Analysis of taxi drivers' behavior with GPS data from a megacity

Extended Abstract

Mikhail Murashkin, Nikolas Geroliminis

Urban Transport Systems Laboratory (LUTS)

Ecole Polytechnique Federale de Lausanne (EPFL)

In large urban networks, taxis are considered an efficient mode of transport that transfers travelers from their preferred origin to destination points. Taxis are spending a significant part of their service to cruise for passengers in urban networks and significantly contribute to congestion. Developing a model to describe how vacant and occupied taxis will cruise in a road network to search for customers and provide transportation services is a challenging research question. Investigating how information services can facilitate the decision making of taxi behavior while searching for customers can influence the quality of service and the associated costs.

During taxi services, once a customer is served and his trip is completed at a destination point, the taxi driver could either stay in the same zone or move to other zones (possibly of higher demand) to search for a new customer. Previous models attempt to identify equilibrium conditions where drivers try to minimize expected search time to meet the next customer (see for example Yang and Wong, 1998). The same paper also investigates the minimum taxi fleet required to ensure the existence of a stationary equilibrium state. Most of the models in the literature integrate concepts of traffic assignment literature or gravity-type models to identify equilibrium conditions for small size networks (see for example Wong et al, 2005, Yang et al. 2002 and others) or investigate the economic consequences of regulations as price control and location restriction (e.g. Cairns and Liston-Heyes, 1996; Arnott, 1996, Yang et al., 2005a). Nevertheless, there is not significant body of literature to develop and validate with real data from large networks the behavior of taxis both during equilibrium and transient conditions. A detailed literature review will be provided in the full version of the paper.

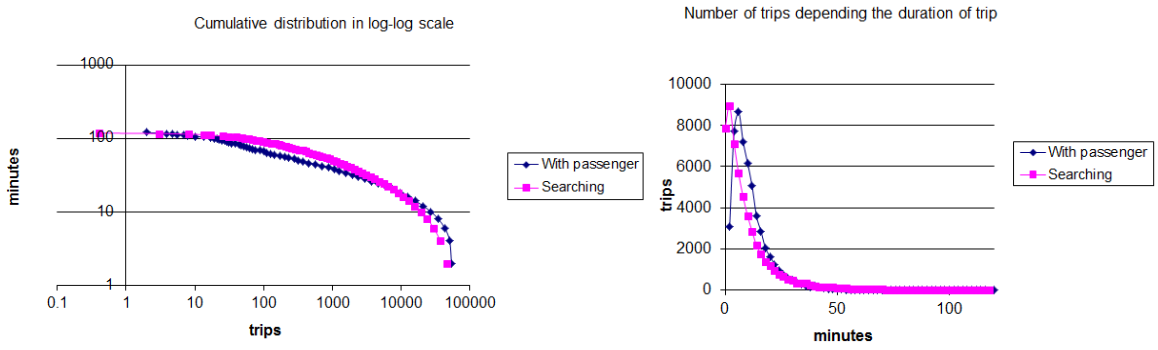
In this paper we analyze a unique dataset from Shenzhen – a fast growing megacity. We utilize GPS data from a large dataset of taxis to identify mobility patterns of taxis and identify aggregated strategies that describe the behavior of taxis when searching for passengers. The data contains the network structure and daily GPS data of taxis for multiple days in the city of Shenzhen in China. Shenzhen is a major city in the south of Southern China's Guangdong Province, situated immediately north of Hong Kong. The rapid foreign investment created one of the fastest-growing cities in the world with a population close to 11 million and, as expected large congestion problems both in the urban and freeway system of the city. Each day, the taxis make records of current time, GPS locations in form of coordinates and passenger information (if the taxi carries a passenger or not). The time of recording is random and the length of interval between two records ranges from 1 second to several hours. In total there are around 30 to 50 million records per day. Data processing is the first and critical step of data

analysis. It plays an essential role by extracting useful information and facilitating the work of later stages in observing the underling patterns and deriving insightful results. Thus the final messages inferred and phenomena revealed are directly related to the quality of data and the effectiveness of the processing methods.

This data gives us an ability to investigate not only the traffic conditions on the roads (Ji and Geroliminis, 2012), but also the properties of trips with passenger and searching trips for passengers. The searching trip starts from the point where a passenger alights and ends in the point where a new passenger boards. However, our analysis should ensure that during this period taxi is in a working shift (e.g. not parked). To extract the GPS points that correspond to the searching trips only, we developed a model which estimates the activity of the taxi in a particular time period. Moreover, this model tries to through away GPS tracks which contain apparent errors.

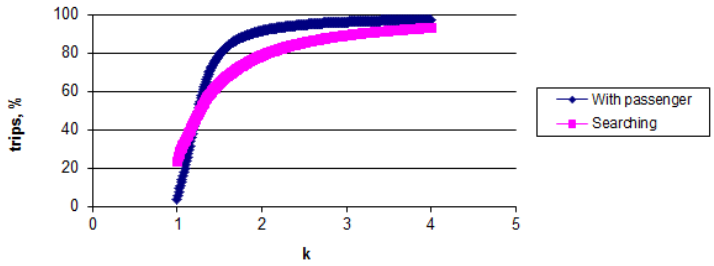
After cleaning the data we investigate the distributions of duration of both types of trips (with passengers and searching) and the circuity of trips for each type. We first observe (Figure 1a) that the distributions of trip duration are similar to a power-law for a range of the distribution. It is very common for big complex systems (there are a lot of examples in Shiode and Batty, 2000). It means that our cleaning model is rather reliable. Since this time we will be focused on the time interval from 8am to 10am (morning peak hour). One of main findings is that searching trips and trips with passenger are very different. For example, the average length of searching trip is 3.0 km and the average length of trip with passenger is 5.0 km, see also Figure 1b (duration of trip). Furthermore, there is a big difference in ratio k= L/S, where L is a trip length and S is distance between the first and last points of the trip (Figure 2). More explanation of the data analysis will be provided in the full paper.

Figure 1 Distribution of duration of trips in the time interval from 8am to 10am



(a) CDF in log-log scale (b) Number of trips (time resolution = 2 minutes)

Figure 2 Cumulative distribution of k



The average value of k for searching trips is 1.77 and for trips with passenger is 1.35, which means that searching trips are less straight than trips with passenger. One of possible explanations of this fact is that after the passenger alights, taxi driver starts to move in some direction, but in order to increase probability to find the new passenger, he tries to visit more roads and therefore his path is less straight.

To investigate the behavior of searching taxis we decided to quantify spatial distributions of different parameters of trips (e.g. length, duration, number of origins, number of destinations…) and analyze them. Given the large size of the studied network, we need first to divide the map of Shenzhen into several regions. Our idea is to do it the way that each region has an area that "attracts" searching taxis which are currently in this region. First, we divided the city into 1 x 1 km squares. Then a clustering algorithm is developed to identify the main regions of attraction. While more details of the algorithm will be presented later, the logic is as follows:

1. For each square we look at 24 neighboring squares (which lie in the 5x5 km square with the chosen square as a center) and calculate number of searching trips that start in this square and end in each of the other 24. Then we draw an arrow from our square to one of these 24 squares which attracts more trips. Finally, these arrows give an oriented graph. Each component of this graph can be considered as a cluster.
2. To keep the number of clusters smaller (that can be utilized later to develop an elegant model) we merge some of the ones created in the previous step. To do this we define an attractor as a vertex of the graph which belongs to the cycle. If we start to move from any vertex by arrows then we will finally reach one of the attractors of the cluster. If two attractors from different clusters belong to two squares with the common side then we merge these two clusters in one big cluster.

This algorithm succeeds in dividing the large network in a small number of clusters (12 in total, see Figure 3). We can now estimate O-D matrices for searching trips and trips with passenger (Figure 4). The rest of the territory was called cluster 0. We will look at all trips excluding trips inside cluster 0 because there are very few trips connecting cluster 0 with clusters 1-12 and we can say that clusters 1-12 are almost isolated. An interesting result is that even at this aggregated scale taxis with and without passengers have a completely different behavior. About 77% of searching trips are internal (lie on the diagonal of the matrix) and only 44% of trips with passenger are internal. It means that we cannot claim that taxi drivers usually return to the place where they took a passenger. This behavior is more complex, and should be further investigated, here we can remember the fact that ratio k for searching trips is higher than for trips with passenger.

In the second part of paper we develop a general approach for modelling the system in an aggregated manner. We try to describe the behavior of taxis in some region in terms of portions of them that make different decisions. We imply that all of them maximize or minimize their objective function (e.g. searching time) – this is a kind of equilibrium model. To simplify the problem we start with a static case, but our main goal is to understand how behavior of taxi drivers depends on changing number of searching taxis and changing traffic conditions. Having such a model we can develop an understanding of possible usage of GPS taxi data for real-time traffic control.

Figure 3 Results of clustering. Big blue points correspond to attractors, color means the number of destinations of searching taxis.

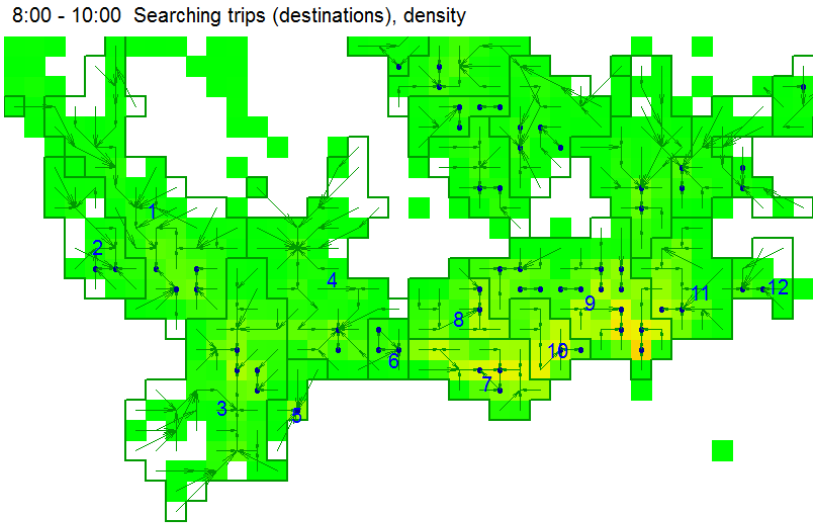


Figure 4 O-D matrices

Destinations

Origins	0	1	2	3	4	5	6	7	8	9	10	11	12	Gran..
0		143	16	18	23	0	2	13	5	70	6	41	20	358
1	30	2002	125	66	16	1	1	2	0	3	0	1	0	2246
2	5	81	246	2	0	0	0	0	0	1	0	0	0	335
3	5	26	2	2948	294	55	16	60	11	4	4	2	0	3428
4	8	11	1	332	1176	76	64	63	35	16	2	4	3	1789
5	0	1	0	32	11	206	1	6	0	2	1	1	0	260
6	1	0	0	39	96	7	165	66	26	5	5	1	1	411
7	6	12	1	32	59	3	55	4394	522	164	607	12	4	5871
8	1	1	1	5	34	0	17	390	1338	193	126	2	2	2110
9	67	5	0	6	17	0	5	129	167	7845	498	737	19	9494
10	23	0	0	5	9	0	2	310	97	478	2231	17	5	3176
11	50	0	0	3	0	0	0	9	3	270	3	1618	32	1988
12	0	0	0	0	0	0	0	1	0	4	0	22	155	184
Grand Total	196	2281	392	3488	1735	350	326	5444	2203	9055	3483	2459	241	31652

(a) Searching trips

Destinations

Origins	0	1	2	3	4	5	6	7	8	9	10	11	12	Gran..
0		53	4	43	37	2	6	89	43	441	106	154	9	987
1	81	1630	192	298	93	20	8	27	13	24	10	0	0	2395
2	19	261	102	18	5	1	0	1	0	0	0	0	0	408
3	72	166	9	2111	529	126	41	216	48	100	35	10	0	3464
4	72	45	4	421	574	44	92	186	80	86	53	4	0	1661
5	9	28	1	135	72	19	9	34	4	9	15	0	0	335
6	11	7	0	39	63	6	48	68	33	25	16	3	0	319
7	141	36	0	207	143	19	96	2767	551	895	804	53	9	5720
8	63	16	0	60	105	3	34	650	607	408	233	26	0	2207
9	433	47	3	112	106	4	30	1023	417	4948	1066	792	53	9035
10	103	10	2	37	40	5	22	934	191	1099	870	46	8	3365
11	161	3	0	18	8	3	5	100	31	1343	112	740	48	2573
12	25	0	1	1	2	0	0	15	3	162	15	81	17	321
Grand Total	1190	2303	318	3500	1778	253	390	6111	2020	9540	3334	1908	145	32790

(b) Trips with passenger

References

Arnott, R. (1996) Taxi travel should be subsidized, *Journal of Urban Economics*, 40 (3) 316‐333

Cairns, R.D. and C. Liston-Heyes (1996) Competition and regulation in the taxi industry, *Journal of Public Economics*, 59 (1) 1‐15

Ji, Y. and N. Geroliminis (2012) On the spatial partitioning of urban transportation networks, *Transportation Research Part B*, 1639-1656.

Shiode, N. and M. Batty (2000) Power Law Distributions in Real and Virtual Worlds, INET'2000, University College London.

Wong, K.I., S.C. Wong, M.G.H. Bell and H. Yang (2005) Modelling the bilateral micro-searching behavior for urban taxi services using the absorbing Markov chain approach, *Journal of Advanced Transportation*, 39 (1) 81‐104

Yang, H. and S.C. Wong (1998) A network model of urban taxi services, *Transportation Research Part B*: Methodological, 32 (4) 235‐246

Yang, H., S.C. Wong and K.I. Wong (2002) Demand-supply equilibrium of taxi services in a network under competition and regulation, *Transportation Research Part B*, 36 (9) 799‐819

Yang, H., M. Ye, W.H. Tang and S.C. Wong (2005) Regulating taxi services in the presence of congestion externality, *Transportation Research Part A*, 39 (1) 17‐40