n}(y_i - \bar{y})^2}}$$

- Relative differences in RMSE (RD):

$$\text{RD} = (\text{RMSE}_i - \text{RMSE}_{\text{MLR}}) / \text{RMSE}_{\text{MLR}}$$

- Training Time and Test Prediction Time

Methods – Model Estimation

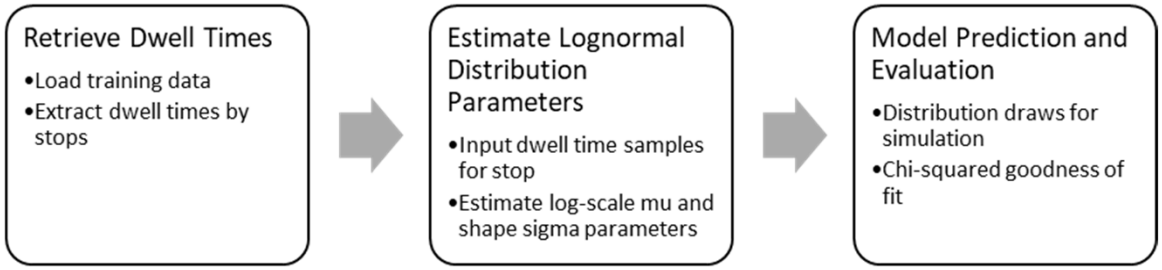
- Running Speed Regression Models
  - Multiple Linear Regression (MLR)
  - Support Vector Machine (SVM)
  - Linear Mixed Effect Model (LME)
  - Regression Tree (RT)
  - Random Forest (RF)

- Dwell Time Model
  - dwell times at transit stops followed the lognormal distribution [1-5]

1. Bellei and Gkoumas, 2010; 2. Li et al., 2012; 3. Meng and Qu, 2013; 4. Rashidi et al., 2014; 5. Zhang Jian and Bai Hai-jian, 2015

Dwell Time Model Estimation

- Program procedure for estimating distribution models



- Dwell Time Model trained in native C#
  - Stop-based models, trained using historical dwell times at the stop
  - Lognormal distribution
  - Open source statistical package (with Math.NET Numerics)

Dwell Time Models

- Model Estimation: Lognormal Distribution
  - Estimation of log mean parameter

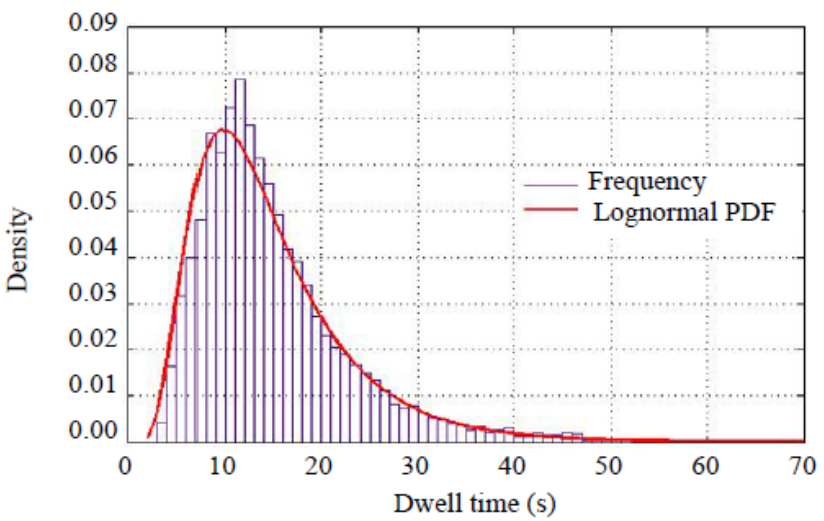
$$\hat{\mu} = \frac{\sum_{i=1}^{k} \ln x_k}{n}$$

  - Estimation of shape parameter

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k} (\ln x_k - \hat{\mu})^2}{n}$$

- Model evaluation: chi-squared goodness of fit

$$\chi^2 = \sum (P_i - O_i)^2 / O_i$$



1

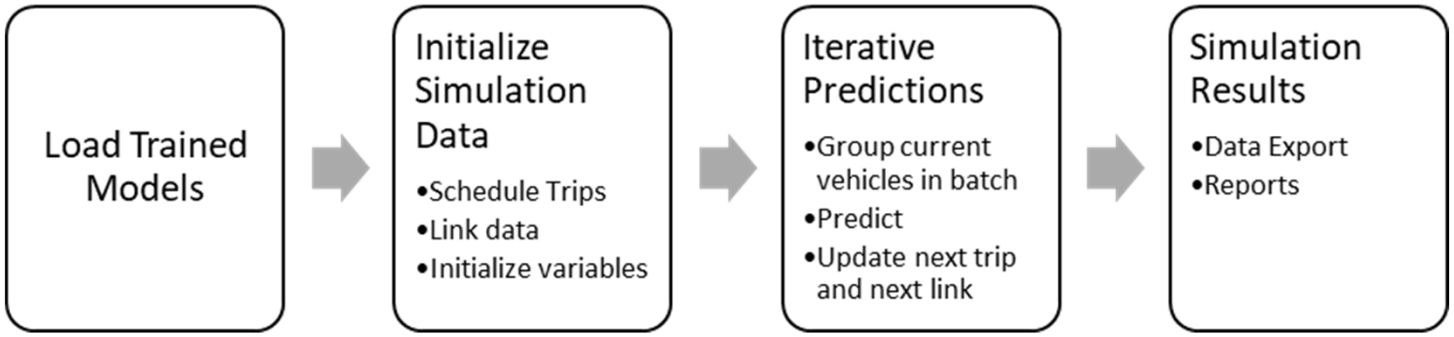
1. Li et al., 2012

Data-driven Simulation

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

Model Simulation

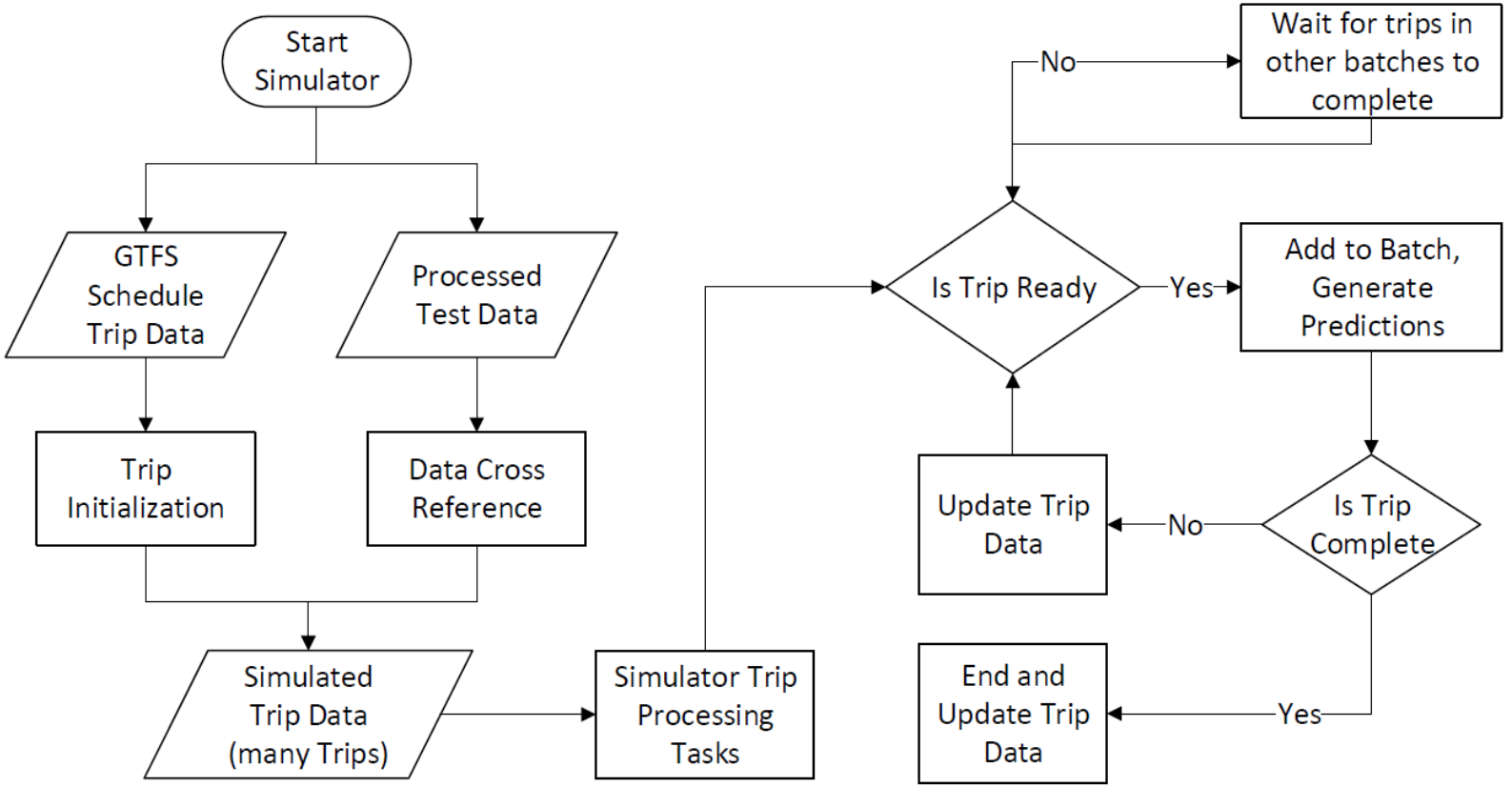
- Program procedure for simulation



- Base case scenario used transit schedule departures from terminals with no short turns

- Running speed and dwell time models predicted mesoscopic transit movements

Model Simulation – Iterative predictions



Results
Case Study: Toronto Transit Commission network

UNIVERSITY OF TORONTO
FACULTY OF APPLIED SCIENCE & ENGINEERING
Transportation Research Institute

Case Study: the TTC network

- The Toronto Transit Commission (TTC) provides public transit in the city of Toronto.
  - Population of Toronto: 2.8 Million
- 4 subway/rail lines, 11 streetcar routes, and over 140 bus routes
- Period of the case study
  - Training Data: 2017-02-28 to 2017-03-02, 6AM to 9AM (AM Peak)
  - Test Data: 2017-03-07 to 2017-03-09, 6AM to 9AM

Summary of Data during study period

- GTFS
  - 8304 Trips (typical weekday AM peak)
- AVL
  - 8381 Trips (Feb 28), 8350 Trips (Mar 1), 8403 Trips (Mar 2), 8428 Trips (Mar 7), 8395 Trips (Mar 8), 8414 Trips (Mar 9)
- Road Restrictions
  - 734 Events (Feb 28 to Mar 2), 766 Events (Mar 7 to Mar 9)
- Weather
  - 72 Records (per day)
- Traffic intersections
  - 2269 Records (intersection volumes)
  - 71 Records (minor intersections)

Running Speed Model Results (Network)

Model Type	MLR	SVM	LME	RT	RF (100 trees)
R Package	MASS	liquidSVM	LME4	RPART	RANGER
R^2	0.277	0.265	0.387	0.225	0.359
MAPE	0.355	0.358	0.311	0.372	0.325
MAE	7.625	7.677	6.902	7.911	7.109
RAE	0.831	0.837	0.752	0.862	0.775
RMSE	9.950	10.035	9.160	10.303	9.366
RRSE	0.850	0.858	0.783	0.881	0.800
Reduction in RMSE	-	-0.9%	7.9%	-3.5%	5.9%
Training Time (min.)	0.419	36.272	2.629	1.021	14.681
Prediction Time (min.)	0.036	3.249	0.049	0.015	0.331

Running Speed Model Results (504-King)

Model Type*	MLR	SVM	LME	RT	RF (100 trees)
R Package	MASS	liquidSVM	LME4	RPART	RANGER
R^2	0.107	0.115	0.223	0.102	0.153
MAPE	0.329	0.326	0.296	0.330	0.318
MAE	5.127	5.077	4.726	5.134	4.982
RAE	0.940	0.931	0.866	0.941	0.913
RMSE	6.816	6.784	6.359	6.834	6.639
RRSE	0.945	0.941	0.882	0.947	0.920
Reduction in RMSE	-	0.5%	6.7%	-0.3%	2.6%
Training Time (sec.)	0.017	31.790	0.327	0.662	3.838
Prediction Time (sec.)	0.011	2.286	0.076	0.012	0.158

* Route-level model trained using data from 504-King only.

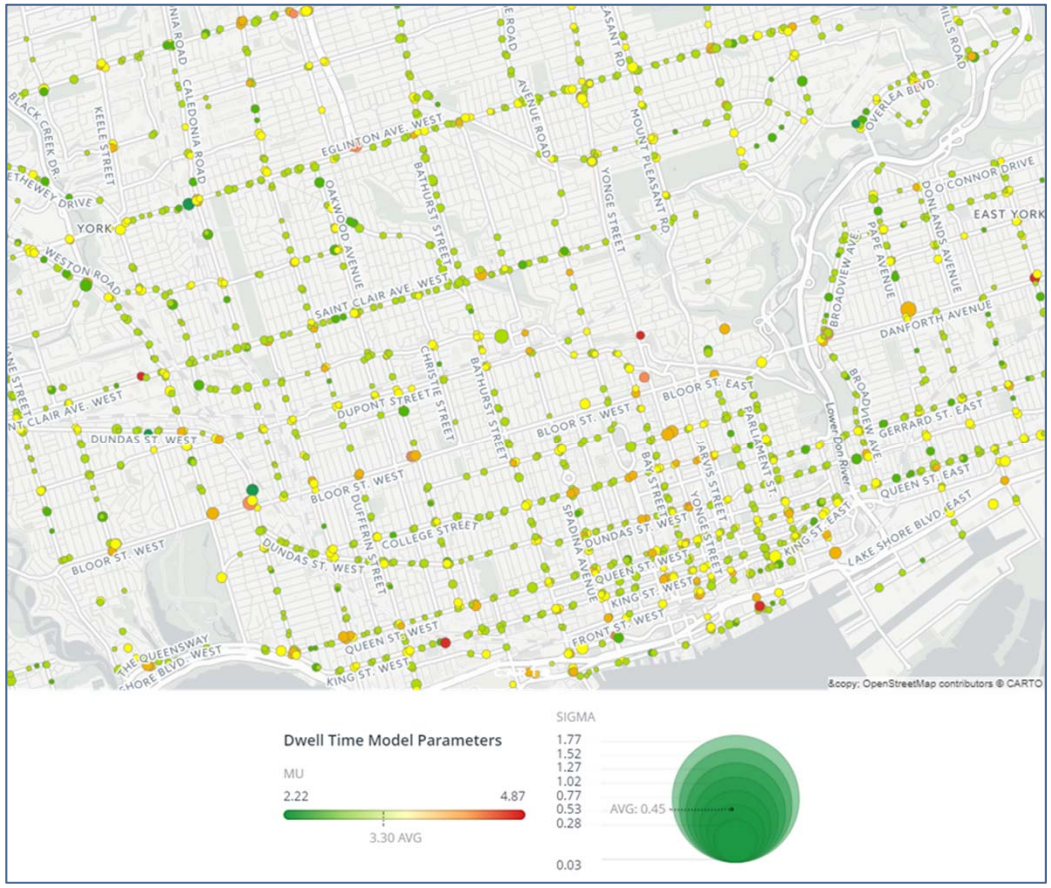
Running Speed Model Results

- Sample size
  - Network level: training =593,234, test = 600,351
  - Route level (504-King only): training = 12,827, test = 12,612

- RT and SVM did not provide improvements over MLR
- SVM provided small improvements over MLR for route level models

- LME model yields the best result with:
  - varying intercept model
  - link identification (link name) as the random effect variable
- RF did well and provided a more flexible implementation
  - allows new links, whereas LME model does not
- LME is more computationally efficient than RF.

Dwell Time Models: Parameters



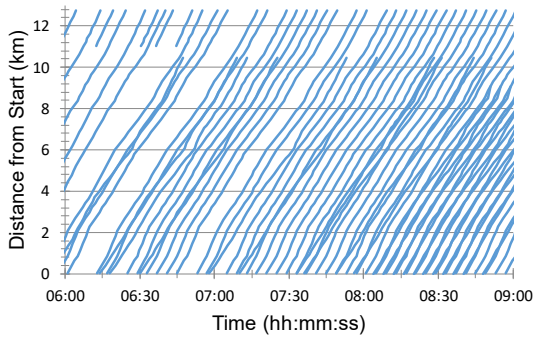
Dwell Time Model: Observed vs Predicted

Simulation Model Results

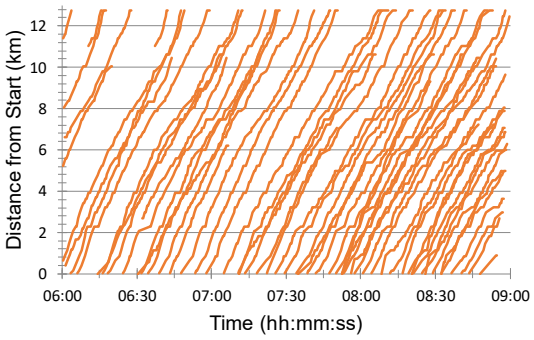
- Simulation scenario:
  - Weekday schedule
  - On-time terminal schedule departure, if possible.
  - No short turns
  - Road conditions from test day: 2017-03-08, 6AM to 9AM
    - 704 road restriction events (Mar 8 only)
    - 72 weather records per day
    - Intersection attribute data for links
- Simulations using RF and LME were generated.
- Comparisons of vehicle trajectories with time-distance diagrams.
- Model Validations
  - Route level with route speeds
  - Stop level with stop delays

Simulation Model Results - RF

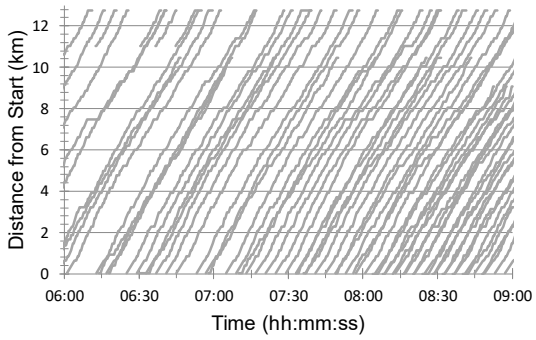
504 KING EB - Scheduled



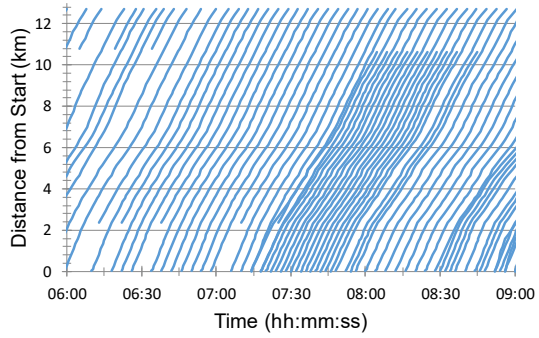
504 KING EB - Observed



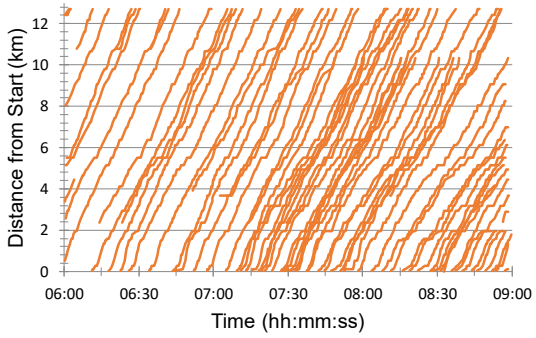
504 KING EB - Simulated



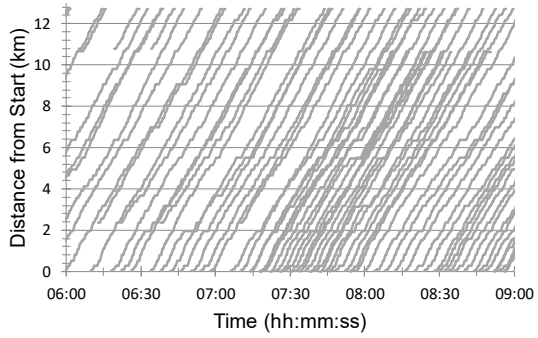
504 KING WB - Scheduled



504 KING WB - Observed

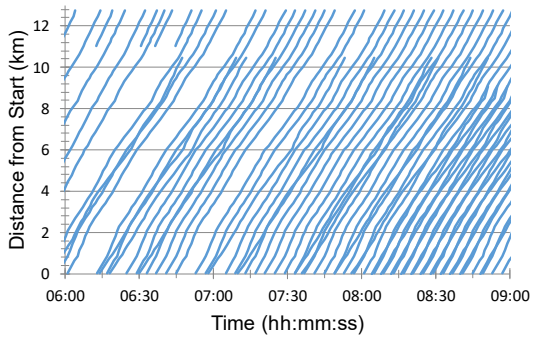


504 KING WB - Simulated

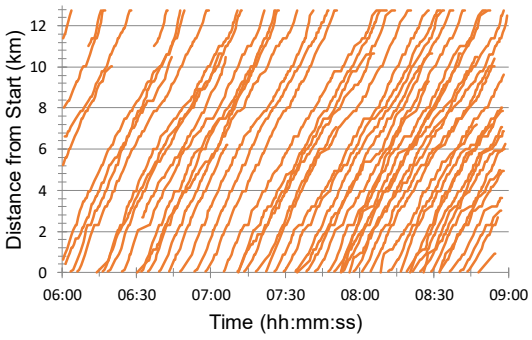


Simulation Model Results - LME

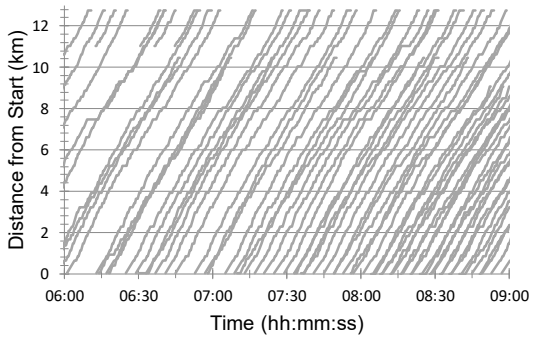
504 KING EB - Scheduled



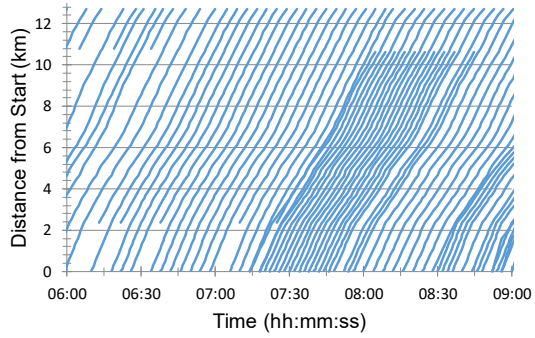
504 KING EB - Observed



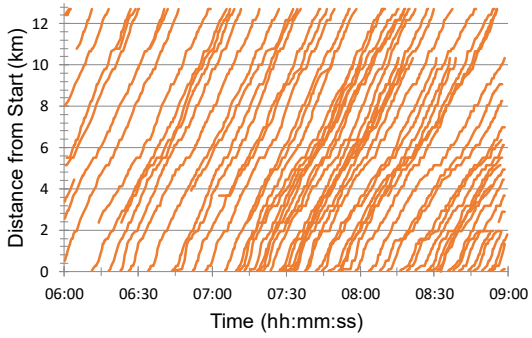
504 KING EB - Simulated



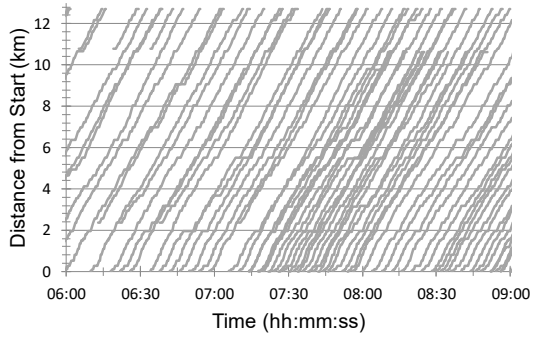
504 KING WB - Scheduled



504 KING WB - Observed

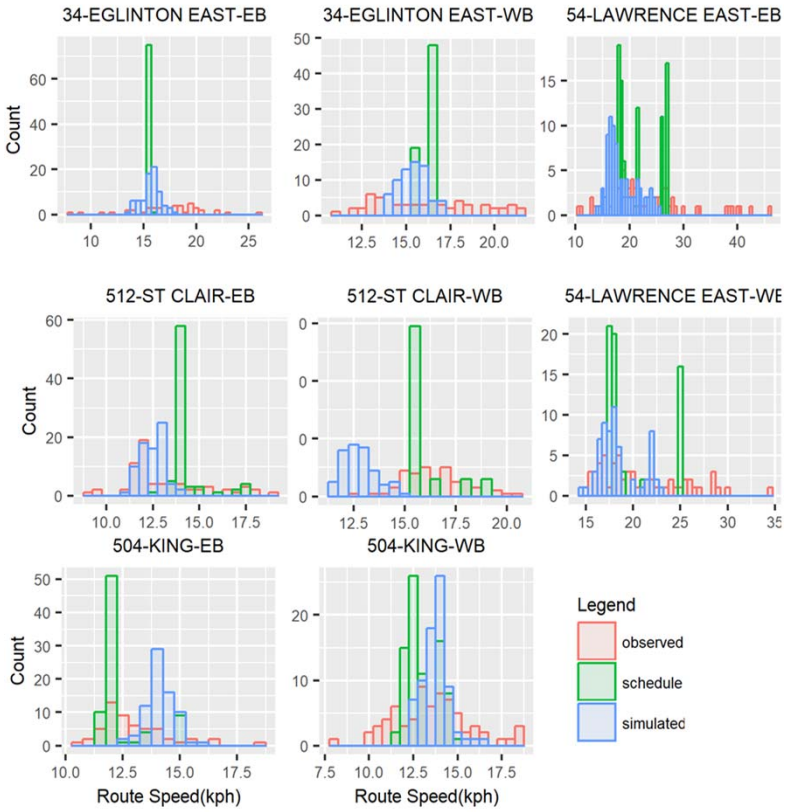


504 KING WB - Simulated

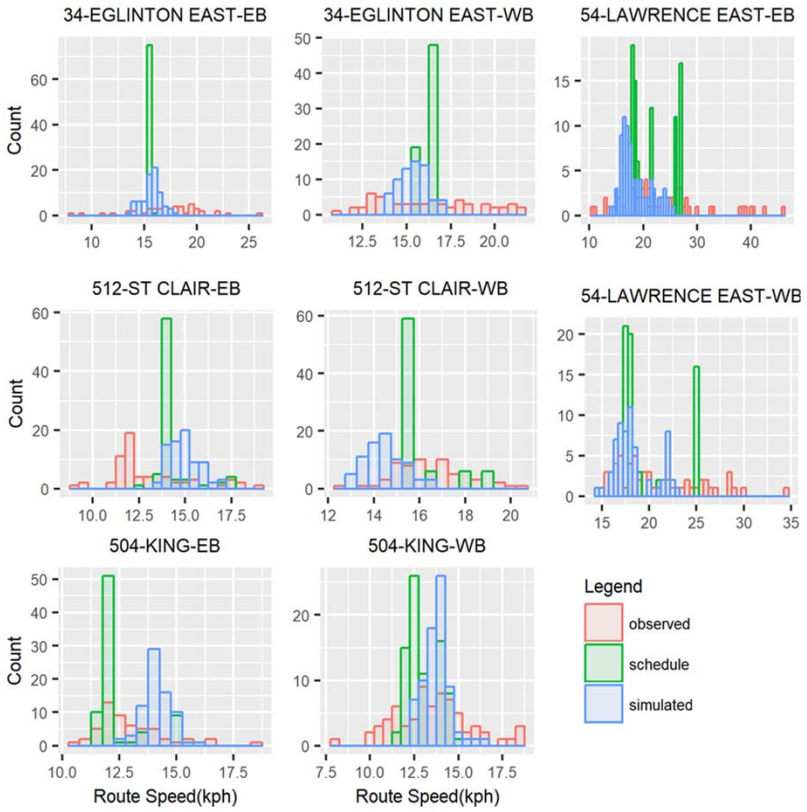


Model Validations – Route Speeds

Random forest

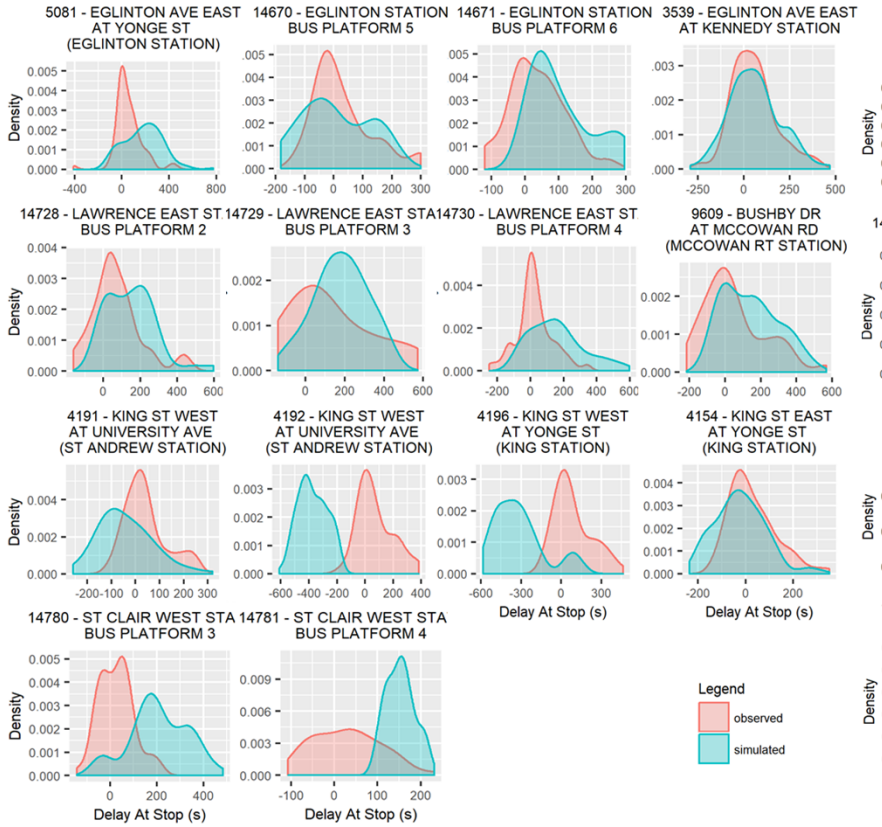


Linear Mixed Effect

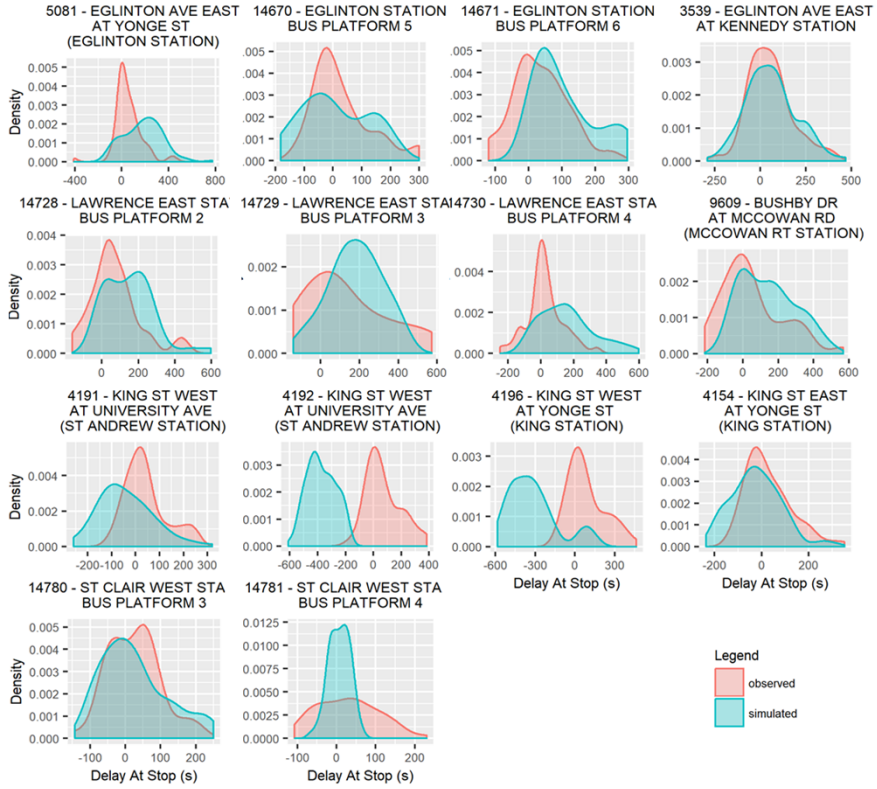


Model Validations – Stop Delays

Random forest



Linear Mixed Effect



Findings

- Running speed model comparisons
  - LME model accuracy outperformed MLR by 8%
  - RF model accuracy outperformed MLR by 6%
  - LME has lower training time, but requires repeated observations from existing links.

- Lognormal dwell time introduce realistic stochasticity into vehicle movements.

- Simulation model prediction runtimes
  - RF (ranger package): 36 minutes
  - LME (lme4 package): 1 minute

Findings

- A data-driven transit simulation model
  - replicated instances of vehicle bunching, distribution of dwell times, and stochastic patterns of delays and headways
- Validation results suggests the need to incorporate:
  - Effect of traffic congestion
  - Signal delays
  - Vehicle short-turns

Future Research

- Model the effects of short-turning vehicles
- Incorporate congestion data
- Advanced dwell time models to incorporate passenger demand
  - Allows reallocation of passenger demand
  - Stop addition, relocation, and removals
- Continuous model training for streaming data

Acknowledgements

ARUP

NSERC
CRSNG

UNIVERSITY OF
TORONTO