ASSESSMENT OF HUMAN MOBILITY FROM TAXI GPS PROBE DATA IN BANGKOK, THAILAND

Songkorn Siangsuebchart, Sarawut Ninsawat, Apichon Witayangkurn, Surachet Pravinvongvuth
Asian Institute of Technology, 58 Moo 9, Km. 42, Paholyothin Highway, Klong Luang,
Pathum Thani, 12120 Thailand
Email: st118748@ait.ac.th, sarawutn@ait.ac.th, apichon@ait.ac.th,
spravinvongvuth@ait.ac.th

KEY WORDS: Human mobility, Taxi GPS Probe, Big data, Hadoop

ABSTRACT: Bangkok is one of the biggest Primate cities in the world (Tangchonlatip, 2007). The urbanization has long been expanding out of the central of Bangkok to the outer zones of Bangkok Metropolitan Region (BMR) such as the suburb area of Bangkok, Samutprakarn, Nonthaburi, Pathum Thani, etc. The transport planner requires the good sources of data to model the transport in BMR regularly. The new source of data, such as taxi GPS probe in BMR, has recently been available. These taxi GPS probe data have been continuously and regularly collected without any bias and cover all BMR. Apart from the taxi GPS probe data, the traffic analysis zone (TAZ) of the Office of Transport and Traffic Policy and Planning (OTP) are used as the base zones.

This paper illustrates the characteristics and structure of the raw GPS probe data. The one month of GPS probe data set of July 2015 contains 1,157 million GPS probe records. There are 10,885 vehicles based on the unique International Mobile Equipment Identity (IMEI) information. With the power of Hadoop Hive, the GPS probe data can be loaded and filtered out the noises and errors such as the error coordinates, duplicated records, etc. The only taxi GPS probes are extracted and taxi trips can be generated from the occupied status. Then the origin and destination of each trips can be derived. The traffic analysis zones (TAZ) of the Office of Transport and Traffic Planning and Policy could be integrated with the origin and destination of all taxi trips by using the ESRI's Hive spatial component. Finally, the approximate 3 million trips from taxi GPS probe are derived from 3,902 taxis.

1. INTRODUCTION

In September 2016, Kasikorn Research Center reveals that Bangkok gets the 12th rank of the most traffic-jammed in the world (Kasikorn Research Center 2016). The report also illustrates that Thailand has the opportunity loss of 60 million baht per day from the traffic jammed problem and this problem has a direct impact on the lives of Bangkokians. The research also says that 27 kilometers is the average distance of daily journey from their home to their workplace in the inner CBDs. Moreover, by this condition of traffic jammed condition during the rush hour such as 6 AM to 9AM, Bangkokians must spend more 35 minutes for regular traveling in weekdays.

The Thai government announced the development plans of transportation infrastructure such as new expressway, new roads, new motorway, and new rail commuters. These new transportation infrastructures certainly revive the above problems of Bangkokians. However, the primate city like Bangkok and BMR are dynamic so that urbanization becomes expanded not only by the expansion of transportation mode but also by uncontrolled expansion or urban sprawl. Understanding the human mobility can reveal the traveling demand of origin and destination (OD) pairs. The current traveling demand can be projected for the future demand for the better future transportation plan.

Taxis are the public transport that people in cities commonly used from their origins to destinations. This can reveal the human mobility in the cities. Fortunately, DLT has announced the regulation in 2015 that all taxis must be equipped with GPS component in order to be tracked the driving behavior of the taxi drivers and ensure the safety of the passengers (DLT, 2015). The coordinates (latitude and longitude) of the taxis are recorded together with the other attributes of the taxi such as unique identifier, meter status, etc. Each record of data is regularly collected such as every 10 seconds. These taxi GPS probe data can be considered as the Big Data by volume, velocity, etc.

In this study, the raw taxi GPS probe data of July 2015 is processed by using Hadoop HiveQL. The raw taxi GPS probe data can be loaded, filtered, and cleaned. The taxi trips or taxi origin and destination pairs can be extracted and spatially joined with the Traffic Analysis Zone (TAZ). Finally, the human mobility from taxi passengers can be visualized in GIS. Moreover, the connectivity between taxi mode and electric rail mode can be figured out from the spatial relationship between taxi origins or destinations and electric rail stations.

2. REQUIRED DATA

2.1 Taxi GPS Probe

TSquare Traffic Information Service of Toyota Tsusho Electronic)Thailand(or TTET has been collecting GPS probe data from 10,000 taxis in Bangkok and suburbs since 2012. The GPS data of the taxis are regularly collected every 3-5 seconds. This produces the Big Data of the taxi GPS probe.

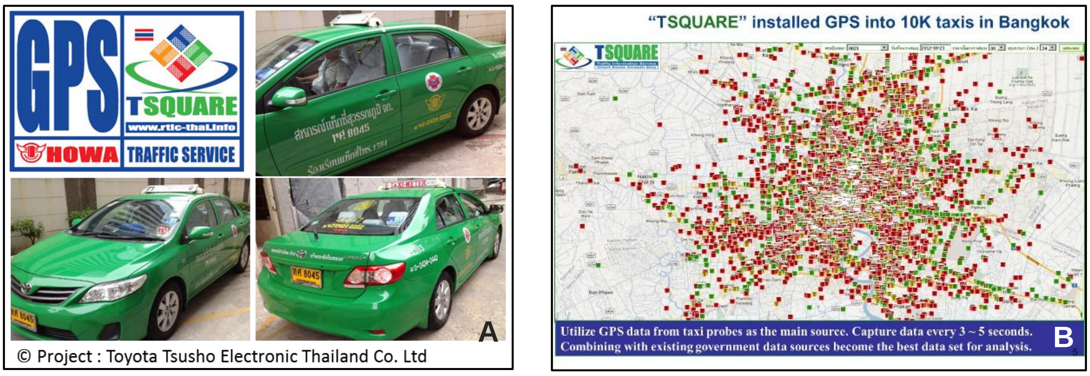
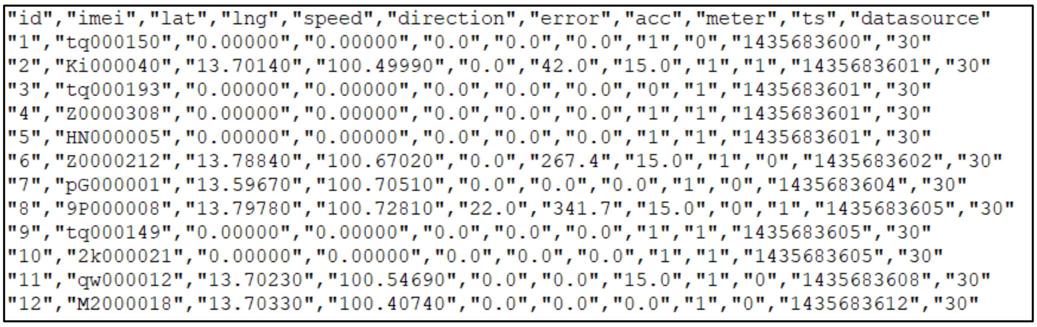

GPS
TSQUARE
www.rtic-thai.info
HOWA
TRAFFIC SERVICE
A
© Project : Toyota Tsusho Electronic Thailand Co. Ltd
“TSQUARE” installed GPS into 10K taxis in Bangkok
TSQUARE
Utilize GPS data from taxi probes as the main source. Capture data every 3 ~ 5 seconds.
Combining with existing government data sources become the best data set for analysis.
B

Figure 1 Taxis equipped with GPS (A). The 10,000 taxis' GPS probe data from the TSquare (B)

The raw data of taxi GPS probe from TTET is in the comma separated value or CSV format. The data of all taxis equipped with the GPS are collected based on one day per CSV file starting from the midnight to the end of the day at the next midnight. Table 1 shows the detail data of the CSV file that contains id, IMEI, GPS coordinates, speed, direction, meter status, etc. In this paper, the past taxi GPS probe data of August 2015 has been using for designing, developing, and testing the processes. The total records of one month data of August 2015 CSV files is 1,156,897,579 records.



```
"id","imei","lat","lng","speed","direction","error","acc","meter","ts","datasource"
"1","tq000150","0.00000","0.00000","0.0","0.0","0.0","1","0","1435683600","30"
"2","Ki000040","13.70140","100.49990","0.0","42.0","15.0","1","1","1435683601","30"
"3","tq000193","0.00000","0.00000","0.0","0.0","0.0","0","1","1435683601","30"
"4","Z0000308","0.00000","0.00000","0.0","0.0","0.0","1","1","1435683601","30"
"5","HN000005","0.00000","0.00000","0.0","0.0","0.0","1","1","1435683601","30"
"6","Z0000212","13.78840","100.67020","0.0","267.4","15.0","1","0","1435683602","30"
"7","pG000001","13.59670","100.70510","0.0","0.0","0.0","1","0","1435683604","30"
"8","9P000008","13.79780","100.72810","22.0","341.7","15.0","0","1","1435683605","30"
"9","tq000149","0.00000","0.00000","0.0","0.0","0.0","1","1","1435683605","30"
"10","2k000021","0.00000","0.00000","0.0","0.0","0.0","1","1","1435683605","30"
"11","qw000012","13.70230","100.54690","0.0","0.0","15.0","1","0","1435683608","30"
"12","M2000018","13.70330","100.40740","0.0","0.0","0.0","1","0","1435683612","30"
```

Figure 2 The raw CSV data of taxi GPS probe data

Table 1 The taxi GPS probe CSV file description

No.	Column	Description
1	id	Unique id of the records in the CSV file
2	IMEI	Unique id of International Mobile Equipment Identity (IMEI) number of each taxi
3	lat	Latitude value of the taxi location
4	lon	Longitude value of the taxi location
5	speed	Driving speed of the taxi at the location
6	direction	Driving direction of the taxi at the location
7	error	GPS error value 0 = no error
8	acc	Engine status of 0 = On, 1 = Off
9	meter	1 = On (with passengers) , 0 = Off (no passengers)
10	ts	Timestamp in Unix epoch format
11	datasource	Taxi GPS probe data source is in 8 or 9

2.2 Traffic Analysis Zones (TAZ)

The Extended Bangkok Urban Model or eBUM is the transport model continuously developed by the Office of Transport and Traffic Policy and Planning(OTP), ministry of Transport since 1995. During 2010 to 2011, eBUM covers 8 provinces including Bangkok, Samutprakarn, Nontaburi, Pathumthani, Nakhon Prathom, Samut Sakorn, Ayutthaya, and Chachoengsao. These 8 provinces are divided into 1,657 TAZs which are smaller than the subdistricts. The human mobility from taxi GPS probe data will be aggregated into these TAZs.

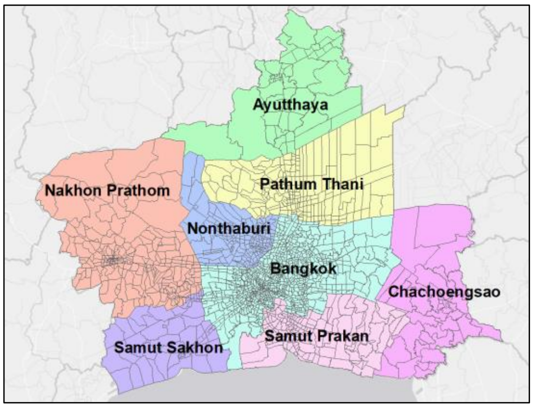


Figure 3 OTP 1,657 TAZs

3. HUMAN MOBILITY IDENTIFICATION

The procedures of taxi GPS probe data processing based on the Hadoop platform has 4 main steps as in the below diagram. It starts from loading GPS Probe data into Hadoop Hive. Typically, GPS probe data contains outliers or errors from the GPS signal or equipment errors. All these outliers and errors must be eliminated in the next step. Not only the errors but also the duplicated records should be removed for compacting the data set. Finally, the taxi trips can be derived into the origin and destination matrix.

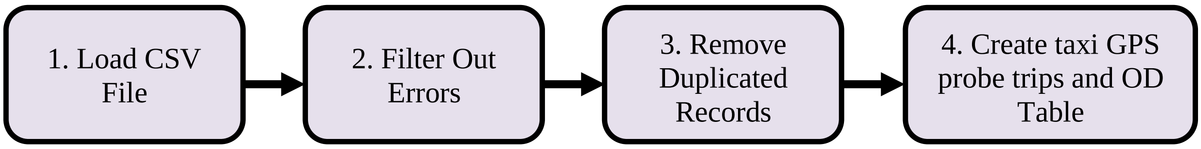
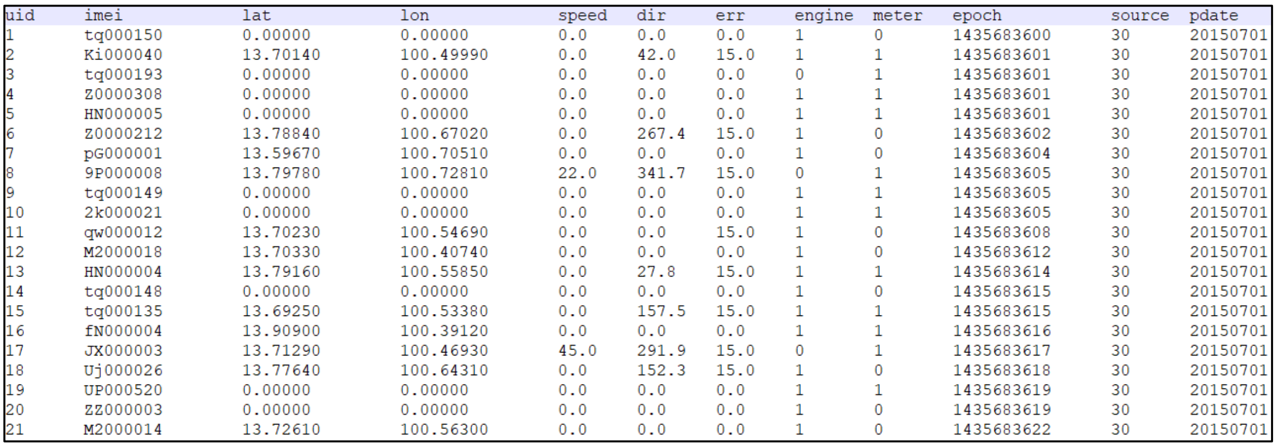

1. Load CSV File
2. Filter Out Errors
3. Remove Duplicated Records
4. Create taxi GPS probe trips and OD Table

Figure 4 Four steps of taxi GPS probe data processing in Hadoop

3.1 Load CSV File into Hadoop Hive

All 31 days CSV files of August 2015 are transferred into the Hadoop system. Then all the files are loaded into the created Hive table partitioned by each date for the efficient usage. There are 1,156,897,579 records loaded into the Hive table as below.



uid	imei	lat	lon	speed	dir	err	engine	meter	epoch	source	pdate
1	tq000150	0.00000	0.00000	0.0	0.0	0.0	1	0	1435683600	30	20150701
2	Ki000040	13.70140	100.49990	0.0	42.0	15.0	1	1	1435683601	30	20150701
3	tq000193	0.00000	0.00000	0.0	0.0	0.0	0	1	1435683601	30	20150701
4	Z0000308	0.00000	0.00000	0.0	0.0	0.0	1	1	1435683601	30	20150701
5	HN000005	0.00000	0.00000	0.0	0.0	0.0	1	1	1435683601	30	20150701
6	Z0000212	13.78840	100.67020	0.0	267.4	15.0	1	0	1435683602	30	20150701
7	pG000001	13.59670	100.70510	0.0	0.0	0.0	1	0	1435683604	30	20150701
8	9P000008	13.79780	100.72810	22.0	341.7	15.0	0	1	1435683605	30	20150701
9	tq000149	0.00000	0.00000	0.0	0.0	0.0	1	1	1435683605	30	20150701
10	2k000021	0.00000	0.00000	0.0	0.0	0.0	1	1	1435683605	30	20150701
11	qw000012	13.70230	100.54690	0.0	0.0	15.0	1	0	1435683608	30	20150701
12	M2000018	13.70330	100.40740	0.0	0.0	0.0	1	0	1435683612	30	20150701
13	HN000004	13.79160	100.55850	0.0	27.8	15.0	1	1	1435683614	30	20150701
14	tq000148	0.00000	0.00000	0.0	0.0	0.0	1	0	1435683615	30	20150701
15	tq000135	13.69250	100.53380	0.0	157.5	15.0	1	1	1435683615	30	20150701
16	fN000004	13.90900	100.39120	0.0	0.0	0.0	1	1	1435683616	30	20150701
17	JX000003	13.71290	100.46930	45.0	291.9	15.0	0	1	1435683617	30	20150701
18	Uj000026	13.77640	100.64310	0.0	152.3	15.0	1	0	1435683618	30	20150701
19	UP000520	0.00000	0.00000	0.0	0.0	0.0	1	1	1435683619	30	20150701
20	ZZ000003	0.00000	0.00000	0.0	0.0	0.0	1	0	1435683619	30	20150701
21	M2000014	13.72610	100.56300	0.0	0.0	0.0	1	0	1435683622	30	20150701

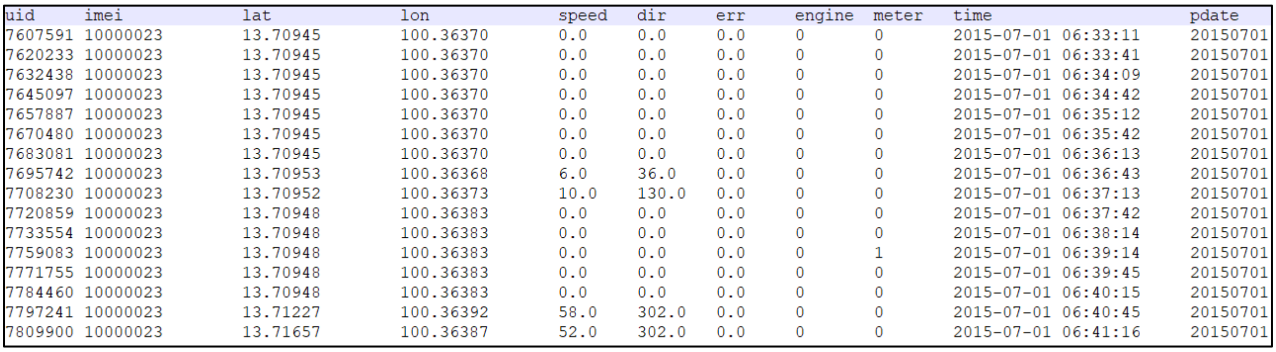
Figure 5 GPS probe raw data in the Hive table

3.2 Filter Out Errors

The raw data table must be processed to get rid of the error and unnecessary data in the raw data source. The number of records after filtering is 403,293,148. The required records' conditions are listed as below. In the meantime, the Unix epoch format is converted into the timestamp format and the output table is ascendingly ordered by date, IMEI, and time.

Table 2 Data condition to select taxi GPS probe data

No.	Data	Required Value
1	Error	0
2	Engine	0
3	Latitude	More than 0
4	Longitude	More than 0
5	Source	Taxi only (8 or 9)



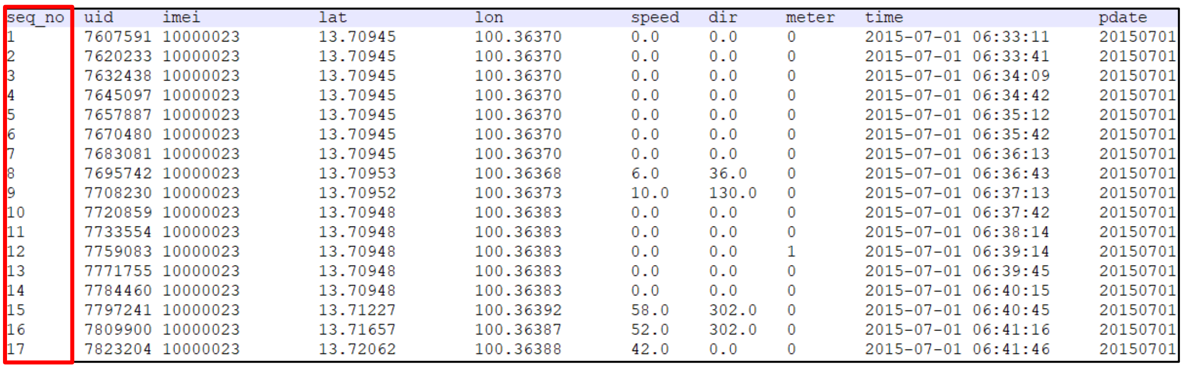
uid	imei	lat	lon	speed	dir	err	engine	meter	time	pdate
7607591	10000023	13.70945	100.36370	0.0	0.0	0.0	0	0	2015-07-01 06:33:11	20150701
7620233	10000023	13.70945	100.36370	0.0	0.0	0.0	0	0	2015-07-01 06:33:41	20150701
7632438	10000023	13.70945	100.36370	0.0	0.0	0.0	0	0	2015-07-01 06:34:09	20150701
7645097	10000023	13.70945	100.36370	0.0	0.0	0.0	0	0	2015-07-01 06:34:42	20150701
7657887	10000023	13.70945	100.36370	0.0	0.0	0.0	0	0	2015-07-01 06:35:12	20150701
7670480	10000023	13.70945	100.36370	0.0	0.0	0.0	0	0	2015-07-01 06:35:42	20150701
7683081	10000023	13.70945	100.36370	0.0	0.0	0.0	0	0	2015-07-01 06:36:13	20150701
7695742	10000023	13.70953	100.36368	6.0	36.0	0.0	0	0	2015-07-01 06:36:43	20150701
7708230	10000023	13.70952	100.36373	10.0	130.0	0.0	0	0	2015-07-01 06:37:13	20150701
7720859	10000023	13.70948	100.36383	0.0	0.0	0.0	0	0	2015-07-01 06:37:42	20150701
7733554	10000023	13.70948	100.36383	0.0	0.0	0.0	0	0	2015-07-01 06:38:14	20150701
7759083	10000023	13.70948	100.36383	0.0	0.0	0.0	0	1	2015-07-01 06:39:14	20150701
7771755	10000023	13.70948	100.36383	0.0	0.0	0.0	0	0	2015-07-01 06:39:45	20150701
7784460	10000023	13.70948	100.36383	0.0	0.0	0.0	0	0	2015-07-01 06:40:15	20150701
7797241	10000023	13.71227	100.36392	58.0	302.0	0.0	0	0	2015-07-01 06:40:45	20150701
7809900	10000023	13.71657	100.36387	52.0	302.0	0.0	0	0	2015-07-01 06:41:16	20150701

Figure 6 The taxi only GPS probe data after applying filter

3.3 Remove Duplicated Records

After investigating the loaded raw Taxi GPS probe data, there are 2 types of duplication in the data set. They are 1) duplicated latitude and longitude for each IMEI and date and 2) duplicated timestamp and IMEI. These duplicated records must be removed by following steps.

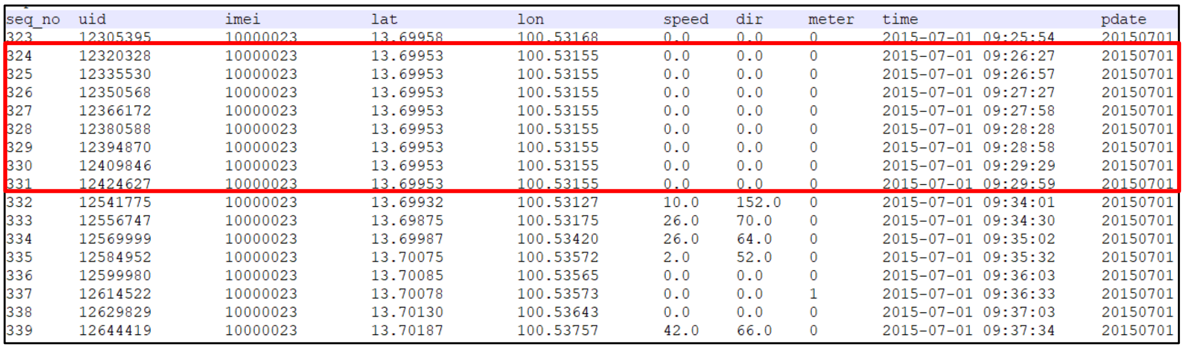
1) Add the sequence number based on date and IMEI
The sequence number will be used for eliminating the duplicated records in the next process. The sequence number can be created based on date and IMEI ordered by the timestamp.



seq_no	uid	imei	lat	lon	speed	dir	meter	time	pdate
1	7607591	10000023	13.70945	100.36370	0.0	0.0	0	2015-07-01 06:33:11	20150701
2	7620233	10000023	13.70945	100.36370	0.0	0.0	0	2015-07-01 06:33:41	20150701
3	7632438	10000023	13.70945	100.36370	0.0	0.0	0	2015-07-01 06:34:09	20150701
4	7645097	10000023	13.70945	100.36370	0.0	0.0	0	2015-07-01 06:34:42	20150701
5	7657887	10000023	13.70945	100.36370	0.0	0.0	0	2015-07-01 06:35:12	20150701
6	7670480	10000023	13.70945	100.36370	0.0	0.0	0	2015-07-01 06:35:42	20150701
7	7683081	10000023	13.70945	100.36370	0.0	0.0	0	2015-07-01 06:36:13	20150701
8	7695742	10000023	13.70953	100.36368	6.0	36.0	0	2015-07-01 06:36:43	20150701
9	7708230	10000023	13.70952	100.36373	10.0	130.0	0	2015-07-01 06:37:13	20150701
10	7720859	10000023	13.70948	100.36383	0.0	0.0	0	2015-07-01 06:37:42	20150701
11	7733554	10000023	13.70948	100.36383	0.0	0.0	0	2015-07-01 06:38:14	20150701
12	7759083	10000023	13.70948	100.36383	0.0	0.0	1	2015-07-01 06:39:14	20150701
13	7771755	10000023	13.70948	100.36383	0.0	0.0	0	2015-07-01 06:39:45	20150701
14	7784460	10000023	13.70948	100.36383	0.0	0.0	0	2015-07-01 06:40:15	20150701
15	7797241	10000023	13.71227	100.36392	58.0	302.0	0	2015-07-01 06:40:45	20150701
16	7809900	10000023	13.71657	100.36387	52.0	302.0	0	2015-07-01 06:41:16	20150701
17	7823204	10000023	13.72062	100.36388	42.0	0.0	0	2015-07-01 06:41:46	20150701

Figure 7 The sequence number based on date and IMEI

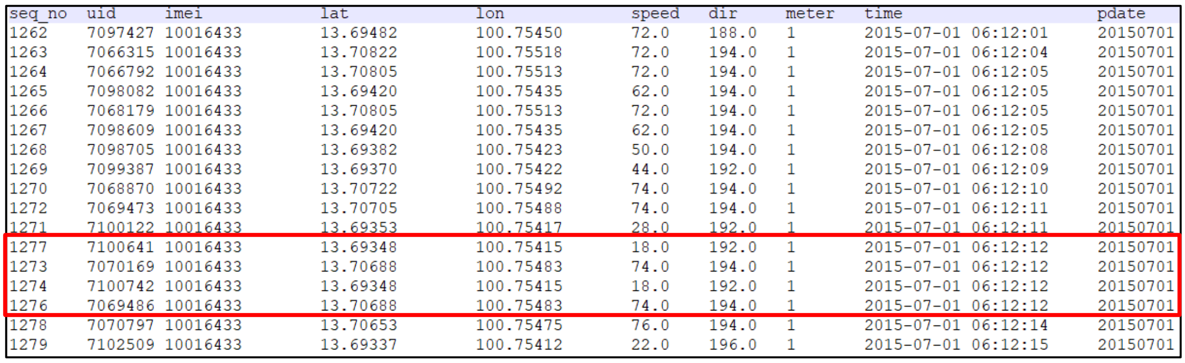
2) Remove duplicated latitude longitude for each IMEI and date
There are the duplicated records of the identical latitude and longitude for each IMEI and date. These records are the overhead for the processing. These duplicated records can be removed by using the sequence number. The output number of records is 394,341,914.



seq_no	uid	imei	lat	lon	speed	dir	meter	time	pdate
323	12305395	10000023	13.69958	100.53168	0.0	0.0	0	2015-07-01 09:25:54	20150701
324	12320328	10000023	13.69953	100.53155	0.0	0.0	0	2015-07-01 09:26:27	20150701
325	12335530	10000023	13.69953	100.53155	0.0	0.0	0	2015-07-01 09:26:57	20150701
326	12350568	10000023	13.69953	100.53155	0.0	0.0	0	2015-07-01 09:27:27	20150701
327	12366172	10000023	13.69953	100.53155	0.0	0.0	0	2015-07-01 09:27:58	20150701
328	12380588	10000023	13.69953	100.53155	0.0	0.0	0	2015-07-01 09:28:28	20150701
329	12394870	10000023	13.69953	100.53155	0.0	0.0	0	2015-07-01 09:28:58	20150701
330	12409846	10000023	13.69953	100.53155	0.0	0.0	0	2015-07-01 09:29:29	20150701
331	12424627	10000023	13.69953	100.53155	0.0	0.0	0	2015-07-01 09:29:59	20150701
332	12541775	10000023	13.69932	100.53127	10.0	152.0	0	2015-07-01 09:34:01	20150701
333	12556747	10000023	13.69875	100.53175	26.0	70.0	0	2015-07-01 09:34:30	20150701
334	12569999	10000023	13.69987	100.53420	26.0	64.0	0	2015-07-01 09:35:02	20150701
335	12584952	10000023	13.70075	100.53572	2.0	52.0	0	2015-07-01 09:35:32	20150701
336	12599980	10000023	13.70085	100.53565	0.0	0.0	0	2015-07-01 09:36:03	20150701
337	12614522	10000023	13.70078	100.53573	0.0	0.0	1	2015-07-01 09:36:33	20150701
338	12629829	10000023	13.70130	100.53643	0.0	0.0	0	2015-07-01 09:37:03	20150701
339	12644419	10000023	13.70187	100.53757	42.0	66.0	0	2015-07-01 09:37:34	20150701

Figure 8 The duplicated latitude and longitude

3) Remove records that have the duplicated timestamp and IMEI
The duplicated records of the identical timestamp and IMEI can be removed for reducing the processing time. The output number of records is 394,340,861.

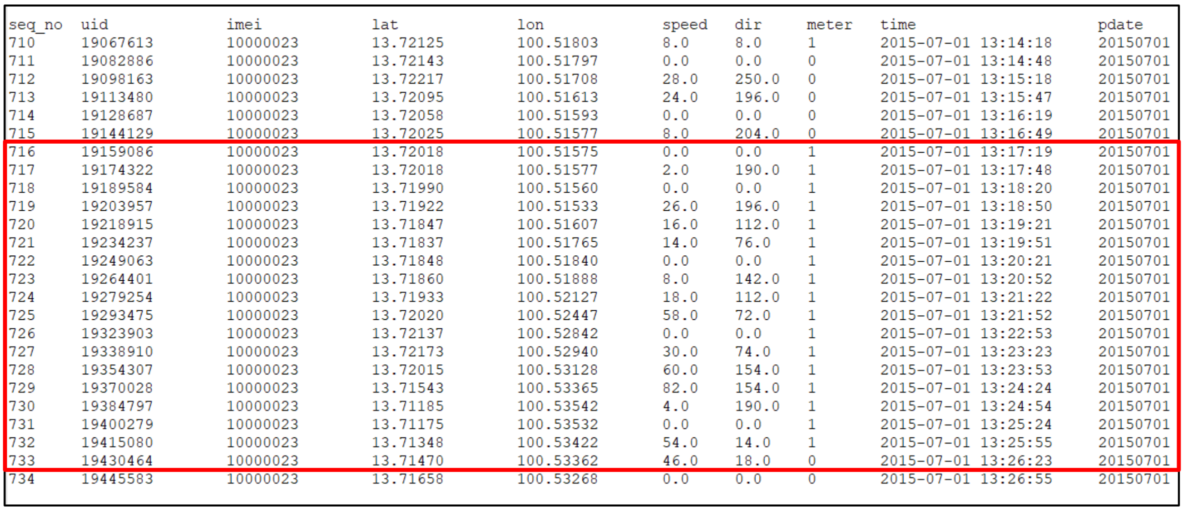


seq_no	uid	imei	lat	lon	speed	dir	meter	time	pdate
1262	7097427	10016433	13.69482	100.75450	72.0	188.0	1	2015-07-01 06:12:01	20150701
1263	7066315	10016433	13.70822	100.75518	72.0	194.0	1	2015-07-01 06:12:04	20150701
1264	7066792	10016433	13.70805	100.75513	72.0	194.0	1	2015-07-01 06:12:05	20150701
1265	7098082	10016433	13.69420	100.75435	62.0	194.0	1	2015-07-01 06:12:05	20150701
1266	7068179	10016433	13.70805	100.75513	72.0	194.0	1	2015-07-01 06:12:05	20150701
1267	7098609	10016433	13.69420	100.75435	62.0	194.0	1	2015-07-01 06:12:05	20150701
1268	7098705	10016433	13.69382	100.75423	50.0	194.0	1	2015-07-01 06:12:08	20150701
1269	7099387	10016433	13.69370	100.75422	44.0	192.0	1	2015-07-01 06:12:09	20150701
1270	7068870	10016433	13.70722	100.75492	74.0	194.0	1	2015-07-01 06:12:10	20150701
1272	7069473	10016433	13.70705	100.75488	74.0	194.0	1	2015-07-01 06:12:11	20150701
1271	7100122	10016433	13.69353	100.75417	28.0	192.0	1	2015-07-01 06:12:11	20150701
1277	7100641	10016433	13.69348	100.75415	18.0	192.0	1	2015-07-01 06:12:12	20150701
1273	7070169	10016433	13.70688	100.75483	74.0	194.0	1	2015-07-01 06:12:12	20150701
1274	7100742	10016433	13.69348	100.75415	18.0	192.0	1	2015-07-01 06:12:12	20150701
1276	7069486	10016433	13.70688	100.75483	74.0	194.0	1	2015-07-01 06:12:12	20150701
1278	7070797	10016433	13.70653	100.75475	76.0	194.0	1	2015-07-01 06:12:14	20150701
1279	7102509	10016433	13.69337	100.75412	22.0	196.0	1	2015-07-01 06:12:15	20150701

Figure 9 The duplicated records of the identical timestamp, and IMEI

3.4 Create Taxi GPS Probe Trips and OD Table

1) Create taxi GPS probe trips
To create the taxi GPS probe trips, the records must be ordered by date, IMEI, and time ascendingly. The taxi trips can be identified from the meter status. The status of 1 means the meter is on and taxi is occupied. The status of 0 means there is no passenger on the taxi. The sample of an occupied taxi trip data is shown in the below figures. This occupied trip is started from 13:17:19 to 13:26:23 of 1st August 2015. In figure 11, the red points are GPS locations of the identical occupied taxi trip. The total records of the table are 394,340,011.



seq_no	uid	imei	lat	lon	speed	dir	meter	time	pdate
710	19067613	10000023	13.72125	100.51803	8.0	8.0	1	2015-07-01 13:14:18	20150701
711	19082886	10000023	13.72143	100.51797	0.0	0.0	0	2015-07-01 13:14:48	20150701
712	19098163	10000023	13.72217	100.51708	28.0	250.0	0	2015-07-01 13:15:18	20150701
713	19113480	10000023	13.72095	100.51613	24.0	196.0	0	2015-07-01 13:15:47	20150701
714	19128687	10000023	13.72058	100.51593	0.0	0.0	0	2015-07-01 13:16:19	20150701
715	19144129	10000023	13.72025	100.51577	8.0	204.0	0	2015-07-01 13:16:49	20150701
716	19159086	10000023	13.72018	100.51575	0.0	0.0	1	2015-07-01 13:17:19	20150701
717	19174322	10000023	13.72018	100.51577	2.0	190.0	1	2015-07-01 13:17:48	20150701
718	19189584	10000023	13.71990	100.51560	0.0	0.0	1	2015-07-01 13:18:20	20150701
719	19203957	10000023	13.71922	100.51533	26.0	196.0	1	2015-07-01 13:18:50	20150701
720	19218915	10000023	13.71847	100.51607	16.0	112.0	1	2015-07-01 13:19:21	20150701
721	19234237	10000023	13.71837	100.51765	14.0	76.0	1	2015-07-01 13:19:51	20150701
722	19249063	10000023	13.71848	100.51840	0.0	0.0	1	2015-07-01 13:20:21	20150701
723	19264401	10000023	13.71860	100.51888	8.0	142.0	1	2015-07-01 13:20:52	20150701
724	19279254	10000023	13.71933	100.52127	18.0	112.0	1	2015-07-01 13:21:22	20150701
725	19293475	10000023	13.72020	100.52447	58.0	72.0	1	2015-07-01 13:21:52	20150701
726	19323903	10000023	13.72137	100.52842	0.0	0.0	1	2015-07-01 13:22:53	20150701
727	19338910	10000023	13.72173	100.52940	30.0	74.0	1	2015-07-01 13:23:23	20150701
728	19354307	10000023	13.72015	100.53128	60.0	154.0	1	2015-07-01 13:23:53	20150701
729	19370028	10000023	13.71543	100.53365	82.0	154.0	1	2015-07-01 13:24:24	20150701
730	19384797	10000023	13.71185	100.53542	4.0	190.0	1	2015-07-01 13:24:54	20150701
731	19400279	10000023	13.71175	100.53532	0.0	0.0	1	2015-07-01 13:25:24	20150701
732	19415080	10000023	13.71348	100.53422	54.0	14.0	1	2015-07-01 13:25:55	20150701
733	19430464	10000023	13.71470	100.53362	46.0	18.0	0	2015-07-01 13:26:23	20150701
734	19445583	10000023	13.71658	100.53268	0.0	0.0	0	2015-07-01 13:26:55	20150701

Figure 10 The occupied taxi trip

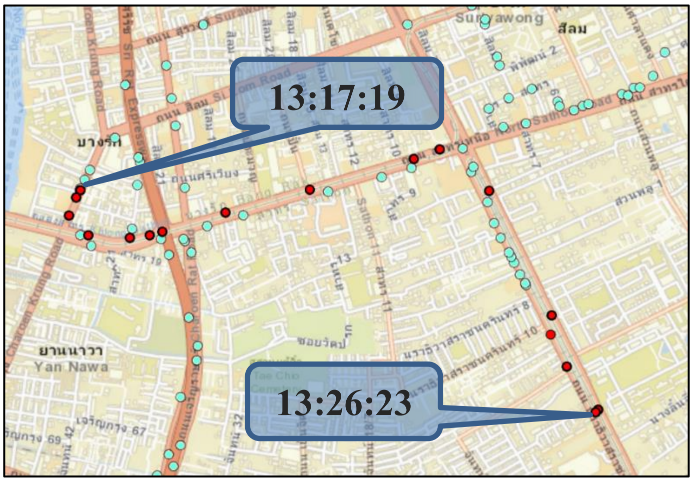
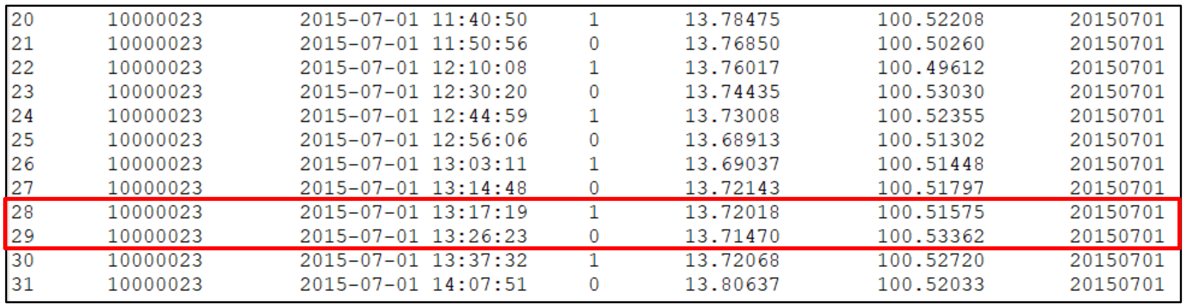


Figure 11 The occupied taxi map

2) Create the taxi trips get in and get off location table

In order to get the OD pair of each taxi trip, the get in and get off location must be derived from the data. From the output table of previous process, the get in (meter status is 1) and get off (meter status is 0) locations are reserved in the output table. The rest of the locations between get in and get off location are eliminated. The total number of records are 3,284,697.

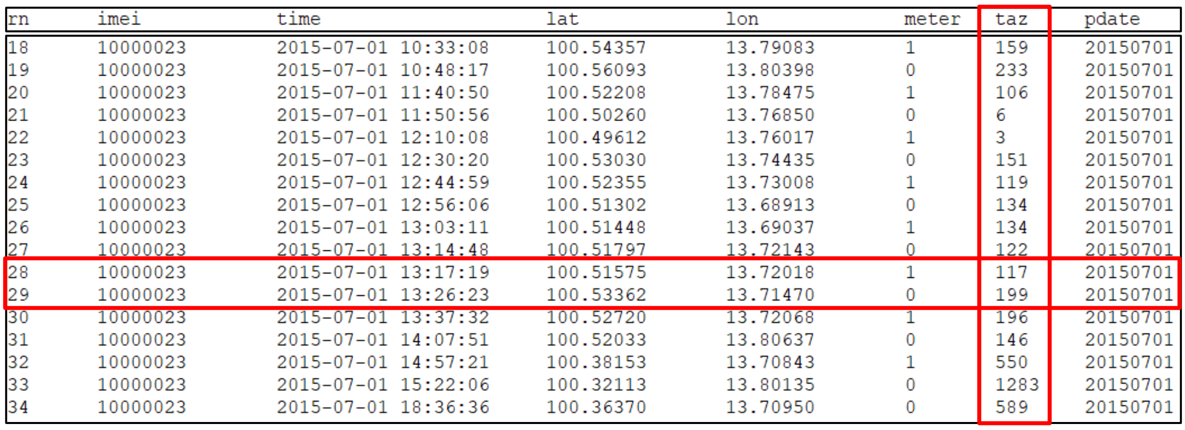


20	10000023	2015-07-01 11:40:50	1	13.78475	100.52208	20150701
21	10000023	2015-07-01 11:50:56	0	13.76850	100.50260	20150701
22	10000023	2015-07-01 12:10:08	1	13.76017	100.49612	20150701
23	10000023	2015-07-01 12:30:20	0	13.74435	100.53030	20150701
24	10000023	2015-07-01 12:44:59	1	13.73008	100.52355	20150701
25	10000023	2015-07-01 12:56:06	0	13.68913	100.51302	20150701
26	10000023	2015-07-01 13:03:11	1	13.69037	100.51448	20150701
27	10000023	2015-07-01 13:14:48	0	13.72143	100.51797	20150701
28	10000023	2015-07-01 13:17:19	1	13.72018	100.51575	20150701
29	10000023	2015-07-01 13:26:23	0	13.71470	100.53362	20150701
30	10000023	2015-07-01 13:37:32	1	13.72068	100.52720	20150701
31	10000023	2015-07-01 14:07:51	0	13.80637	100.52033	20150701

Figure 12 The get in (origin) and get off (destination) locations

3) Get the TAZ number from the TAZ geometry table

To get the OD table based on TAZ, the TAZ number must be spatially joined into the get in and get off locations table. In this case, ESRI ST_Geometry functions for Hadoop are used. The records that beyond the TAZ polygons are automatically eliminated. The remaining records are 3,242,424. Below is the output table with TAZ number for each get in and get off location.

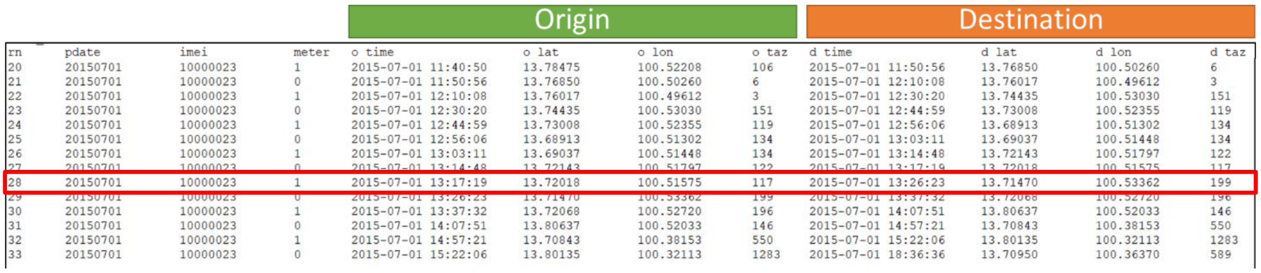


rn	imei	time	lat	lon	meter	taz	pdate
18	10000023	2015-07-01 10:33:08	100.54357	13.79083	1	159	20150701
19	10000023	2015-07-01 10:48:17	100.56093	13.80398	0	233	20150701
20	10000023	2015-07-01 11:40:50	100.52208	13.78475	1	106	20150701
21	10000023	2015-07-01 11:50:56	100.50260	13.76850	0	6	20150701
22	10000023	2015-07-01 12:10:08	100.49612	13.76017	1	3	20150701
23	10000023	2015-07-01 12:30:20	100.53030	13.74435	0	151	20150701
24	10000023	2015-07-01 12:44:59	100.52355	13.73008	1	119	20150701
25	10000023	2015-07-01 12:56:06	100.51302	13.68913	0	134	20150701
26	10000023	2015-07-01 13:03:11	100.51448	13.69037	1	134	20150701
27	10000023	2015-07-01 13:14:48	100.51797	13.72143	0	122	20150701
28	10000023	2015-07-01 13:17:19	100.51575	13.72018	1	117	20150701
29	10000023	2015-07-01 13:26:23	100.53362	13.71470	0	199	20150701
30	10000023	2015-07-01 13:37:32	100.52720	13.72068	1	196	20150701
31	10000023	2015-07-01 14:07:51	100.52033	13.80637	0	146	20150701
32	10000023	2015-07-01 14:57:21	100.38153	13.70843	1	550	20150701
33	10000023	2015-07-01 15:22:06	100.32113	13.80135	0	1283	20150701
34	10000023	2015-07-01 18:36:36	100.36370	13.70950	0	589	20150701

Figure 13 The spatially joined TAZ number for each get in and get off location

4) Create Taxi OD Table

Finally, the taxi OD table is then created. Each record indicates a taxi trip from an origin TAZ to a destination TAZ together with the get-in location and time and get-off location and time. There are both occupied and vacant trips indicated by the meter status. There are 3,129,329 taxi trips in July 2015 including both 1,553,467 vacant trips and 1,575,862 occupied trips.



				Origin				Destination			
rn	pdate	imei	meter	o time	o lat	o lon	o taz	d time	d lat	d lon	d taz
20	20150701	10000023	1	2015-07-01 11:40:50	13.78475	100.52208	106	2015-07-01 11:50:56	13.76850	100.50260	6
21	20150701	10000023	0	2015-07-01 11:50:56	13.76850	100.50260	6	2015-07-01 12:10:08	13.76017	100.49612	3
22	20150701	10000023	1	2015-07-01 12:10:08	13.76017	100.49612	3	2015-07-01 12:30:20	13.74435	100.53030	151
23	20150701	10000023	0	2015-07-01 12:30:20	13.74435	100.53030	151	2015-07-01 12:44:59	13.73008	100.52355	119
24	20150701	10000023	1	2015-07-01 12:44:59	13.73008	100.52355	119	2015-07-01 12:56:06	13.68913	100.51302	134
25	20150701	10000023	0	2015-07-01 12:56:06	13.68913	100.51302	134	2015-07-01 13:03:11	13.69037	100.51448	134
26	20150701	10000023	1	2015-07-01 13:03:11	13.69037	100.51448	134	2015-07-01 13:14:48	13.72143	100.51797	122
27	20150701	10000023	0	2015-07-01 13:14:48	13.72143	100.51797	122	2015-07-01 13:17:19	13.72018	100.51575	117
28	20150701	10000023	1	2015-07-01 13:17:19	13.72018	100.51575	117	2015-07-01 13:26:23	13.71470	100.53362	199
29	20150701	10000023	0	2015-07-01 13:26:23	13.71470	100.53362	199	2015-07-01 13:37:32	13.72068	100.52720	196
30	20150701	10000023	1	2015-07-01 13:37:32	13.72068	100.52720	196	2015-07-01 14:07:51	13.80637	100.52033	146
31	20150701	10000023	0	2015-07-01 14:07:51	13.80637	100.52033	146	2015-07-01 14:57:21	13.70843	100.38153	550
32	20150701	10000023	1	2015-07-01 14:57:21	13.70843	100.38153	550	2015-07-01 15:22:06	13.80135	100.32113	1283
33	20150701	10000023	0	2015-07-01 15:22:06	13.80135	100.32113	1283	2015-07-01 18:36:36	13.70950	100.36370	589

Figure 14 The origin and destination TAZ table

4. RESULTS

4.1 Statistics of the Hadoop Hive processing

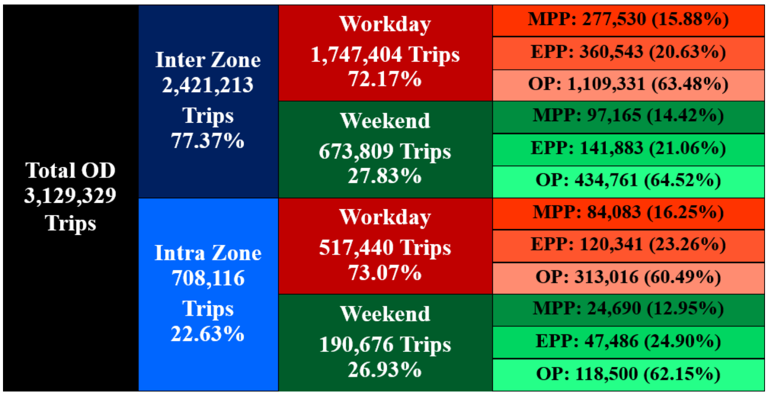
The below table shows the summary of the pre-processing process of taxi GPS probe data in term of records, processing time, and number of IMEI.

Table 3 Hadoop Hive processing summary

No.	Process	Input	Output	HH	MM	SS	Imei
1	Import CSV to hadoop hive partitioned by date	1,156,897,579	1,156,897,579	0	39	47	10,885
2	Filter out non-taxi, error, and engine stop records	1,156,897,579	403,293,148	0	46	37.428	4,460
3	Create sequence no based on each date and imei	403,293,148	403,293,148	0	45	27.947	4,460
	Remove duplicated lat and lon for each imei	403,293,148	394,341,914	0	59	2.466	4,460
	Analyze Table	394,341,914	394,341,914	0	7	37.328	
	Remove duplicated timestamp for each imei	394,341,914	394,340,861	1	2	29.414	4,460
	Analyze Table	394,340,861	394,340,861	0	4	5.27	
4	Create taxi trips and remove one record imei	394,340,861	394,340,011	0	49	39.344	4,409
	Analyze Table	394,340,011	394,340,011	0	4	9.294	
	Create the get in and get off locations	394,340,011	3,284,697	0	35	7.348	4,409
	Get the TAZ number	3,284,697	3,242,424	0	48	38.722	3,915
	Create the OD Table	3,242,424	3,129,329	0	0	31.414	3,902
				6	43	12.975	

4.2 Human mobility from taxi OD

From the OD TAZ trips of 3,129,329 trips are divided into inter-zones and intra-zone. Both inter-zone and intra-zone are categorized into working day (Monday to Friday) and Weekend (Saturday and Sunday). Then the Moring Peak Period (MPP) during 6 AM to 9 AM, Evening Peak Period (EPP) during 4 PM and 7 PM, and other periods (OP) are derived. The below table is the summary of the data.



Total OD 3,129,329 Trips

Inter Zone 2,421,213 Trips 77.37%

Workday 1,747,404 Trips 72.17%

MPP: 277,530 (15.88%)

EPP: 360,543 (20.63%)

OP: 1,109,331 (63.48%)

Weekend 673,809 Trips 27.83%

MPP: 97,165 (14.42%)

EPP: 141,883 (21.06%)

OP: 434,761 (64.52%)

Intra Zone 708,116 Trips 22.63%

Workday 517,440 Trips 73.07%

MPP: 84,083 (16.25%)

EPP: 120,341 (23.26%)

OP: 313,016 (60.49%)

Weekend 190,676 Trips 26.93%

MPP: 24,690 (12.95%)

EPP: 47,486 (24.90%)

OP: 118,500 (62.15%)

Figure 15 The summary of OD data based on the inter-zone and intra-zone

Due to this one month data set, the ratios of workday and weekend of the inter-zone and intra-zone trips are similar that is around 73% of workday trips and 27% trips of weekend. The trips' ratios of morning peak period, evening peak period, and other periods are in the same shape of all categories.

When investigating the origin TAZ and destination TAZ from the OD table, the daily average trips and morning peak period average trips are considered. The following table shows top ten of the highest average trips TAZ for both daily and morning peak period.

Table 4 The top ten TAZ of average daily and morning peak period origin and destination trips on workdays

Rank	Daily				Morning Peak Period			
	O_Zone	O_AVG_Trips	D_Zone	D_AVG_Trips	O_Zone	O_AVG_Trips	D_Zone	D_AVG_Trips
1	231	699.26	231	699.13	266	138.61	724	147.52
2	266	691.40	266	691.52	724	131.61	266	146.39
3	207	625.96	207	628.48	231	109.52	231	109.96
4	724	606.21	724	607.96	183	78.52	183	83.83
5	154	531.96	154	532.43	154	77.69	154	81.43
6	233	418.35	233	419.41	42	71.87	42	79.65
7	12	411.78	12	412.00	207	62.74	233	67.26
8	156	369.22	156	370.69	233	61.78	207	65.21
9	183	368.87	183	369.30	716	59.17	716	61.82
10	412	538.09	412	359.04	156	54.82	156	58.91

The top three TAZ zones for daily origin and destination trips are 231, 266, and 207. For the morning peak period, the top three TAZ zone are 266, 724, and 231. The important point of interests in these four TAZ zones are listed in the below table.

Table 5 The important point of interests in the top four TAZ zones

Zone	Important POI
231	BTS and MRT electric rail interchange stations
266	Donmuang airport
207	MRT, BTS stations, office buildings, hospital
724	Suvannabhumi airport

Below maps illustrates the daily and morning peak period origin and destination trips in each zone. The number of origin and destination trips for each zone are categorized into 7 ranks for easier visualization. The above 4 TAZ zones are labeled the TAZ number in below figure.

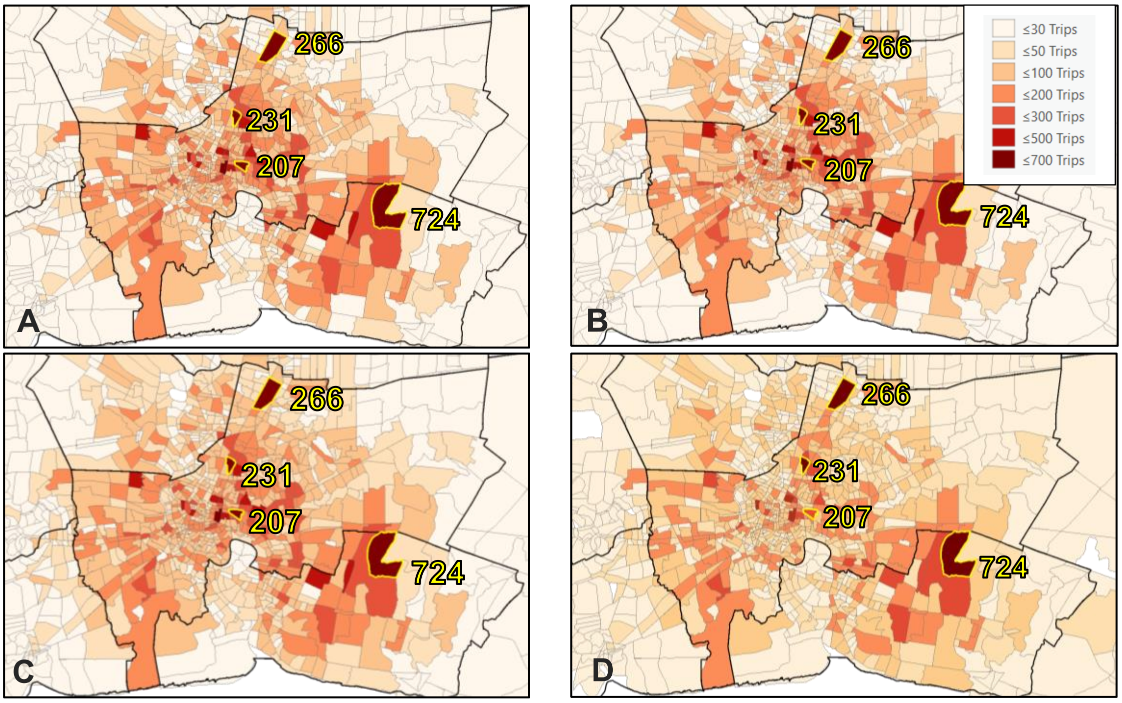


Figure 16 The average daily origin trips (A). The average daily destination trips (B). The average morning peak period origin (C). The average morning peak period destination (D).

From the taxi OD table, the OD links between origin to destination zone can be created and visualized. The following map shows the trips from TAZ 231 (BTS and MRT interchange stations) to others and the trips from other zones to TAZ 231.

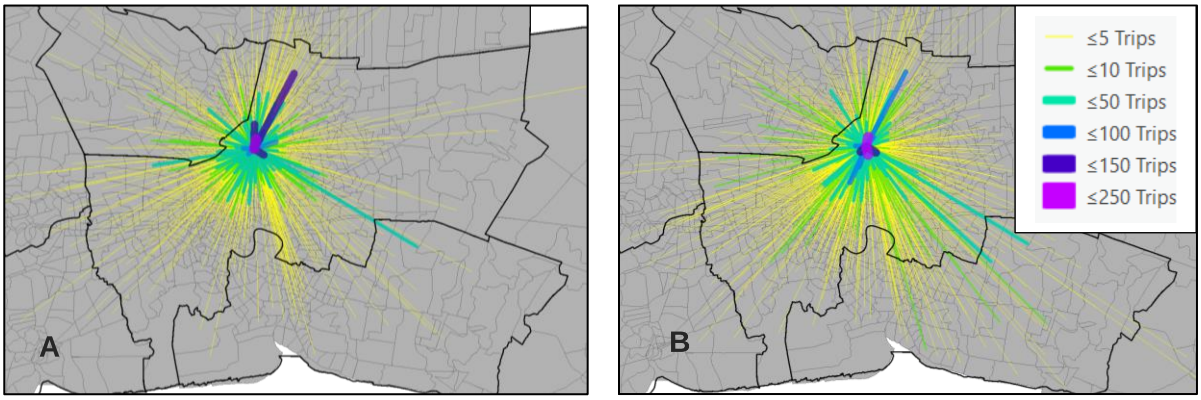


Figure 17 The trips from TAZ zone 231 to other zones (A). The trips to TAZ zone 231 from other zones(B).

4.3 Taxi and electric rail mode connectivity

For the connectivity from taxi get in and get off locations to electric rail BTS, MRT, and ARL station locations, the spatial relationship functions are used to search for the get in and get off locations that are within 100 meters from these stations. The below table shows top ten BTS and MRT stations that have the highest number of connections between taxi trips and BTS and MRT. BTS Bangwa has the highest connection with the taxi trips. The Get-In Taxi column shows the number of taxi trips that have the station location as the origin. On the other hand, the Get-Off taxi column means the number of taxi trips that have the station location as the destination. The ratios between get in taxi and get off taxi is very close.

Table 6 The top ten stations that connect to the taxi get in and get off locations

Rank	Station Name	Get-In Taxi	Get-Off Taxi
1	BTS Bangwa (E12)	5,826	5,817
2	BTS Prakanong (N8)	4,036	4,037
3	BTS On-nut (E9)	3,873	3,882
4	MRT Chatuchak (CHA) (Blue Line)	3,768	3,765
5	BTS Bearing (E14)	3,424	3,424
6	BTS Siam (CS)	3,375	3,385
7	BTS Saladaeng (S2)	3,121	3,130
8	BTS Asoke (E4)	2,693	2,697
9	BTS Victory monument (N3)	2,621	2,624
10	MRT Praram 9 (RAM) (Blue Line)	2,503	2,506

The below map shows the get off taxi locations and get in taxi locations within 100 meters from the location of BTS Prakanong.

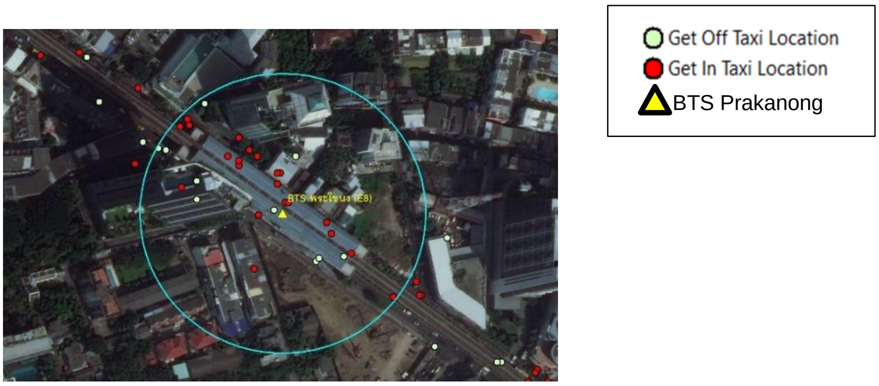


Figure 18 The get off taxi and get in taxi locations within 100 meters around BTS Prakanong on 01/07/2015

5. CONCLUSION

The taxi GPS probe data contains the detail location and other attributes of all taxi trips. It is in the form of spatio-temporal data that can be used for constructing each taxi trajectory and trips based on the specific time period such as morning peak period or daily period. Moreover, the characteristics of the traffic can be derived such as driving speed. Hadoop Hive is an efficient tool for processing the big taxi GPS probe data. The occupied taxi trips can be derived from the data set. When spatially joining with the traffic analysis zones layer, the origin and destination matrix of the taxi passengers, that reveals the human mobility, can be produced. The major taxi origins and destinations zones indicate that a lot of passengers traveled to and from Donmuang airport, Suvannabhumi airport, and BTS and MRT stations for their traveling purpose. Donmuang airport and Suvannabhumi airport has both domestic and international flights so that they become both origin and destination for Bankokians and foreigners. While the BTS and MRT stations are the important origin and destination especially the stations in the CBD or interchange between the electric rail system. The electric rail systems are the major public commuters for Bangkokians to travel from home to workplace in the morning and from workplace to their home in the evening. From the taxi GPS probe data set and the electric rail stations' location, the connectivity between taxi and electric rail commuters can be gained by using the taxi get-in and get-off locations and the electric rail stations. This connectivity could help the transport planner for preparing the sufficient infrastructure such as taxi parking area for each electric rail station to avoid the traffic jam problem in the morning and evening peak period.

6. REFERENCES

Tangchonlatip, K., 2007. Bangkok, the unlimited primate city of Thailand. In IPSR 2007 Annual Conference III : Urbanization and Urbanism, (pp. 1–19).

Kasikorn Research Center, 2016. Critical Traffic Problem - The impact on the economy and Bangkokians' lives.

DLT, 2015. Department of Land Transport's announcement of the type of public transport vehicles that must be equipped with GPS before registration, Retrieved August 8, 2019, from http://gps.dlt.go.th/?page_id=136.