Modelling the taxi drivers' passenger pickup location strategies using GPS data

Asiri Senasinghe
Dr. Merkebe Demissie
Dr. Lina Kattan

May 18, 2018

Outline

1. Background
2. Problem formulation
3. Data development
4. Exploratory data analysis
5. Model estimation
6. Next steps

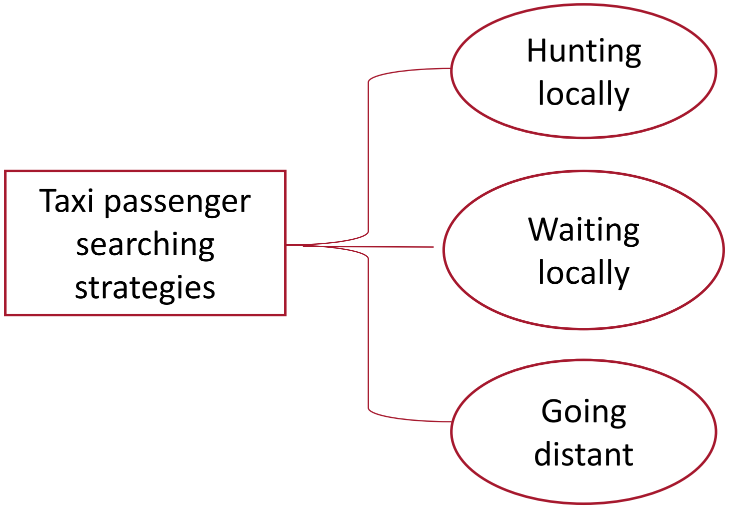
Background

- Taxi services are available in most of cities of the world, accounting for a small yet significant part of the daily trips;
- In recent years, the taxi industry is struggling to keep its market share;
- One of the reasons is the emergence of new transport alternatives (e.g.: Uber, Lyft, etc.);
- Transportation systems are increasingly being augmented with a range of information and communication technologies that make them smarter, safer, and more efficient;
- Taxi companies are implementing on-board GPS devices that produce large sets of data useful for strategic planning;
- Currently, there have not been enough efforts made to process the available GPS data to understand travel demand and this affects the ability to optimize the taxi service.

Background

One important issue for a taxi driver is reaching passengers;

- A driver with great experience will target specific taxi stands or street sections depending on the time of the day;
- A driver with less experience will wander on the street, empty, while looking for a client;
- If taxi drivers were more aware of travel demand and how it varies through space and time, they would make more trips with passengers and reduce their unproductive time and kilometers.

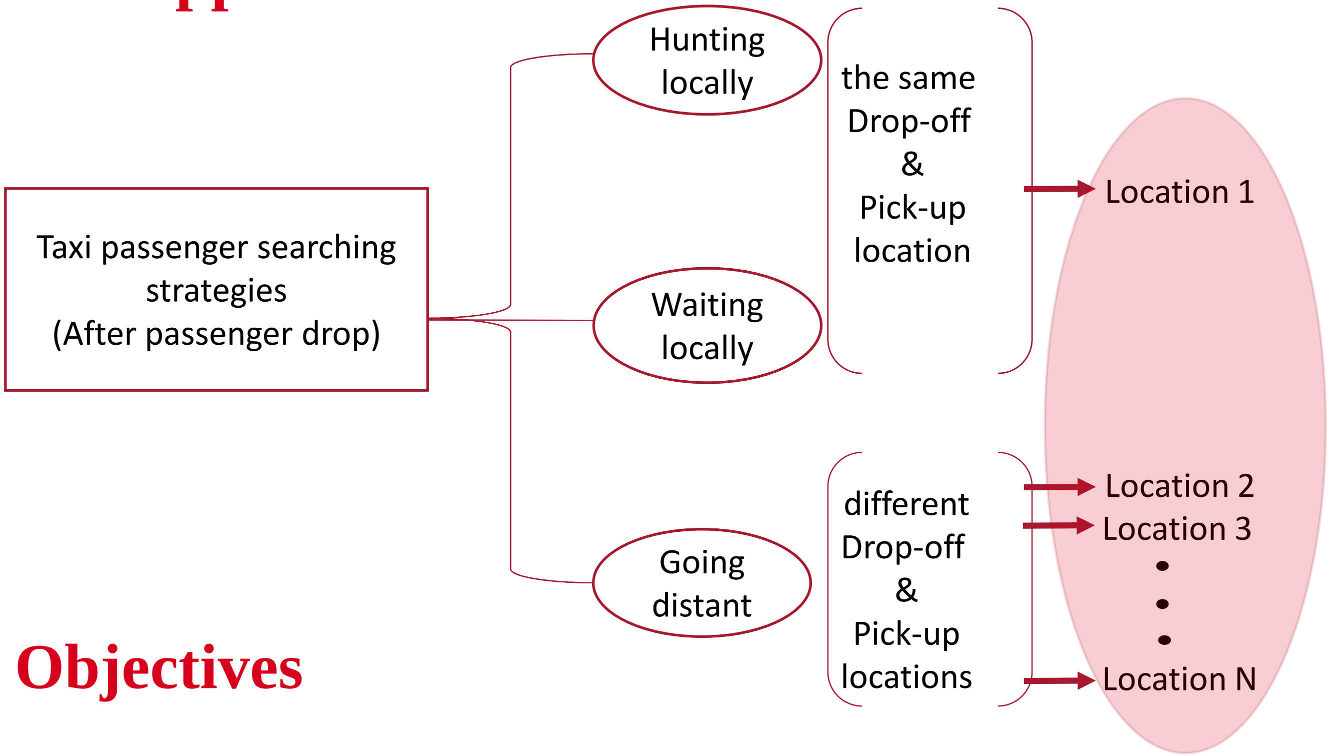


Literature Review

No	Study	Topic Addressed	Data	Method	Findings
1	Lacombe et al. (2015) Yang et al. (2014)	Taxi trip generation	Taxi GPS Census GTFS Weather	Multiple linear regression models	Daily and hourly trip generation models for pickup and drop-off events.
2	Moreira-Matias et al. (2012)	Short term taxi demand prediction	Taxi GPS	Time series forecasting techniques: Time varying poisson Weighted time varying poisson and Autoregressive integrated moving average models	Predicted the best taxi stand to go after a passenger drop-off in a given location and time.
3	Wen et al. (2013) Chang et al. (2010)	Taxi hotspots	Taxi GPS	Data mining: DBSCAN K-means Agglomerative hierarchical	Identified taxi hotspots for pickup and drop-off events at different weather conditions and time of day.
4	Yuan et al. (2011)	Recommender system	Taxi GPS	Data mining : OPTICS clustering Method Interactive Voting-based Map-Matching (IVMM) Probabilistic model	Developed a recommender system for both taxi drivers (Pickup locations) and passengers (vacant taxi).
5	Zhang et al. (2014) Zhang et al. (2011)	Taxi service strategies	Taxi GPS	Data mining: Support Vector Machine (SVM)	Identified efficient and inefficient taxi service strategies based on the generated revenue.

NOTE: Regarding passenger searching strategies, all reported works have used GPS data to study which factors affect the taxi driver mobility intelligence and consequently the choice about the best route and pickup location.

Our approach



Objectives

- Explore the feasibility of developing the taxi drivers' passenger pickup location strategies with discrete choice model structure.
- Besides impendence and size variables, test the explicit inclusion of socio-economic, geographic and other drivers of taxi demand variables in the utility function.

Destination Choice Model: MNL Model formulation

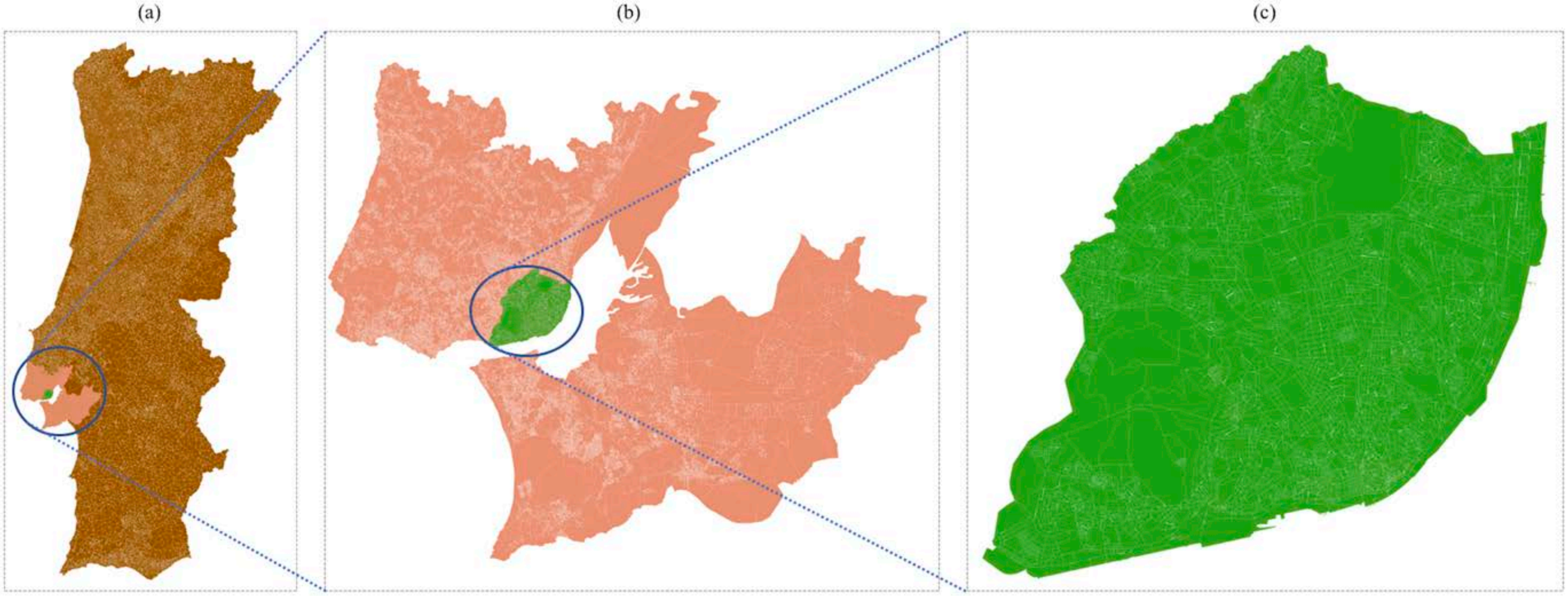
- The most common implementation of destination choice is the multinomial logit form;
- The trip makers are hypothesized to choose the destination that maximizes their utility;
- The utility of a destination is a function of trip maker preferences, the attractiveness of the destination zone, and other unknown, un-included attributes of the trip maker or the destination zone.
- The probability that trip produced in zone i chooses destination zone j is given by the utility of zone j and the utility of all other possible destinations (P_{ij}):

$$P_{ij} = \frac{e^{V_{ij}}}{\sum_{k=1}^{N} e^{V_{ik}}} \qquad V_{ij} = \alpha t_{ij} + \beta \ln\left(S_j\right) + \gamma Z_{ij}$$

Where,

V_{ij} = measurable utility of TAZ j as a destination perceived by trip from origin TAZ i;
N = total number of TAZs in the study area;
t_{ij}= travel time/distance/ or other form of impendence variable from the origin TAZ i to destination TAZ j;
S_j = size variable for destination TAZ j;
Z_{ij} = other explanatory variables (e.g.: dummy variables);
α, β, γ = coefficients for the corresponding explanatory variables.

Case study region: Lisbon (which is the capital and the largest city of Portugal)



Portugal | Lisbon metropolitan area | Lisbon

Acquired Data

GeoTaxi	Statistics Portugal	SAPO (Internet services Company)	Foursquare	Google
Taxi GPS data: - Location (lat, long) - Taxi status (free, occupied) - Taxi ID - Time and Date	Census data: - Population number - Education - Employment - Shapefiles	Point of Interest (POI): - Services - Education - Health - Recreation - Shopping	Data on where and when users check-in: - Venue name - Venue category - Geo-referenced location - Number of unique visitors - Number of total check-ins	Google distance matrix API: - Travel time - Travel distance

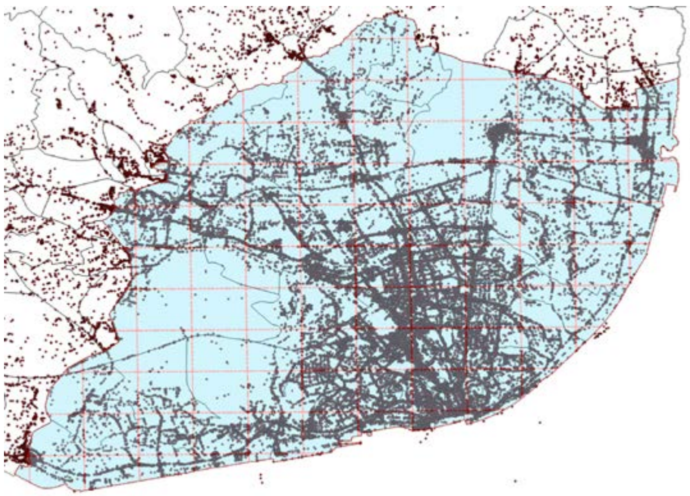
GeoTaxi GPS Data

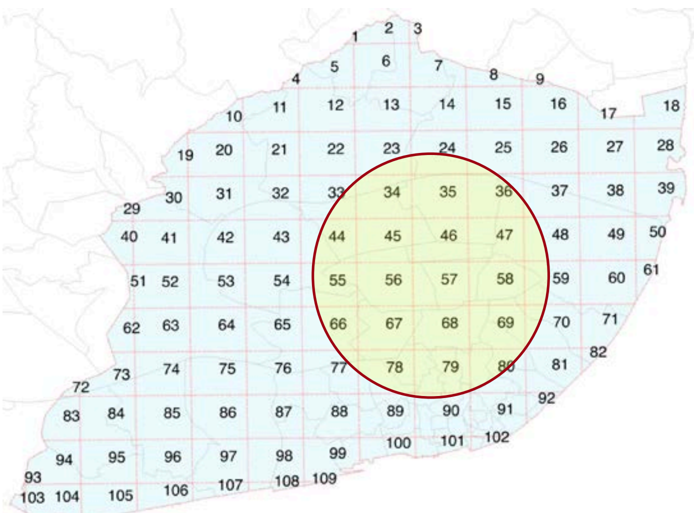
Attributes	Description
Spatial coverage	Pop= 552,700; Area=100.05km.sq.
Temporal coverage	September 2009
	October 2009
# of records	2.83 million
# of unique taxi IDs	253
Market share	15% in 2009

Important Point of Interests (POIs):

25: Lisbon Portela Airport
28: Oriente train and bus terminal
55: Sete Rios train and bus terminal
90: Rossio train terminal
92: Santa Apolonia train terminal
101: Cais do Sodre train terminal

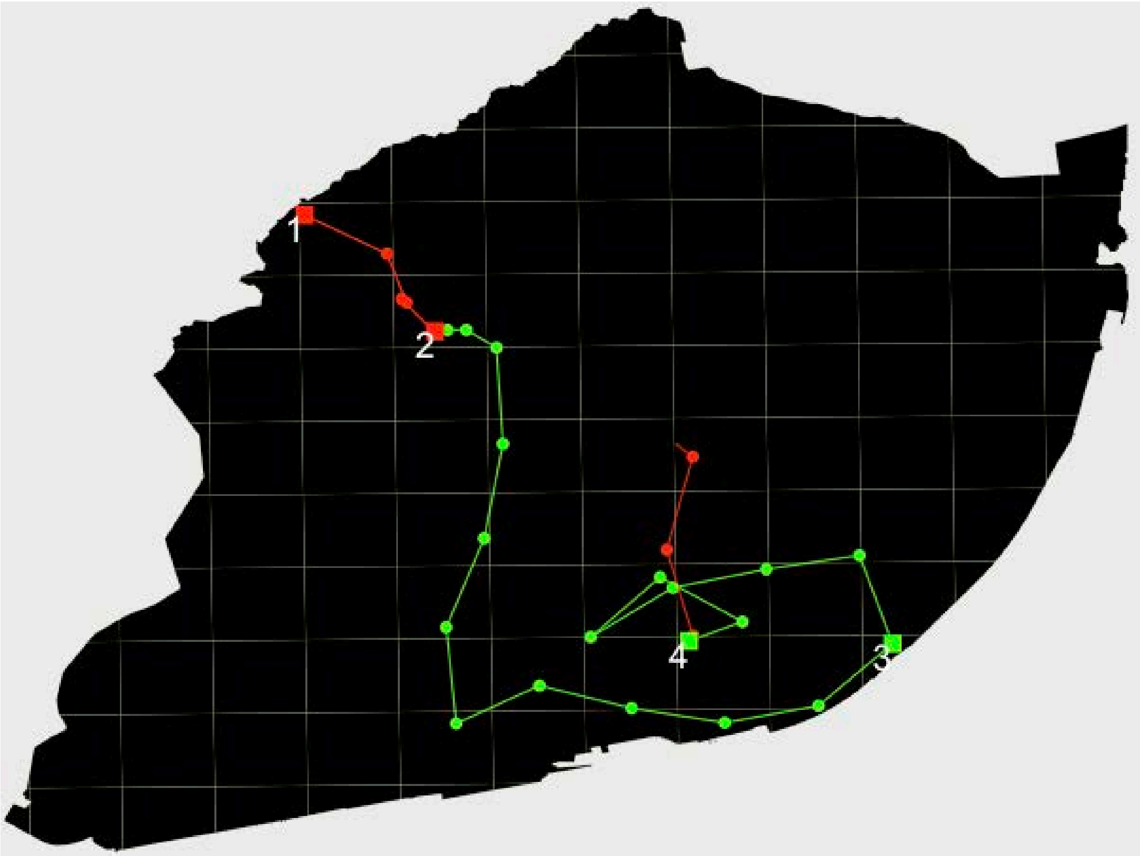
Major transport hubs





Central Business District (CBD)

Sample Taxi GPS traces



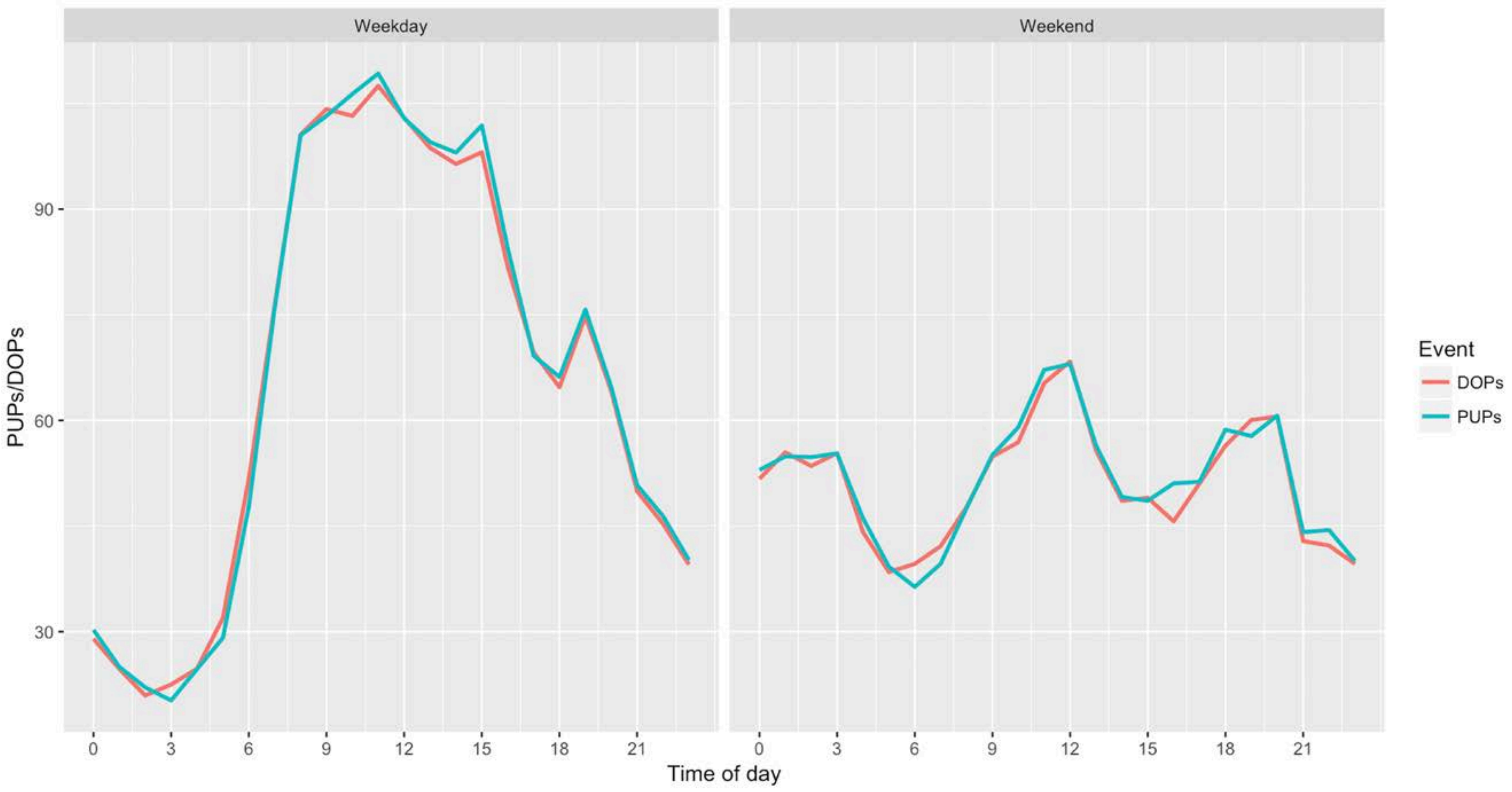
1: Taxi in occupied status
2: Passenger drop-off event
3: Taxi stopped (waited?) for 23min
(unsuccessful passenger pick-up attempt)
4: Passenger pick-up event
(successful passenger pick-up attempt)

Red segment: passenger delivery
Green segment: passenger searching
Solid circles: GPS records

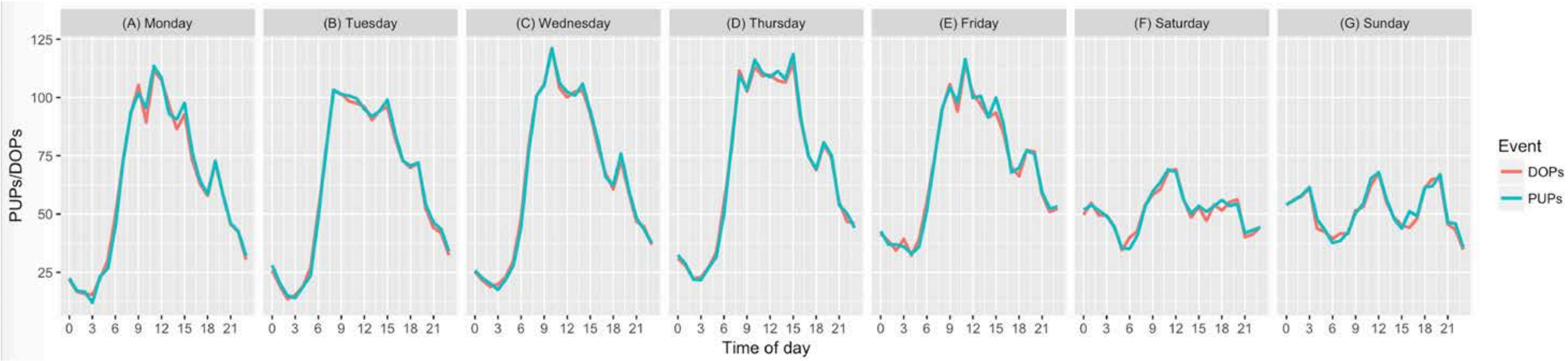
The Google Maps Distance Matrix API

- Provides travel distance and time for a matrix of origins and destinations;
- Max. of 2,500 free elements per day (O*D);
- Maximum of 25 origins or 25 destinations per request;
- Maximum 100 elements per request.

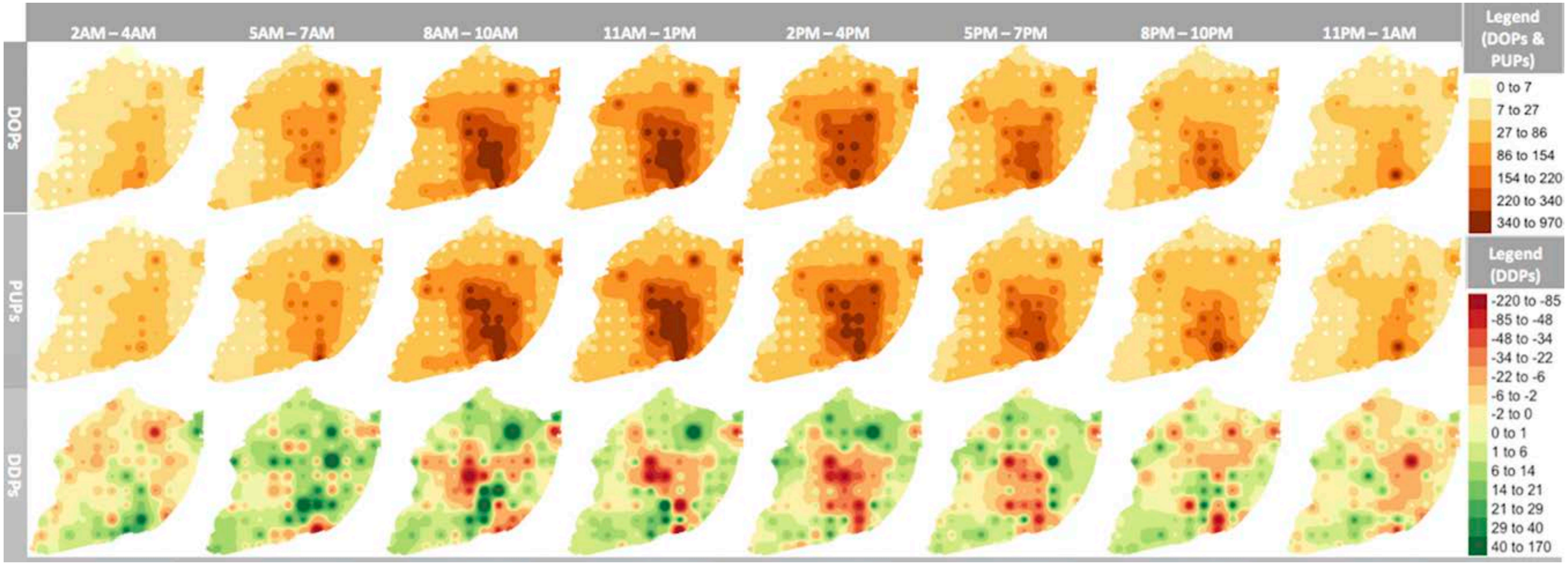
Citywide weekday and weekend average Drop-off (DOPs) and Pick-up (PUPs) patterns



Citywide daily average Drop-off (DOPs) and Pick-up (PUPs) patterns



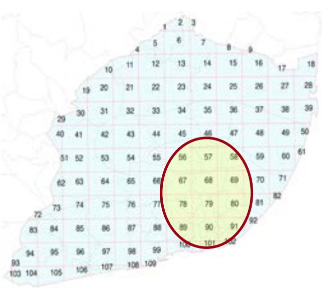
Citywide average weekday patterns: Drop-off (DOPs); Pick-up (PUPs); and DDPs = DOPs- PUPs



DDPs = DOPs- PUPs

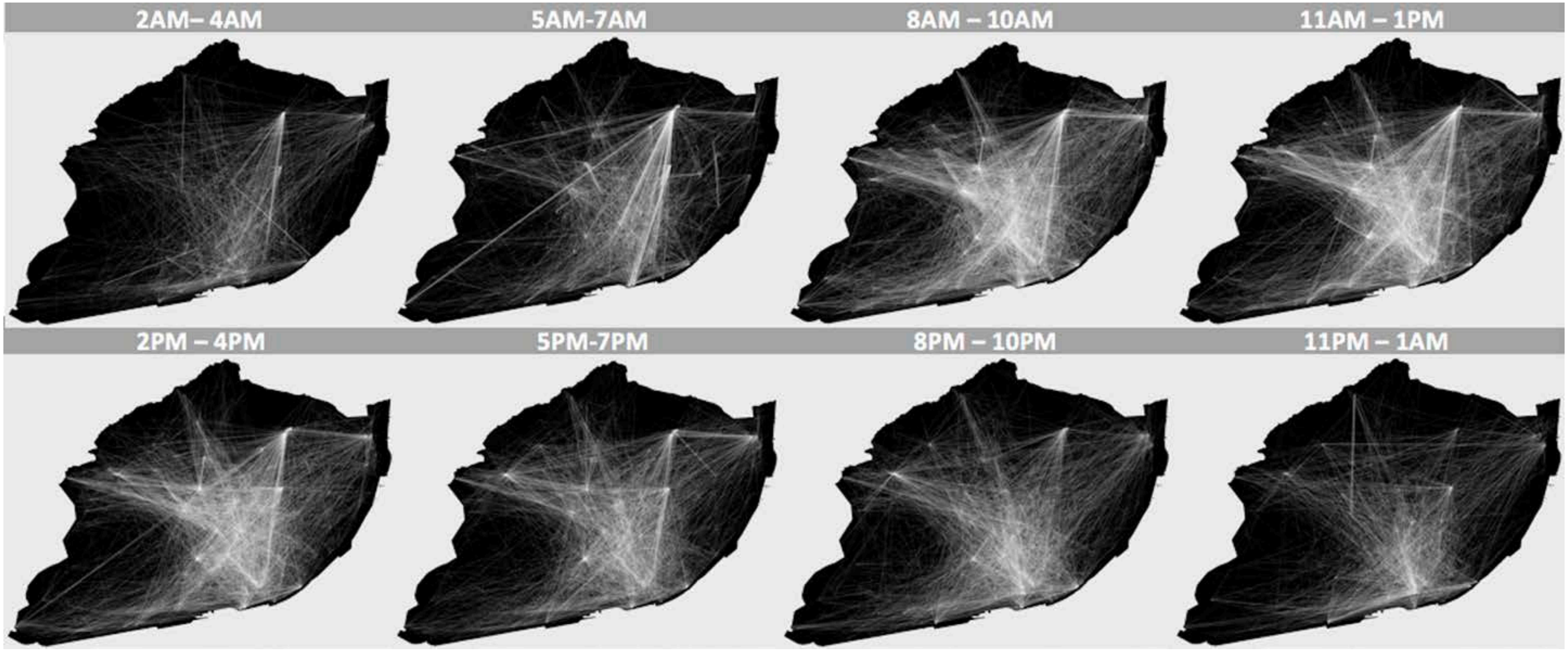
Red DDPs = more taxi passenger departures
Green DDPs = more taxi passenger arrivals

Important Point of Interests (POIs):
25: Lisbon Portela Airport
28: Oriente train and bus terminal
55: Sete Rios train and bus terminal
90: Rossio train terminal
92: Santa Apolonia train terminal
101: Cais do Sodre train terminal

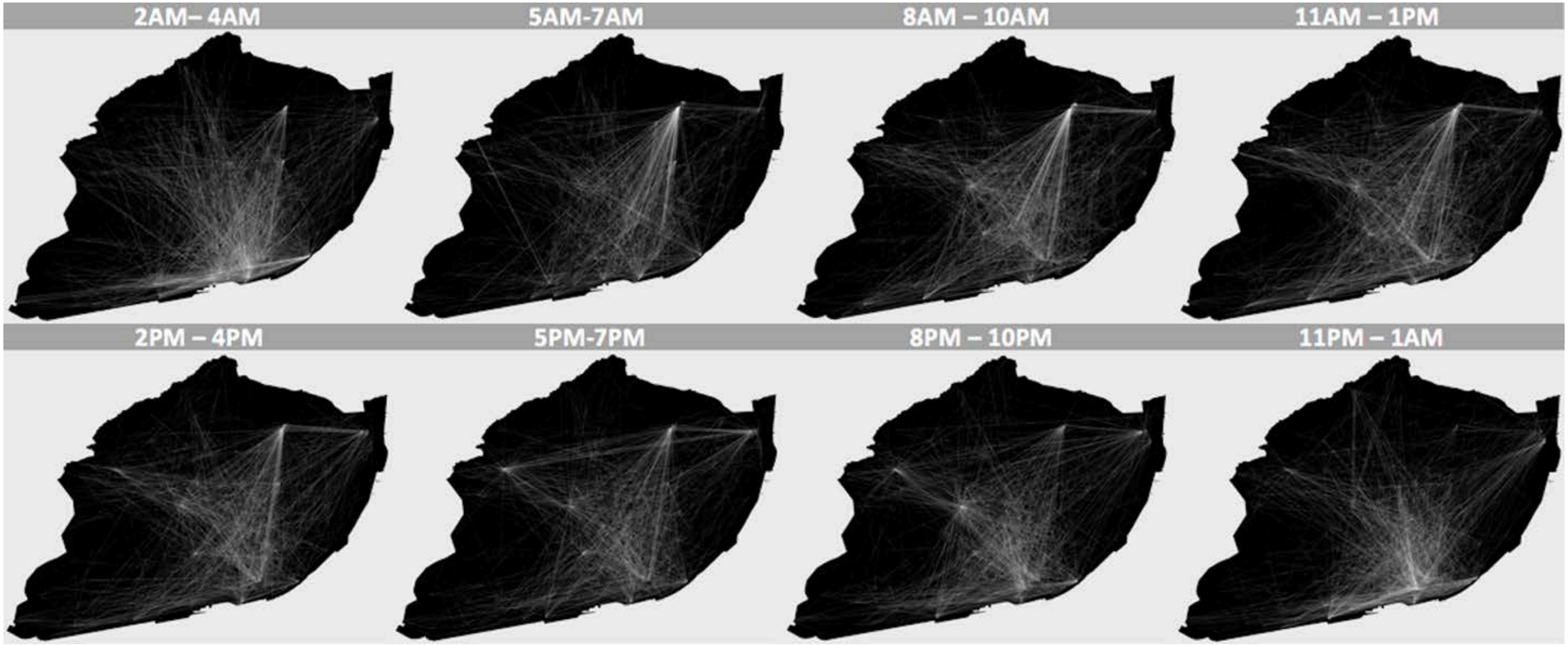


Central Business District (CBD)

Average weekday Drop-off (DOPs)-to-Pick-up (PUPs) patterns



Average weekend Drop-off (DOPs)-to-Pick-up (PUPs) patterns



Utility Function:

$$
\begin{aligned}
V_{ij} = \alpha_1 tt_{ij} + \alpha_2 td_{ij} + \alpha_3 ttr_{ij} + \alpha_4 tdr_{ij} + \alpha_5 wt_j + \beta \ln(S_j) \\
&+\gamma_1 dummy_{high\_employment} \\
&+\gamma_2 dummy_{medium\_employment} \\
&+ \gamma_3 dummy_{high\_education} \\
&+ \gamma_4 dummy_{medium\_education} \\
&+\gamma_5 dummy_{high\_hotspot} \\
&+\gamma_6 dummy_{medium\_hotspot} \\
&+\gamma_7 dummy_{high\_people\_presence} \\
&+\gamma_8 dummy_{medium\_people\_presence} \\
&+\gamma_9 dummy_{transport\_hubs}
\end{aligned}
$$

Employment and education dummies indicating portion of the people who are Educated and employed

Presence of people and hotspot dummies
Indicating the attractiveness of pickup areas

Transport dummy representing major Transport hubs (air port, train, and bus)

Where,

- wt_j- taxi drivers' wait time; tt_{ij} - is travel time; td_{ij}- trip length; ttr_{ij} and tdr_{ij} are individual specific variables representing the ratio of travel time/trip length to average travel time/trip length to a taxi driver usual pickup location;
- The $low\_employment, low\_education, low\_hotspot, and\ low\_people\_presence$ are used as the reference and model results obtained for the other combinations should be interpreted relative to the reference.

Estimated destination choice (pickup location) models for taxi movements

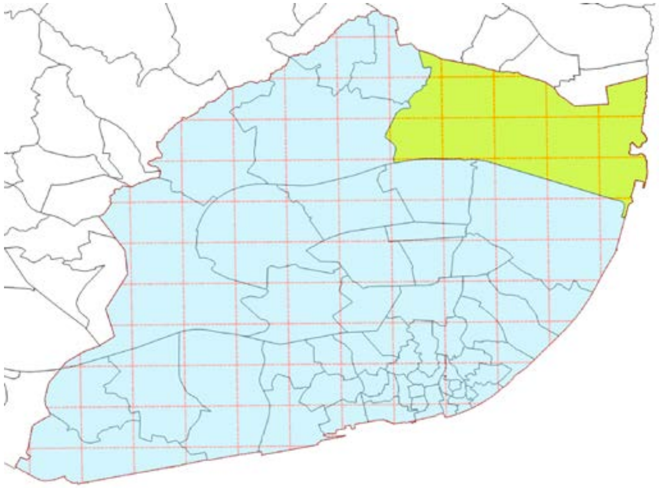
Variable	Models (Weekday)			
	Model 1		Model 2	
	Coefficient	P-value	Coefficient	P-value
S_j $(weekday)$	0.1975	< 0.001	0.2067	< 0.001
travel time (tt)	-0.1225	< 0.001	- -	- -
trip length (td)	-0.2921	< 0.001	- -	- -
travel time ratio (ttr)	- -		-1.9175	< 0.001
trip length ratio (tdr)	- -		-2.3561	< 0.001
high employment	-0.0869	< 0.001	-0.0971	< 0.001
medium_employment	0.0397	0.004	0.0326	0.017
high_hotspot	1.3814	< 0.001	1.3915	< 0.001
medium_hotspot	0.7801	< 0.001	0.7817	< 0.001
high_people_presence	0.6511	< 0.001	0.6429	< 0.001
medium_people_presence	0.3912	< 0.001	0.3853	< 0.001
transport_hubs	0.4324	< 0.001	0.4417	< 0.001
initial log-likelihood	-258280		-258280	
final log-likelihood	-202500		-201180	
Rho-squared (ρ^2)	0.2160		0.2211	

- Models underestimated long pickup trips;
- The estimated coefficients for the education dummy variables are not significantly different from zero. (t-statistics are less than 1.96, which is the critical level for 95% confidence level);
- The estimated coefficient for the taxi driver's wait time variable (at the pickup location) does not have the expected sign (estimated coefficients have +ve sign).

Next Step

Proper delineation of traffic analysis zones (Choice set formation)

- The 1km by 1km grid size was considered to establish the traffic analysis zone (TAZ);
- One of the assumptions is that a MNL model can only be applied to situations in which alternatives are totally independent (mutually exclusive alternatives);
- The 1km by 1km TAZs may not satisfy this assumption, thus use of a simple MNL model seems inappropriate;
- Previous studies suggested a nested logit formulation with two levels of decision for destination (Suarez et al., 2004).
- This means that each taxi driver is assumed to first choose an urban level (e.g. a small administrative district (freguesias)) and then, within that broad spatial zone, to choose a precise pick-up location (1km by 1km zone).



Thank you!

CONTACTS:

asiri.senasinghe2@ucalgary.ca
merkebe.demissie@ucalgary.ca
lkattan@ucalgary.ca