

Inferring the Purposes of Using Ride-Hailing Services through Data Fusion of Trip Trajectories, Secondary Travel Surveys, and Land-Use Attributes

UT ITE Seminar
February 14, 2020

Sanjana Hossain, M.Sc.
Supervisor: Khandker Nurul Habib, PhD, PEng

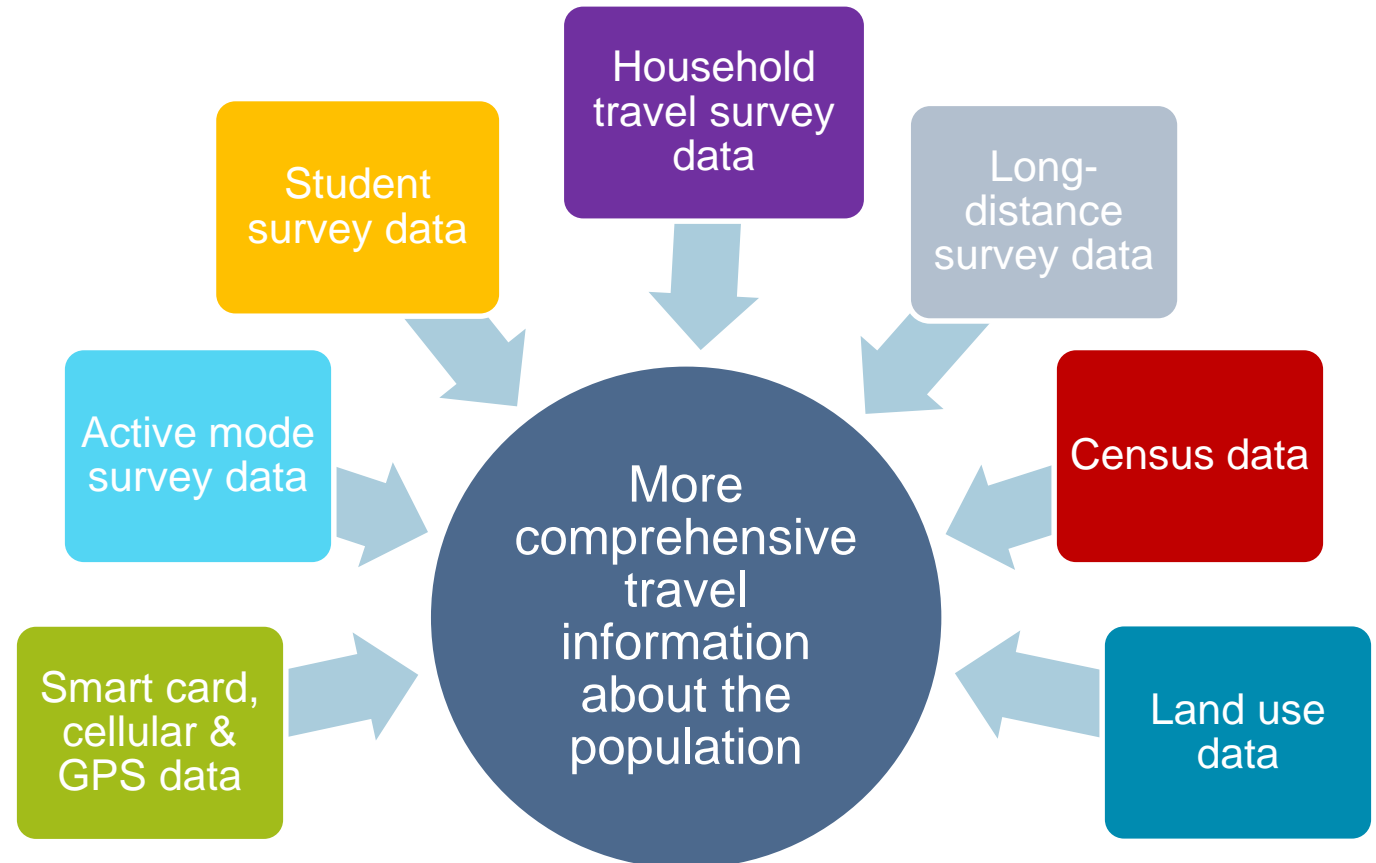
Outlines

- Thesis framework
 - Background
 - Conceptual framework
 - Objectives
- Empirical investigation: Ride-hailing trip purpose inference
 - Background and research motivation
 - Purpose inference methodology
 - Data for empirical investigation
 - Model estimation and results
 - Validation of inferred trip purposes
 - Key findings and conclusions

Data fusion for travel demand analysis

■ Data fusion

- enrich the quality of a sample of travel data by combining it with other data sources
- either to add variables or to update the sample



Need for data fusion

Growing methodological issues of HTS

- incomplete sample frames
- low response rates
- under-representation of certain sub-populations
- reporting errors

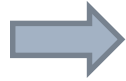
More detailed data requirements of advanced TDM

- multi-day information
- flexible mobility options (AV, MaaS) affecting
 - mobility tool ownership
 - vehicle allocation
 - feasible choice sets of modes and locations
 - user values of time
 - parking costs

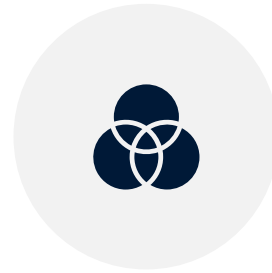
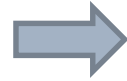
The data fusion process



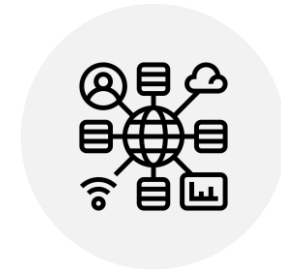
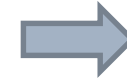
IDENTIFY APPROPRIATE
DATASETS BASED ON
PURPOSE OF FUSION



EXAMINE DATA
CHARACTERISTICS OF
EACH OF THE SOURCES



IDENTIFY COMMON (OR
SIMILAR) DATA
ELEMENTS THAT
FACILITATE DATA FUSION



ANALYZE AND
INTEGRATE DATASETS
USING APPROPRIATE
FUSION TECHNIQUE

Challenges of fusing travel data

- Data incompatibilities in different contexts
 - Spatial
 - Temporal
 - Semantic: Household vs Individual travel surveys
- Choice of matching variables
- Non-response bias
- Other uncertainties
 - Input uncertainties: Random/systematic measurement uncertainty, Scenario uncertainty on ultimate model forecasts
 - Model uncertainties: Model specification uncertainty, Parameter uncertainty

Objectives of the thesis

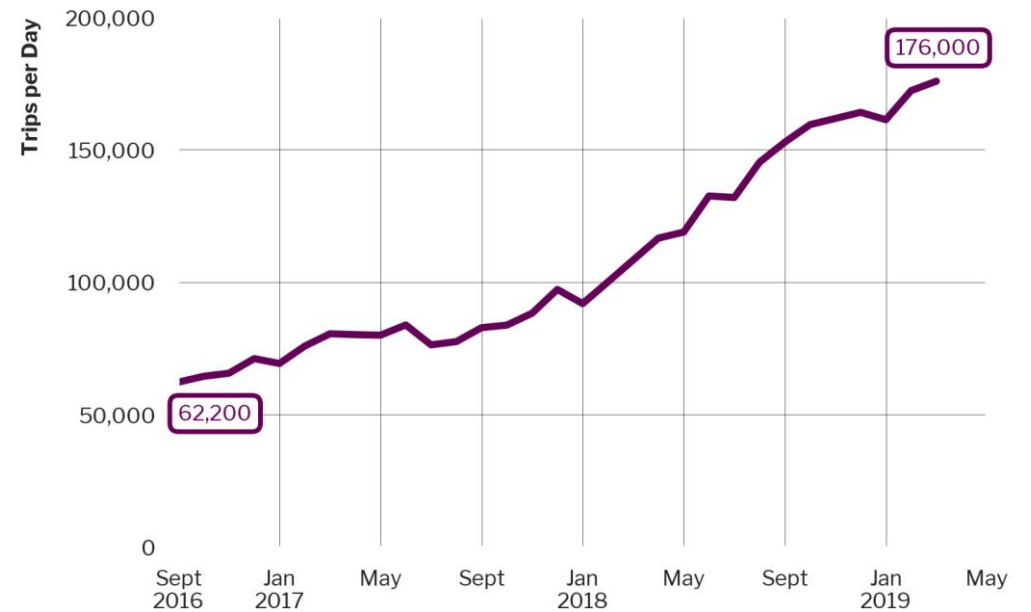
- To develop innovative methods for fusing passive data sources with traditional data sources to facilitate the analysis of travel behavior
 - Ride-hailing trajectory data
 - Smart card transaction data
- To investigate the necessity of fusing data from different time periods to account for changing travel patterns due to (i) seasonal variation and (ii) weekday versus weekend variation in data sets
 - Applicability of the continuous passive data fused with additional variables
- To develop methods for optimizing the performance of demand models using a combination of data sources

- Inferring the Purposes of Using Ride-Hailing Services through Data Fusion of Trip Trajectories, Secondary Travel Surveys, and Land-Use Attributes



Background

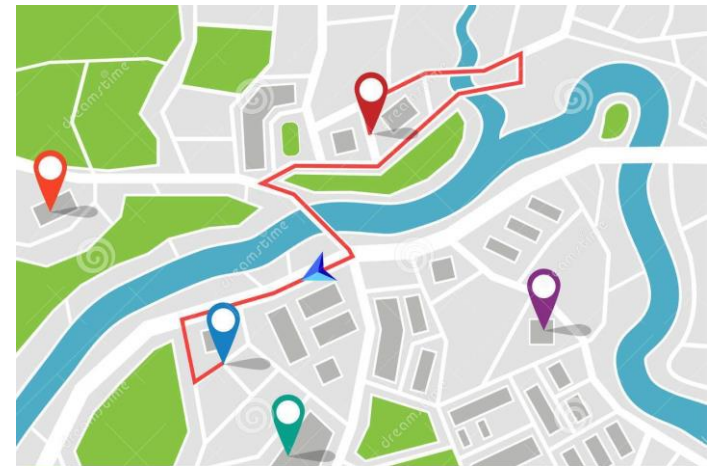
- Ride-hailing services are growing rapidly
 - flexibility
 - reliability
 - cost-effectiveness
- Need to understand the characteristics of these trips and how the services are changing the travel behaviour of people




Source: The Transportation Impacts of Vehicle for Hire Report by the Big Data Innovation Team of the City of Toronto

Research Motivation

- Trip purpose relates to the activities for which ride-hailing is used
 - Thus provides important context of travel demand generated by the services
- GPS trajectory contain when and where passengers move in a high resolution
- But it does not have trip purposes




Trade-off between trajectory and survey data



Travel survey

- detailed trip purposes
- small sample size and inaccuracies



Trajectory data

- rich spatial and temporal information
- no trip purposes

- Leverage both of the information sources (along with land use data) to infer ride-hailing trip purposes

Previous works on Trip Purpose Inference

Passive data sources

GPS based travel surveys

AFC/Smart card transaction data

Mobile phone CDR

Taxi trajectory

Ride-hailing trajectory

Methodology

Rule-based method (land use and purpose matching tables, heuristic rules, closest POI matching etc.)

Probabilistic methods (MNL, NL, probability calculation based on distance etc.)

Machine learning methods (decision trees, random forest etc.)

Input variables

Land use and POI information

Activity duration

Trip start and end times

Frequent activities

Key addresses

Demographic data

Social network check-in data

Data Fusion Methodology

Travel survey data

Trip ID	Start time	Origin DA	Destination DA	Trip day	Trip Length (km)	Trip purpose
1	8:30	35202880	35203542	1	17.27	work
1096	9:45	35202932	35202910	1	7.333	shopping

Enhanced Points of Interest data

DA	# of EPOI per unit area (sq. km)					
	Agriculture	...	Manufacturing	...	Professional	...
35202516	0		14.3		14.3	
35202551	0		11.8		23.5	

Census data

DA	# of private dwellings per unit area (sq. km)
35202247	3640
35202251	4800

Fusing land use information

Estimation

Discrete Choice Model
that predicts trip purpose
Input variable categories: trip attributes, land use data

Prediction

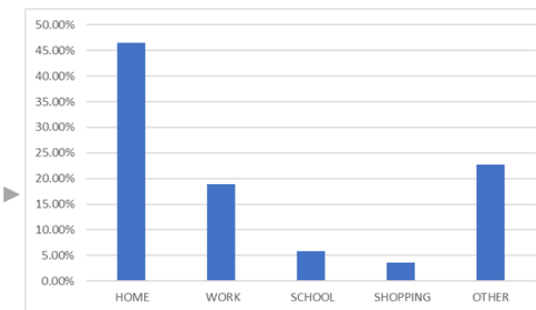
Ride-hailing trip records
Contain time-stamped pick-up and drop-off coordinates (to the nearest intersection)

EPOI data

Fusing land use information

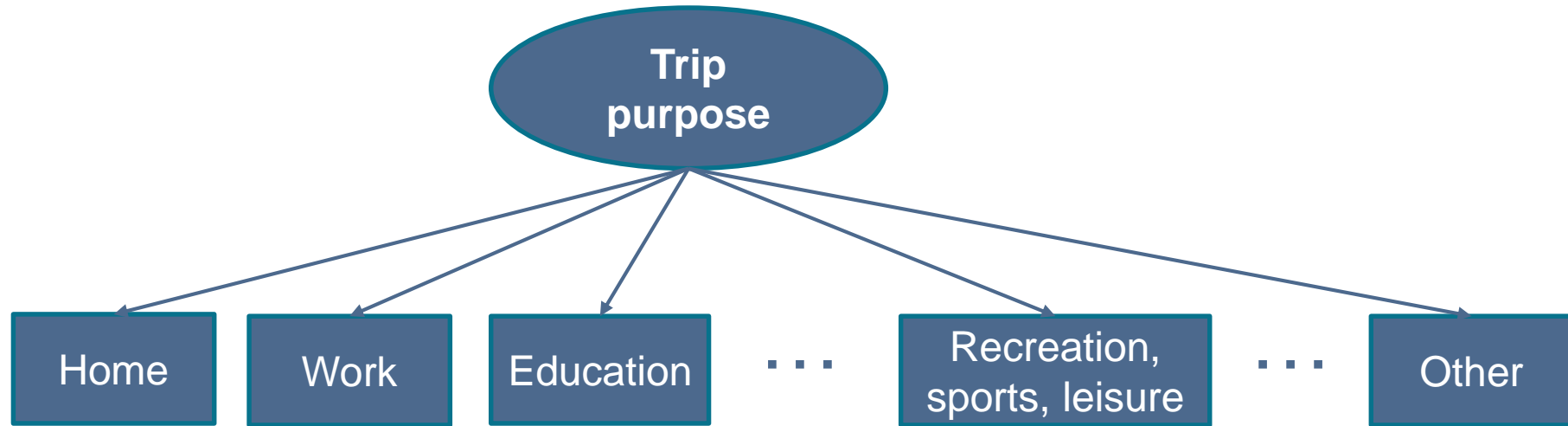
Census data

Inferred trip purpose distribution



Discrete choice models tested (1)

- Multinomial logit model

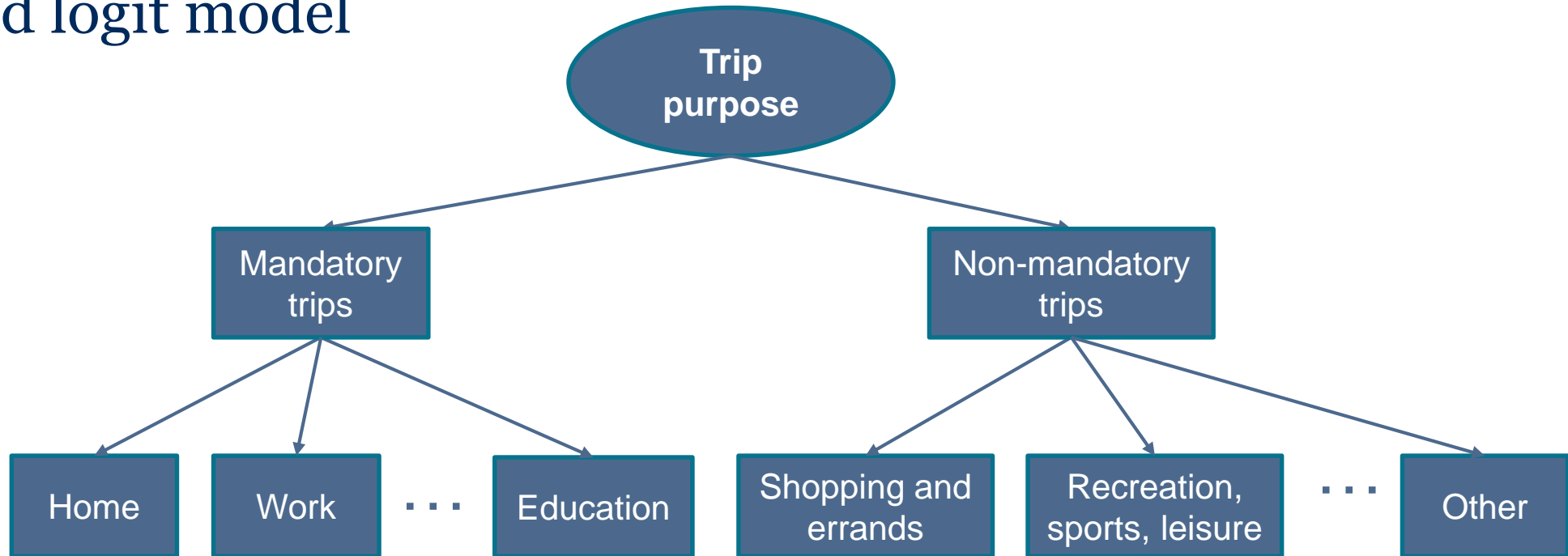


$$- P_{in} = \frac{e^{\mu V_{in}}}{\sum_J e^{\mu V_{Jn}}}$$

- Classical maximum likelihood estimation

Discrete choice models tested (2)

- Nested logit model



$$- P_{in} = \frac{e^{\mu_M V_{in}}}{\sum_m e^{\mu_M V_{mn}}} \frac{\frac{\mu_R}{e^{\mu_M}} \ln(\sum_m e^{\mu_M V_{mn}})}{\frac{\mu_R}{e^{\mu_M}} \ln(\sum_m e^{\mu_M V_{mn}}) + \sum_{J-m} e^{\mu_R V_{(J-m)n}}}$$

$$- P_{ln} = \frac{e^{\mu_R V_{ln}}}{\frac{\mu_R}{e^{\mu_M}} \ln(\sum_m e^{\mu_M V_{mn}}) + \sum_{J-m} e^{\mu_R V_{(J-m)n}}}$$

Discrete choice models tested (3)

- Mixed multinomial logit

- $U_{in} = V_{in} + \eta_{in} + \varepsilon_{in}$

- A heteroskedastic MMNL was found to be valid for the estimation data

$$P_{in} = \frac{1}{D} \sum_{d=1}^D \frac{e^{\mu(\beta X_{in} + \sigma_i \xi_{in}^d)}}{\sum_J e^{\mu(\beta X_{iJ} + \sigma_J \xi_{Jn}^d)}}$$

- Maximum simulated likelihood estimation

- Error simulated using Halton draws

Empirical Analysis for the City of Toronto

- City of Toronto's vehicle for hire bylaw review
- In partnership with UTTRI
- Provided anonymized ride-hailing trajectory data



Data sources

- Ride-hailing trip records from the City of Toronto for September 2016 – September 2018
 - More than 17 million trips



PICK UP AND DROP OFF
LOCATIONS GIVEN TO
NEAREST INTERSECTION



TIMESTAMPS TO NEAREST
MINUTE (HOUR FROM
APRIL 2017)



NO ANONYMIZED USER
IDS

Data sources

■ Person trip survey data

- Web-based survey conducted in summer and fall of 2017
- Collected travel diaries, home and work locations, and socio-demographics
- Subset of 5,065 trips originating and terminating within Toronto
- Detailed trip purpose categories



HOME



WORK



EDUCATION



DAYCARE



FACI. PASS.



SHOP, ERRANDS



EAT OUT



RECREATION,
SPORTS,
LEISURE



ARTS, HEALTH,
PERSONAL CARE



SERVICES



VISITING
FRIENDS, FAMILY



WORSHIP,
RELIGION



OTHER

Data sources

- Enhanced Points of Interest (POI) data from DMTI Spatial
 - Geocoded locations of POI along with their NAICS codes

NAICS major code	Sector name
Sector 31-33	Manufacturing
Sector 44-45	Retail Trade
Sector 52	Finance and Insurance
Sector 54	Professional, Scientific, and Technical Services
Sector 61	Educational Services
Sector 62	Health Care and Social Assistance
Sector 71	Arts, Entertainment, and Recreation
Sector 72	Accommodation and Food Services
Sector 81	Other Services (except Public Administration)
Sector 92	Public Administration

Data sources

- 2016 Canadian Census data
 - Number of private dwellings in each Dissemination Area
- 2016 Transportation Tomorrow Survey (TTS) data
 - Large-scale household travel survey in the Greater Toronto and Hamilton Area
 - Provided a sample of 1264 ride-hailing trips in the City with seven categories of reported trip purposes
 - Used for validating the performance of the inference model

Contextual variables used

Trip attributes	
Start time	Morning (06:01-10:00) Midday (10:01-15:00) Afternoon (15:01-20:00) Evening (20:01-24:00) Overnight (00:01-06:00)
Trip day	Weekday Weekend
Season	Fall Summer
Trip distance	Euclidean distance (in km) between origin and destination of a trip

Contextual variables used

Land use attributes	
NAICS Major Industry Category	Number of different types of business establishments per unit sq. km of trip origin & destination DA
Occupied private dwellings	Number of private dwellings per unit sq. km of trip origin & destination DA

Trip purpose inference model estimation results

	Multinomial Logit	Nested Logit	Mixed Logit
LL-final	-7525.07	-7505.42	-7430.71
# of parameters	65	66	77
R-squared-bar	0.4158	0.4172	0.4221
AIC	15180.14	15142.84	15015.42
BIC	15290.94	15255.34	15146.67

Model estimation results: Land use variables



- Private dwellings in destination DA
- Manufacturing POIs in origin DA
- Educational POIs in origin DA



- Manufacturing POIs in destination DA
- Finance & insurance POIs
- Professional, scientific, & technical POIs
- Public administration POIs



- Educational POIs



- Private dwellings density in origin DA



- Retail trade POIs



- Accommodation and Food Services POIs



- Arts, Entertainment, and Recreation POIs



- Health Care and Social Assistance POIs



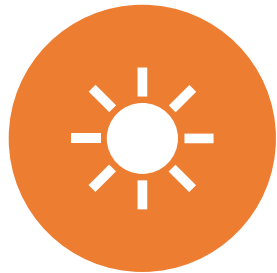
- Finance and Insurance POIs
- Other Services POIs



- Private dwellings density

Model estimation results: Trip start times

- Separate coefficients estimated for each time period to capture their specific effects on trip purpose



Morning trips are destined for some out-of-home activity location



Trips starting later in the day have lower probability of being work trip, and higher probability of being discretionary trip

Model estimation results: Day & Season

Weekday coefficients

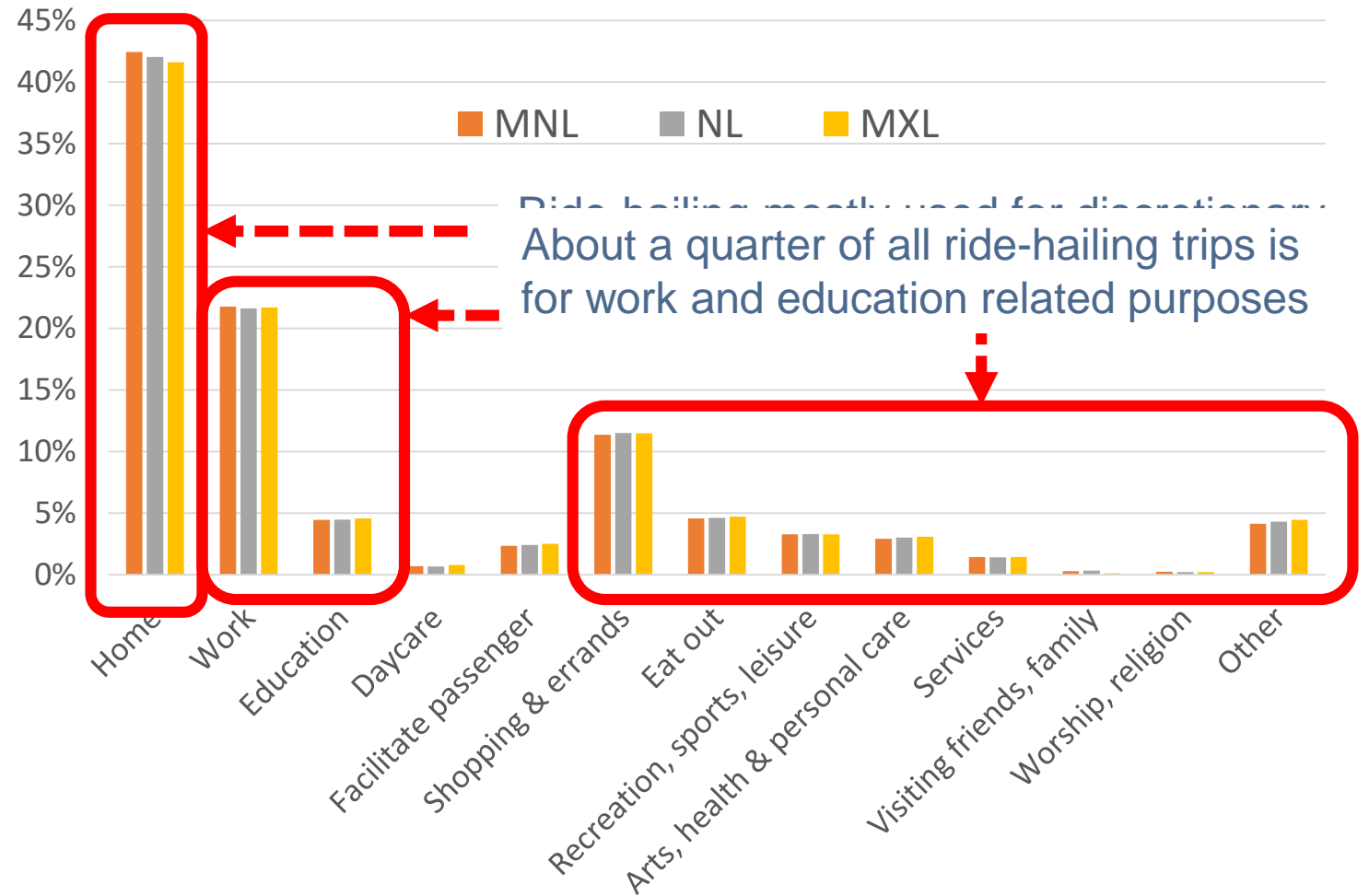
- +ve for work
- -ve for worship

Fall season coefficients

- +ve for education
- -ve for recreation and social visits

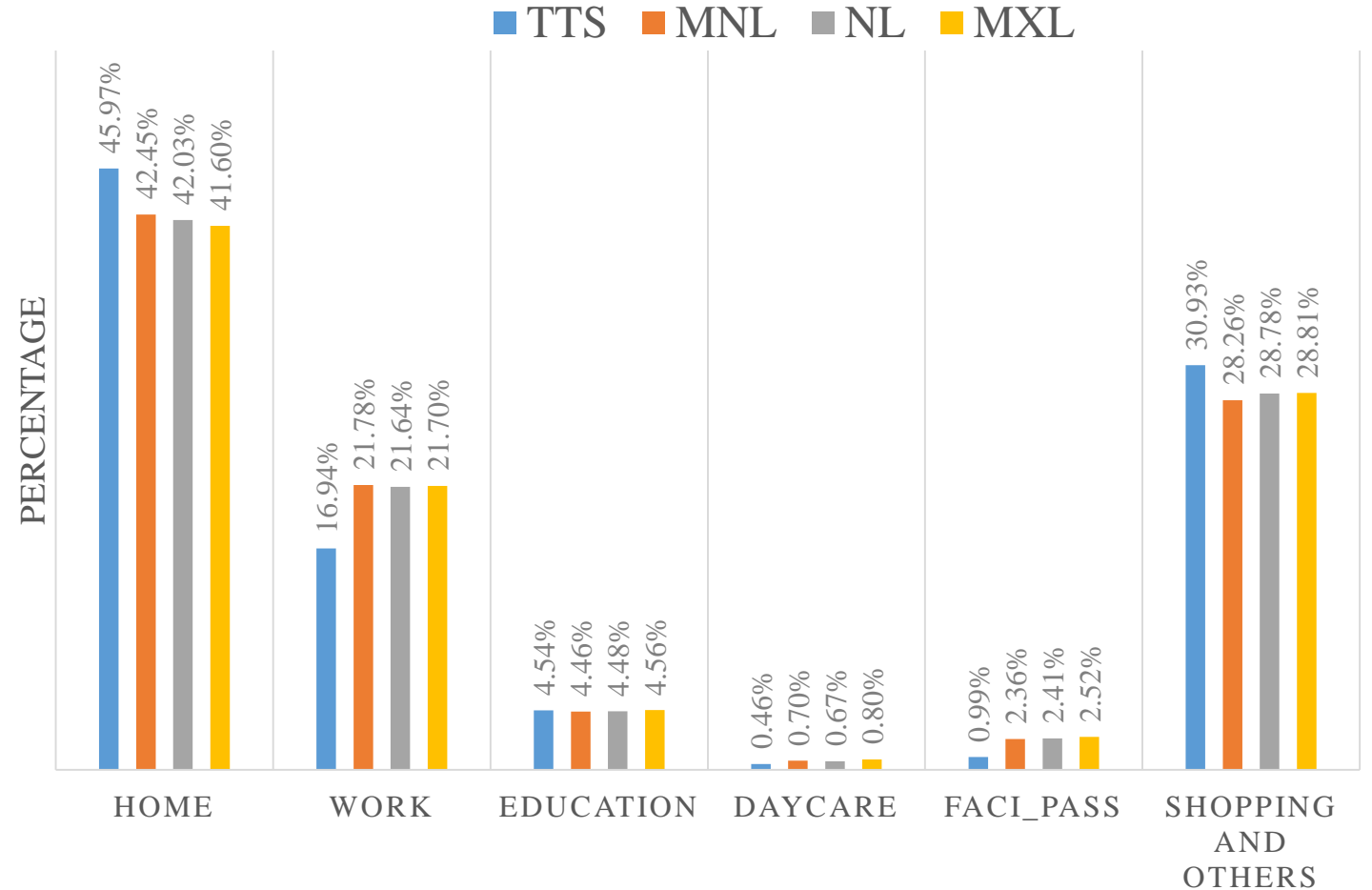
Inferring Ride-hailing Trip Purposes

- Estimated models applied to 20% of all ride-hailing trip trajectories within September and December 2016 augmented with land use information
- Generated the most probable purpose distributions for the 1,390,527 ride-hailing trips



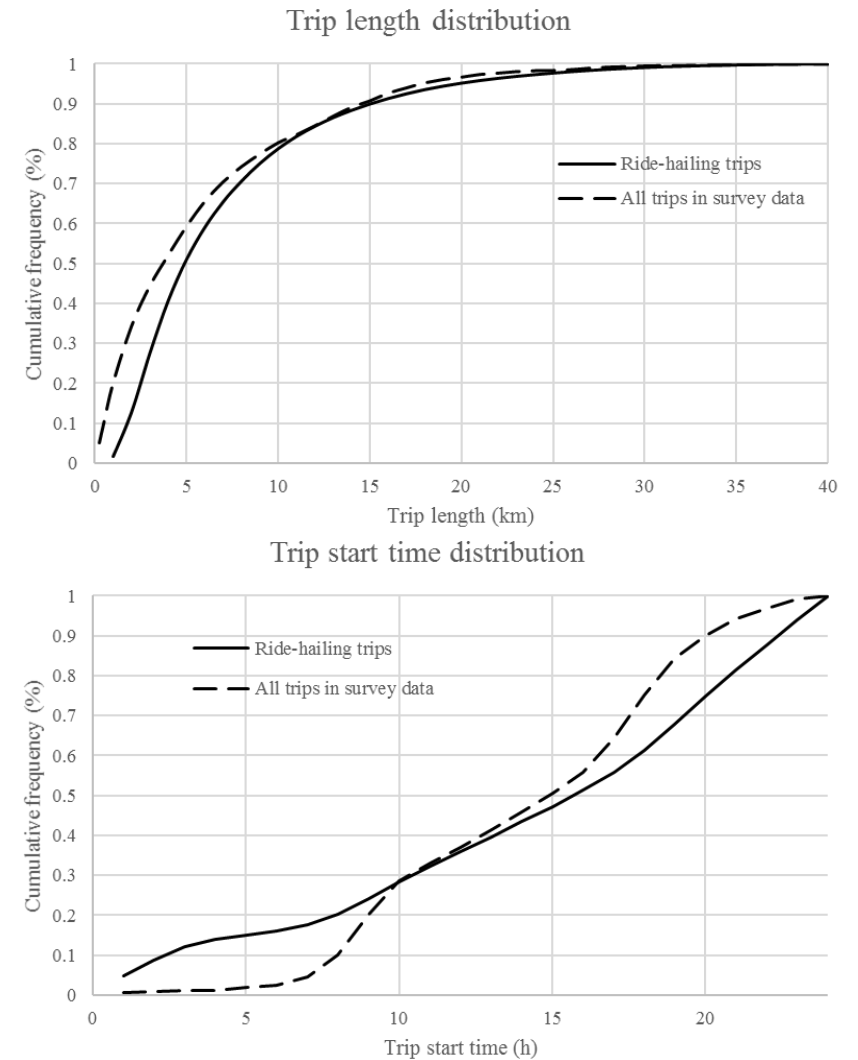
Validation

- Inferred weekday trip purposes are validated against TTS data
- Discretionary purposes are merged to make categories compatible



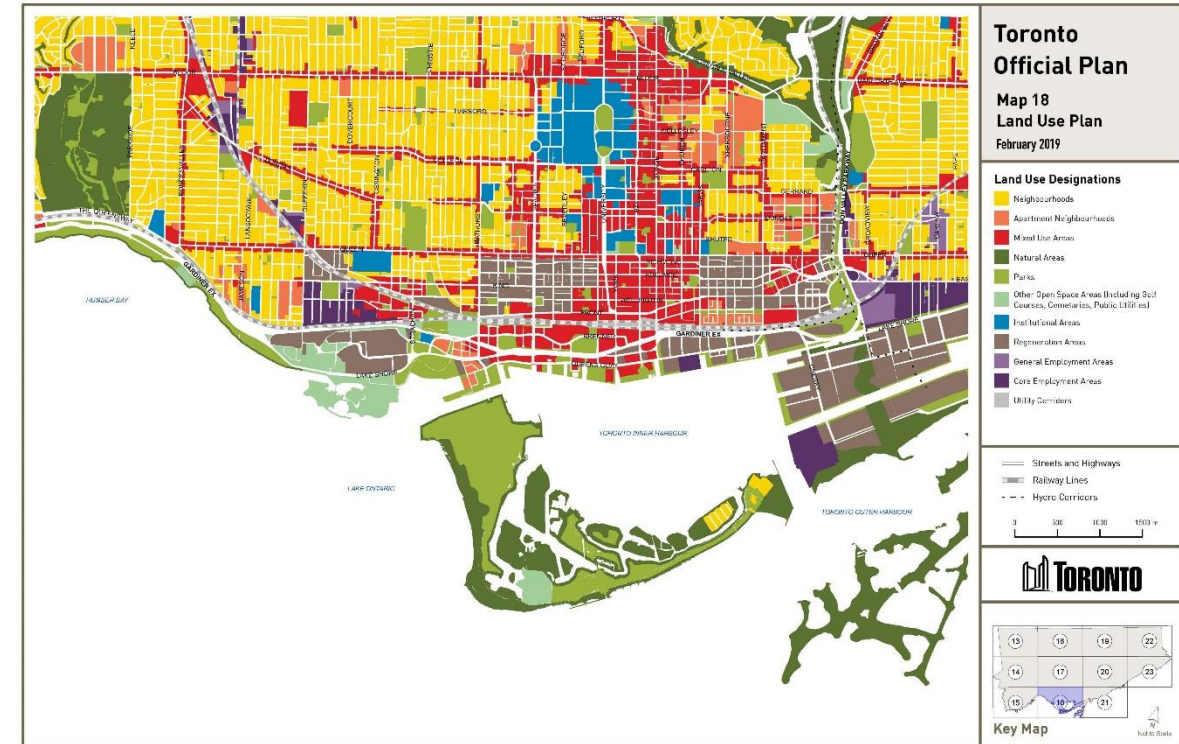
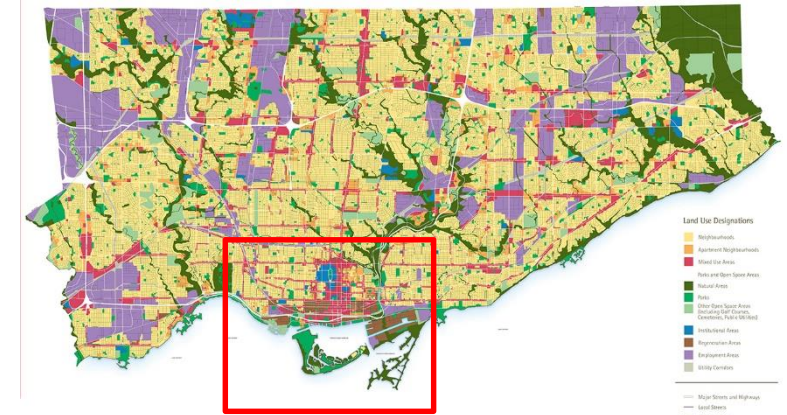
Validation

- Results are quite encouraging, given that
 - Trips in the estimation data have somewhat different spatial and temporal characteristics than the ride-hailing trip records

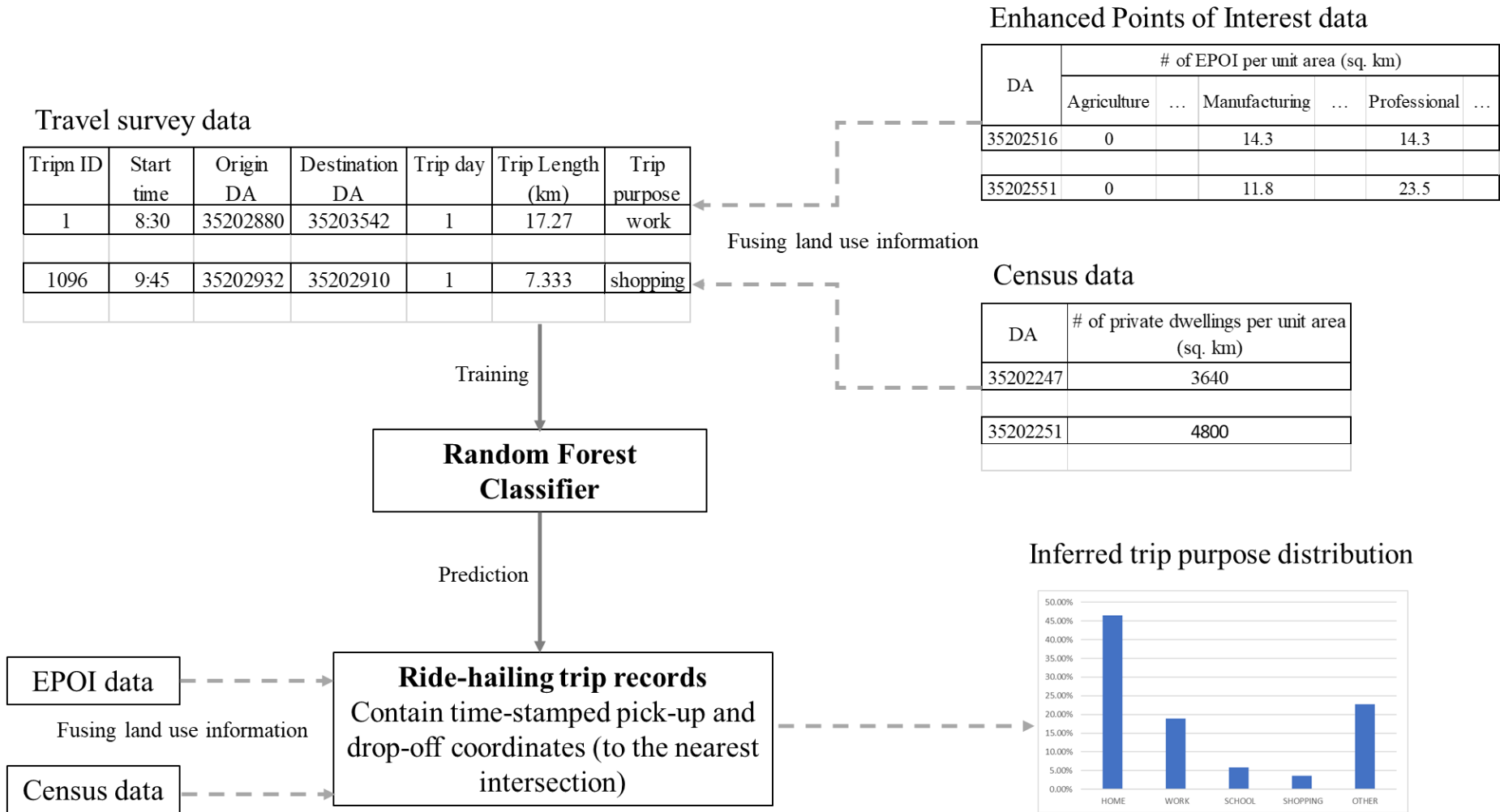


Validation

- Results are quite encouraging, given that
 - The study area has mixed-use land parcels, which has always been as a major challenge for trip purpose imputation



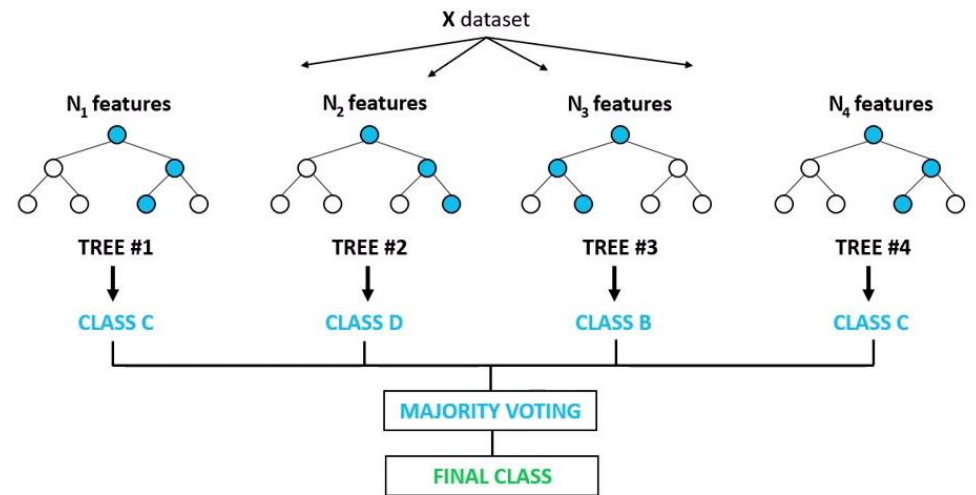
Purpose inference by Random Forest Classifier



Random Forest Classifier

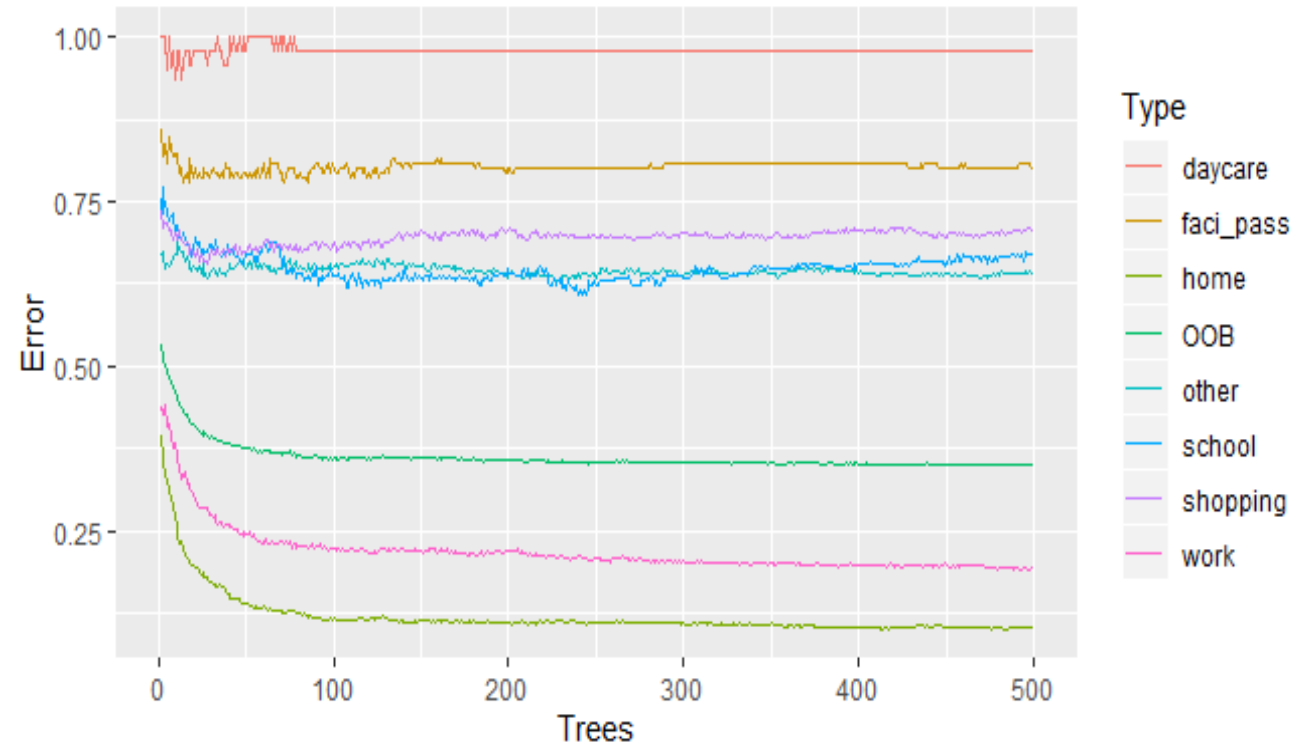
- An ensemble learning approach
- Predictions made based on votes from multiple decision tree structures
 - Random sampling of training data points when building trees
 - Random subsets of features considered when splitting nodes
- Less prone to errors in prediction due to overfitting compared to individual decision trees

Random Forest Classifier

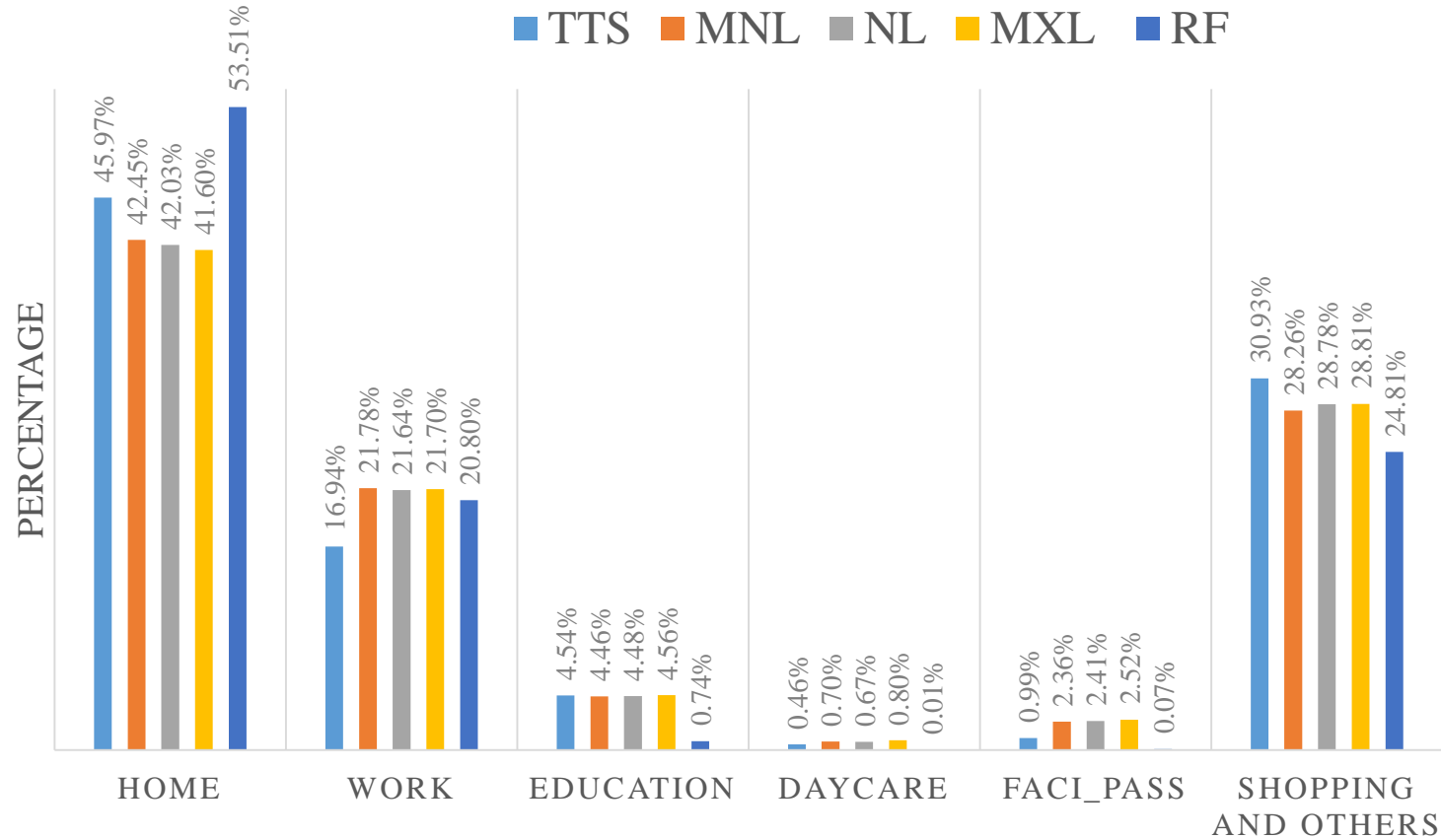


Training the Random Forest model

- Model was trained and tested for aggregated purposes
 - During training, 500 trees were grown for each forest with up to 7 input variables tried at each split
- The purpose categories with smaller shares have high prediction errors

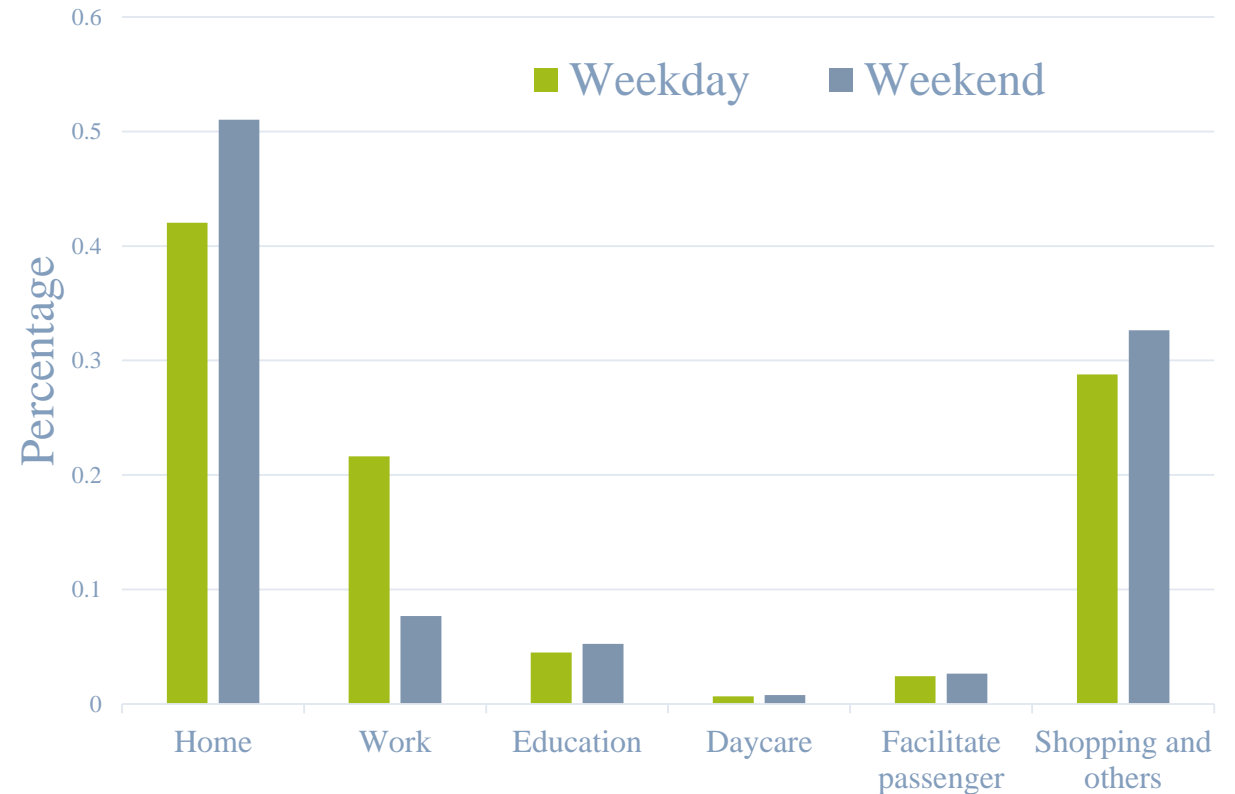


Comparing Predictions of Econometric models and Random Forest Classifier



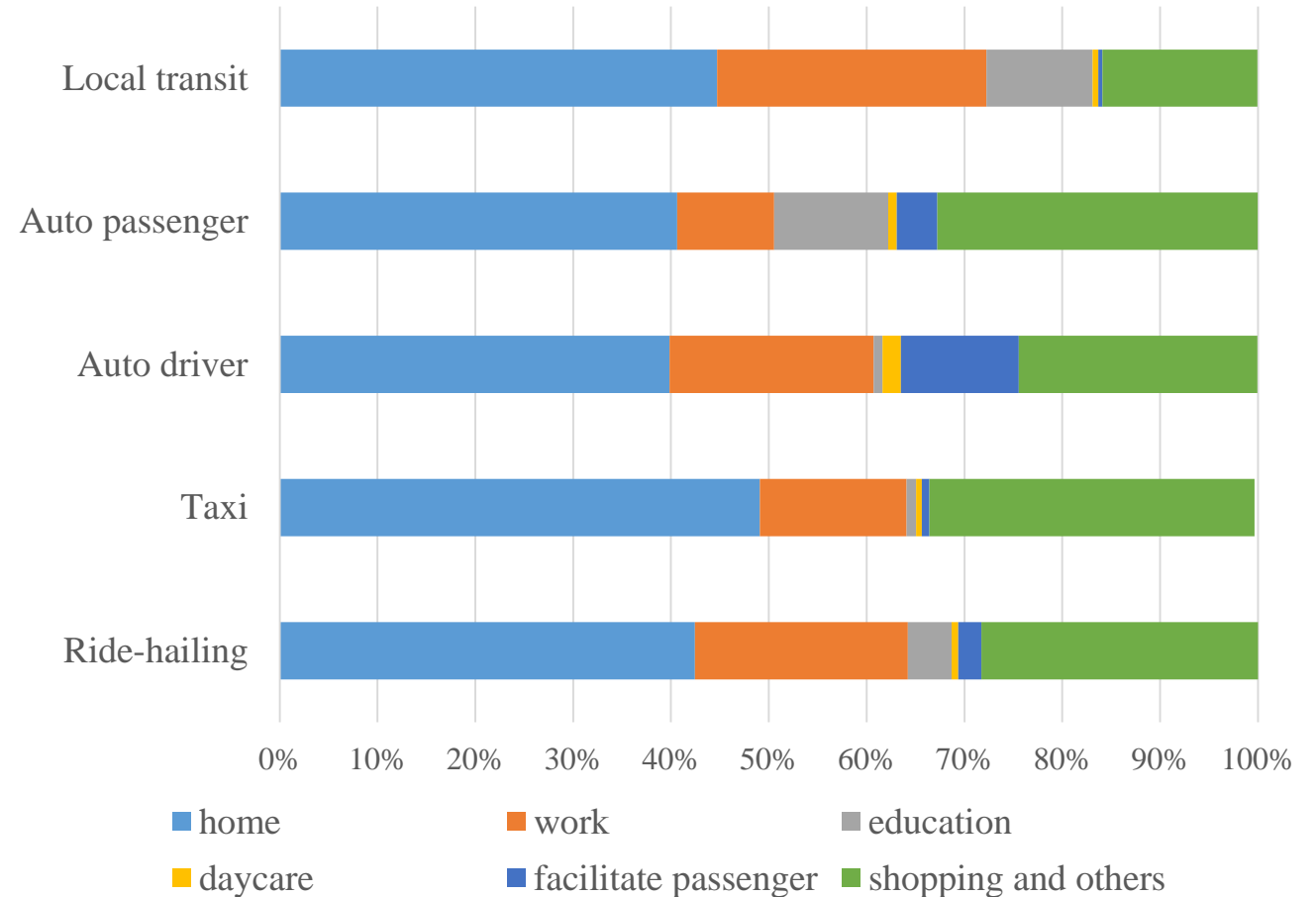
Characteristics of ride-hailing trip purposes

- Weekday vs weekend ride-hailing trips
- More 'return home' and 'shopping and others' trips are made by ride-hailing over the weekends



Characteristics of ride-hailing trip purposes

- Proportion of trip purposes for different travel modes
- Strong modal competition between taxi and ride-hailing
- ‘Work’ and ‘education’ constitute higher percentage of total ride-hailing trips than taxi



Limitations and Future Research

- Assumption: ride-hailing trips have the same conditional probability as the trips in the survey data.
 - What happens if ride-hailing is used to access transit?
- Improve prediction accuracy using social network check-in data, Google Places API, hours of operation of POI etc.

Key Findings & Conclusions

- Most probable trip purpose distribution inferred from ride-hailing trajectory data using limited context-specific variables
- Land use characteristics and trip start times are good contextual variables
- Ride-hailing is mostly used for discretionary activities and for returning home; it also plays an important role in daily commuter travel
- Efficient policies should be mandated to support the benefits of ride-hailing, but not at the expense of increased congestion and reduced transit ridership

Thank You

sanjana.hossain@mail.utoronto.ca