Travel pattern in pp

1 Data cleaning method & result

1.1 Original data (notes for myself)

The GPS data collect location information of users in Phnom Penh in second week of each month in 2023, including time, latitude and longitude. As shown in the data sample, each row is the location information of a user at that time. The location sequence based on time order can reflect the user's activity trajectory.

Table 1 Data sample

	User ID	Time	Longitude	Latitude
1	ZnB1dG5mZXY2c3MwZDo0Z2syMGwxZTFpaHFv	1675510497	104.84183	11.57090
2	ZnB1dG5mZXY2c3MwZDo0Z2syMGwxZTFpaHFv	1675510505	104.84189	11.57092
3	NWpwZGFqN2tsOW9sdjpiZzVrbWp2Z2Uwa2c0	1675630030	104.83921	11.56933
4	bDV0bDJqcmhycDZwOjU5MTdpNWNhdG5mMHA=	1675500788	104.83402	11.57049
	……	……	……	……

Note 1: merge tables to get complete data

Location information was collected in tables for each day. For each month there are 7 tables (7 days of the second week), but each daily table contains data of adjacent days (Table 2). Thus, these tables should be merged as one table by month for further analysis for two reasons: (1) obtain complete data for the day; (2) some trips may span one day (from the evening of the current day to the early morning of the next day), identify trips based on merged table can better ensure accuracy.

Table 2 Data distribution of table for Feb 06.

Date	UNIX time	Amount
Mon Jan 23 2023	(1674406800, 1674493200]	1
Tue Jan 24 2023	(1674493200, 1674579600]	1
Wed Jan 25 2023	(1674579600, 1674666000]	0
Thu Jan 26 2023	(1674666000, 1674752400]	0
Fri Jan 27 2023	(1674752400, 1674838800]	0
Sat Jan 28 2023	(1674838800, 1674925200]	0
Sun Jan 29 2023	(1674925200, 1675011600]	0
Mon Jan 30 2023	(1675011600, 1675098000]	17897
Tue Jan 31 2023	(1675098000, 1675184400]	271811
Wed Feb 01 2023	(1675184400, 1675270800]	982701
Thu Feb 02 2023	(1675270800, 1675357200]	441234
Fri Feb 03 2023	(1675357200, 1675443600]	18714
Sat Feb 04 2023	(1675443600, 1675530000]	1276
Sun Feb 05 2023	(1675530000, 1675616400]	6782
Mon Feb 06 2023	(1675616400, 1675702800]	1271925
Tue Feb 07 2023	(1675702800, 1675789200]	194196

Note 2: cannot compare trips directly

Data amount of each day is shown in table 3. Most of users only have records in specific date, so cannot compare how personal travel pattern changed on different dates. The amount of each day can vary greatly, so the output (number of trips) of different dates needs to be divided by number of users to compare.

Table 3 Data amount of each date.

Month	Type	Date							Total
		Mon	Tue	Wed	Thu	Fri	Sat	Sun	
Jan	User	10983	11915	11984	11916	11798	11610	11453	19380
	Record	1971097	2157194	1987763	1894967	1973320	1330997	1657994	12973332
Feb	User	19010	20155	19832	14512	9734	9364	9606	31576
	Record	5661654	5701100	4724580	1464633	1138701	1160580	1190441	21041689
March	User	18892	19622	14718	9993	9625	9560	9461	27078
	Record	6237693	5723713	2133063	797553	1093596	998223	994131	17977972
Apr	User	22385	28494	28154	20399	16890	16765	14789	55091
	Record	5665799	5548164	4529120	2836024	1532284	1257130	928042	22296563
May	User	26943	29904	29801	25233	22626	22625	22143	55027
	Record	5743800	5756855	4911745	3216987	2509110	2470383	2120166	26729046
Jun	User	29737	32715	27882	23927	24139	23703	23097	56419
	Record	5361781	5211339	3480383	2803646	2288482	2013589	1952655	23111875
Jul	User	30345	32780	33309	28909	24750	24612	24495	61094
	Record	4246960	4811559	3981153	2985514	2631634	2746344	2615358	24018522
Aug	User	21997	19926	13564	13621	10043	1737	530	61192
	Record	7908934	4733199	2959232	1445743	433344	310017	19668	17810137
Sep	User	32024	31780	37014	36018	27888	27106	17395	73377
	Record	6384270	5841244	6855977	5033447	3035981	2520095	440639	30111653
Oct	User	63640	56671	55846	45794	45732	45384	34451	153241
	Record	11855337	10575542	9312588	6299827	3270058	2538861	604271	44456484

Note 3: Study time selection

For the merged monthly tables, we can find out that each month they actually gave us first two weeks data. Table 4 shows the data of August, it is not the second week that has the biggest amount of data highest and data quality, but the middle 7 consecutive days of the dataset. So, to ensure the accuracy and rate of identification, we should choose the middle days for each month to analysis and try to avoid public holiday.

Table 4 Data of each day in August table.

31/07, Mon	01/08	02/08	03/08	04/08	05/08	06/08
21,355	1,068,600	3,287,747	7,723,083	10,735,706	10,067,898	7,497,923
07/08, Mon	08/08	09/08	10/08	11/08	12/08	13/08
7,908,934	4,733,199	2,959,232	1,445,743	433,344	310,017	19,668

The specific selected research date base on original data volume and trip identification rate is shown in Table 5. For January, February, May, June and July, the second week of these months is still selected because original data volume in these days is significantly higher than other days in the month, and the trip identification rate is also the highest. For March and April, there were some dates in the second week where the original data volume dropped sharply and the trip identification rate was low, so the first week was used as a partial replacement for those dates. For august, September and October, 7 consecutive days with the highest original data volume and highest travel identification rate are selected.

Table 5 Selected study time of each month.

Jan	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…

Feb	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
Mar	1	2	3	4	5	6	7	***8***	9	10	11	12	13	14	15	16	17	18	19	20	…
Apr	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
May	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
Jun	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	***18***	19	20	…
Jul	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
Aug	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
Sep	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
Oct	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
Nov	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
Dec	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	…
* Pink grid is selected date; yellow grid is Monday.																					

1.2 Data cleaning method

Two major anomaly problems in the original data need to be cleaned: (1) redundant data, including users with only one record and users recorded in two locations at the same time; (2) drifting data, records that do not conform to the driving rules and clearly deviate from the driving trajectory.

For redundant data, users with only one data cannot form activity trajectory, so delete the data. users recorded two locations at the same time usually happens when user was in the middle of two location, sense the accuracy of the dataset is 1 meter, can keep only the first record. Meanwhile, after merging all the tables for each week, there are some repeat records, only keep one of these data.

For drifting data, use “traj_clean_drift” function in python package TransBigData to identify and delete drift data. The function defines drift data by speed, distance and angle limit. Sense there is no metro in pp, the speed limit here is 120km/h, distance limit is 1000m by default. Drifting record usually is clearly off trajectory, thus the angle of drift point with the previous point and the next point is usually very small. The angle limit is set 30 by default.

After data cleaning, valid data amount of each date is shown in table 6.

~~Table 6 Valid data amount of each date after data cleaning. Haven’t finished yet~~

Month	Type	Date							Total
		Mon	Tue	Wed	Thu	Fri	Sat	Sun	
Jan	User	10889	11787	11849	11771	11659	11347	11115	18725
	Record	1876502	2071336	1931959	1847234	1922391	1304465	1613347	12567234
Feb	User	18861	19947	19415	14185	9582	9188	9356	30710
	Record	5222250	5296713	4451734	1419558	1115535	1134826	1164372	19804988
Mar	User	18745	19308	14509	9881	9540	9394	9251	26394
	Record	5633902	5244979	2035138	781600	1069050	978770	972539	16715978
Apr	User	21964	27654	27244	19838	16452	16090	13761	52403
	Record	5069828	5021434	4093159	2576169	1474572	1228740	905794	20369696
May	User	26524	29336	28963	24573	22021	21880	20997	51861
	Record	5315246	5374649	4672600	3092871	2459801	2419796	2084212	25419175
Jun	User	29149	31795	27025	23256	23428	22835	21961	52746
	Record	4960912	4864850	3287825	2702275	2249444	1986122	1922958	21974386
Jul	User	29893	32349	31857	27841	24085	23796	23322	57373
	Record	4092673	4671449	3885713	2934669	2594211	2704285	2573810	23456810
Aug	User	21794	19797	13143	12801	8577	1626	457	58654
	Record	6202487	3374569	2225974	1237068	402500	298681	18544	13759823

Sep	User	31400	30930	36767	35034	26368	24778	15884	68483
	Record	5946366	5542970	6517710	4885186	2984489	2477766	421263	28775750
Oct	User	60451	54167	52615	44849	43898	38065	26702	135670
	Record	10578877	9491485	8608801	5977036	3183274	2487628	563052	40890153

1.3 Trip identification method

Considering there might be trips across 12:00 midnight, trips of the date are defined as trips started within this day. A trip is defined as the movement in space from one stopover or activity to another. Therefore, the user's continuous movement in the same place or within the same range is not considered a trip, such as shopping in a mall, moving at home, etc. Moreover, trips are considered to last for a certain duration and distance, and short-term stops during travel, such as waiting for traffic lights, asking for directions and looking at maps, are not considered stops.

Based on the definition of trip, trip identification is performed by four steps(Fang *et al.*, 2018): (1) convert location records into segments; (2) identify status of segments; (3) merge adjacent segments of the same status; (4) adjust status of short trajectories and merge into trips.

(1) Convert location records into segments

GPS data consists of a series of user's location records, arranged in chronological order to reflect how user moved in the city. The first step is to convert point records into segments by join every adjacent points pair. Each segment is associated with moving attributes such as distance, duration and speed.

(2) Identify status of segments

Then, status of each segment can be identified as moving or stopping by comparing segment speed with speed threshold. According to Wang and Li, the range of walking speed is 0.83–1.67 m/s. Since the straight-line distance of segments is usually lower than the actual distance along road network (thus the avg speed will be lower than actual speed along roads), we define the minimum moving speed as 0.5m/s. If segment speed is higher than the 0.5m/s, the segment is considered moving, otherwise it is stopping.

(3) Merge adjacent segments of the same status

Adjacent segments may be identified as the same status, they may appear as a trajectory or a cluster of stopping segments. Thus, we need to merge all the adjacent segments with the same status into one trajectory, until the status of adjacent trajectories is distinct from one another.

(4) Adjust status of short trajectories

Based on the definition of trips, both of moving or stopping status should last for a certain duration or distance. After testing, we define time and distance thresholds of stopping as 3min and 100m, which means stop for less than 3 min and 100m is considered temporary stop during one trip. Thus the status of such trajectory should be adjust into moving. For moving trajectories, time and distance thresholds are defined as 30s and 100m, and trajectories with time and distance less than this threshold will be modified to the opposite status.

Moreover, A trip is defined as the movement in space from one stopover or activity to another, which means people usually don't take long detours to get to their destinations. Thus, we introduced detour index, the ratio of the trajectory distance to the OD straight-line distance, to further filter trips. Through this index, continuous movement within a certain range, such as shopping back and forth in a business district, will be filtered out.

After completing this adjustment process, the status of some adjacent trajectories will be the same again. Therefore, we need to merge those short trajectories with the same status as we did in substep 1.3.3, until we get alternating stopping and moving trips.

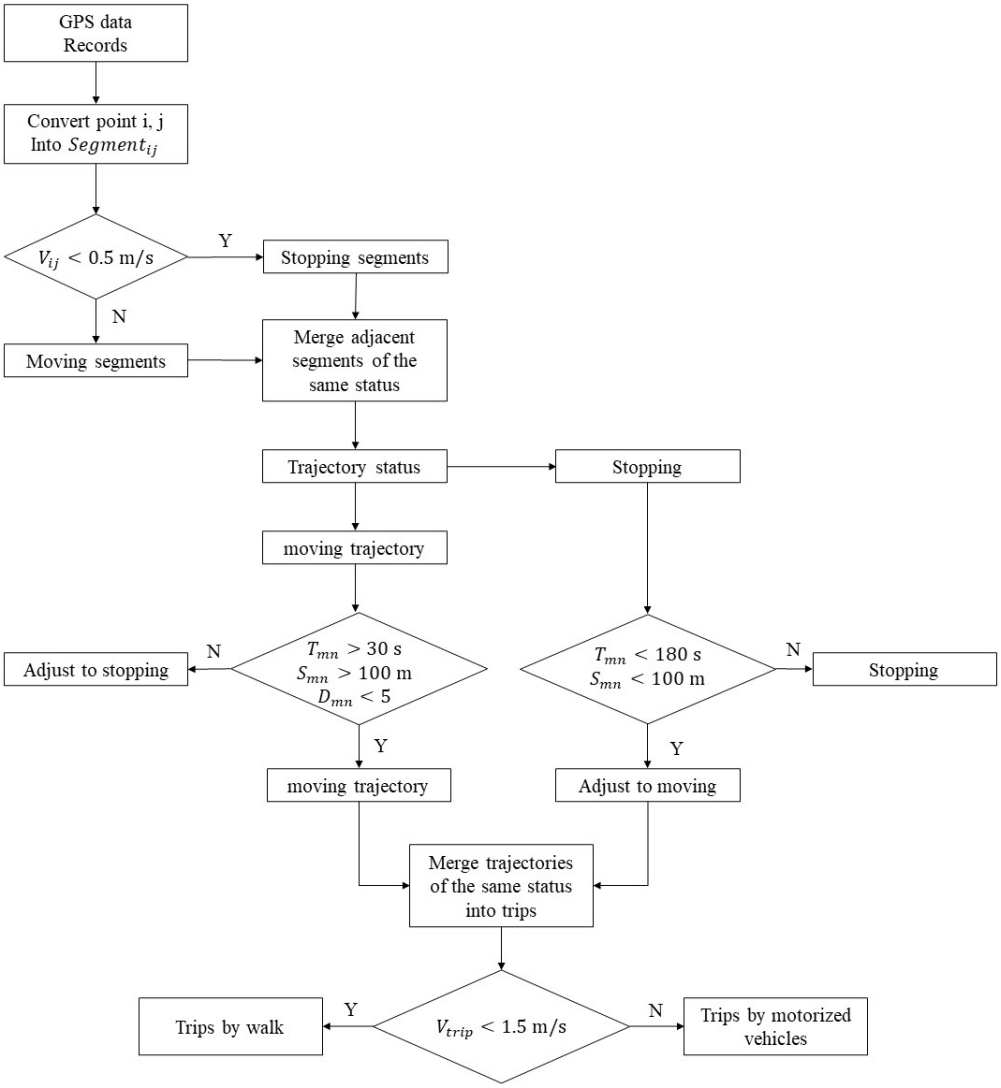

Figure 1 Trip identification method

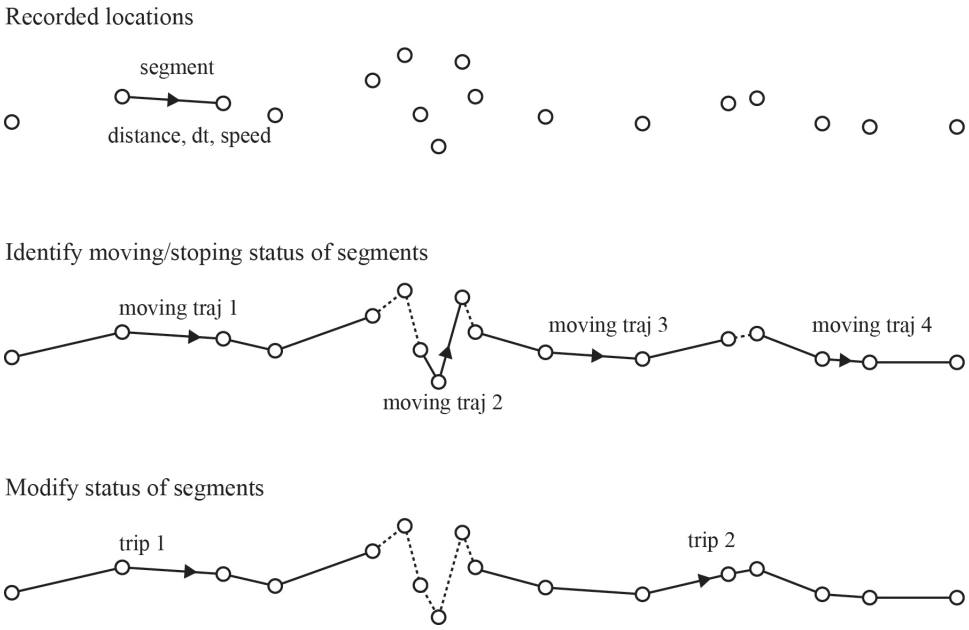

Figure 2 Trip identification method

1.4 Trips identification result

The result of trip identification is shown in Table5. The number of trips identified is affected not only by travel behaviour of residents, but also by original data amount (number of users and records) and identification rate (affected by how many records valid per person and per trip). In order to eliminate the impact of the original data amount on the results, the average trip amount per 10,000 people is shown in Table 6 to compare the difference between days and months.

There are two key findings from the tables: (1) trips by motorized vehicles are approximately three times as walking; (2) people travel more during weekdays than weekends. (From Table6, we can find average trips per person decreased from January to March. Not sure it is because of travel behaviour changed or because of data.)

Table 7 Trips per 10000 people.

Month	Mode	Date							Total
		Mon	Tue	Wed	Thu	Fri	Sat	Sun	
Jan	Walking	7220	7060	7077	6824	6921	5729	5913	46743
	Motorized vehicles	24631	23774	24255	23528	25259	19169	21114	161730
Feb	Walking	7641	8119	7705	4228	5552	5073	5878	44195
	Motorized vehicles	19978	19605	20222	13667	19411	19450	21534	133867
March	Walking	6530	6306	3994	4129	5018	4719	4768	35464
	Motorized vehicles	16508	15723	14481	13434	18435	17474	17345	113401
May	Walking								
	Motorized vehicles								
June	Walking								
	Motorized vehicles								
July	Walking								
	Motorized vehicles								

Average time and distance of two trip method during weekdays and weekends are shown in Table 9 and Table 10. For trips by walking, average distance is between 650 to 800 meters, and average time is between 11 to 14 minutes. Since trips by walking is mainly for short distance, the median distance and time are smaller than the average. The median distance is between 400-550 meters and the time is between 7-10 minutes. There is no significant difference in walking distance and time between

weekends and weekdays. However, travel time and distance in rainy season (May to October) are lower than in dry season (November to next April), with distance reduced by about 100-150 meters and time shortened by about 1 to 2 minutes. **(check it again after the November and December data comes out.)**

For trips by motorized vehicles, average distance is above 3000 meters, and time above 11min. The median distance and time are lower than the average by about 1000 meters and 3-4 minutes. In dry season, travel distance and time during weekends are obviously higher than during weekday. However, in rainy season there is no significant difference between weekday and weekends **(Perhaps residents' interest in mid- to long-distance weekend trips were affected by the rainy season, or maybe due to the decrease in tourists during rainy season).** In addition, travel distance and time in rainy season are slightly shorter than in dry season, with the average distance being about 500-1000 meters lower and the median distance being about 500 meters lower.

Table 8 Average and median time and distance of trips by walking.

Month	Date	Average		Median	
		Time / min	Distance / m	Time / min	Distance / m
Jan	Weekdays	13.33	783.24	9.75	522.49
	Weekends	13.61	792.21	9.87	516.92
Feb	Weekdays	12.90	748.10	9.33	488.94
	Weekends	13.26	766.72	9.65	514.33
Mar	Weekdays	11.74	669.42	8.73	442.57
	Weekends	14.88	865.89	10.87	605.53
Apr	Weekdays	12.48	719.43	9.08	458.27
	Weekends	14.48	840.03	10.47	559.18
May	Weekdays	12.61	727.94	9.30	478.09
	Weekends	13.04	754.61	9.68	497.95
Jun	Weekdays	13.65	793.77	9.98	513.13
	Weekends	13.54	782.51	10.03	511.09
Jul	Weekdays	12.96	752.99	9.45	487.99
	Weekends	12.41	715.26	9.13	475.14
Aug	Weekdays	11.35	634.60	8.47	431.71
	Weekends	10.35	574.17	7.68	399.57
Sep	Weekdays	11.26	654.70	7.70	428.40
	Weekends	11.01	633.07	7.58	422.95
Oct	Weekdays	11.29	645.81	7.73	431.23
	Weekends	11.34	649.87	7.75	431.38
Nov	Weekdays				
	Weekends				
Dec	Weekdays				
	Weekends				

Table 9 Average and median time and distance of trips by motorized vehicles.

Month	Date	Average		Median	
		Time / min	Distance / m	Time / min	Distance / m
Jan	Weekdays	12.97	3751.66	9.38	2579.40
	Weekends	13.85	4201.82	10.13	2892.97
Feb	Weekdays	12.56	3458.81	9.07	2352.19
	Weekends	12.99	4132.85	9.50	2933.12
Mar	Weekdays	10.95	3110.80	7.98	2006.21
	Weekends	12.99	3416.37	9.55	2469.98
Apr	Weekdays	10.82	2602.72	7.77	1861.92
	Weekends	12.49	3187.62	9.40	2197.86
May	Weekdays	12.64	2937.70	9.42	2010.27
	Weekends	13.14	3206.31	9.90	2126.41
Jun	Weekdays	13.65	3003.54	10.20	2049.53
	Weekends	13.89	2970.52	10.52	2037.64
Jul	Weekdays	13.11	2910.94	9.68	1977.51

	Weekends	12.83	2831.22	9.65	1999.00
Aug	Weekdays	8.71	2468.05	6.68	1637.68
	Weekends	8.09	2006.24	6.18	1461.51
Sep	Weekdays	12.15	2914.01	8.27	1978.79
	Weekends	13.27	3281.51	9.38	2126.03
Oct	Weekdays	10.99	2953.80	7.48	2071.75
	Weekends	10.99	3112.42	7.58	2146.25
Nov	Weekdays				
	Weekends				
Dec	Weekdays				
	Weekends				

2 Temporal distribution of trips

Plot number of trips identified every day in line chart, with the x axis being time, and the y axis being the number of trips departed within that duration. The number of trips identified (blue line in chart) each day differs greatly due to the large difference in the amount of original data per day. In order to eliminate the impact of the original data amount on the results, the average travel volume per 10,000 people (red dash line in chart) is also shown in the chart to compare the difference between days and months.

In terms of temporal dynamic of daily trips, the distribution of trip time shows the similar trend and pattern. During weekday, from 0 to 6 AM is the time when the number of trips was lowest each day. After 6 AM, there were two obvious rush hours, morning rush hours from 7 to 10 AM and evening rush hours from 5 to 8 PM. The number of trips dropped slightly after the morning rush hour and picked up during lunch time. In addition, the travel volume during the evening rush hours was significantly higher than that during the morning rush hours. After evening rush hours, travel volume showed a nearly linear downward trend until 12 PM. On weekends, travel volume increased slowly after dawn and remained stable after 10 AM. There was no obvious morning travel peak, but an evening peak from 5 to 8 PM.

Compare travel distribution by different modes

Compare travel volume in three months

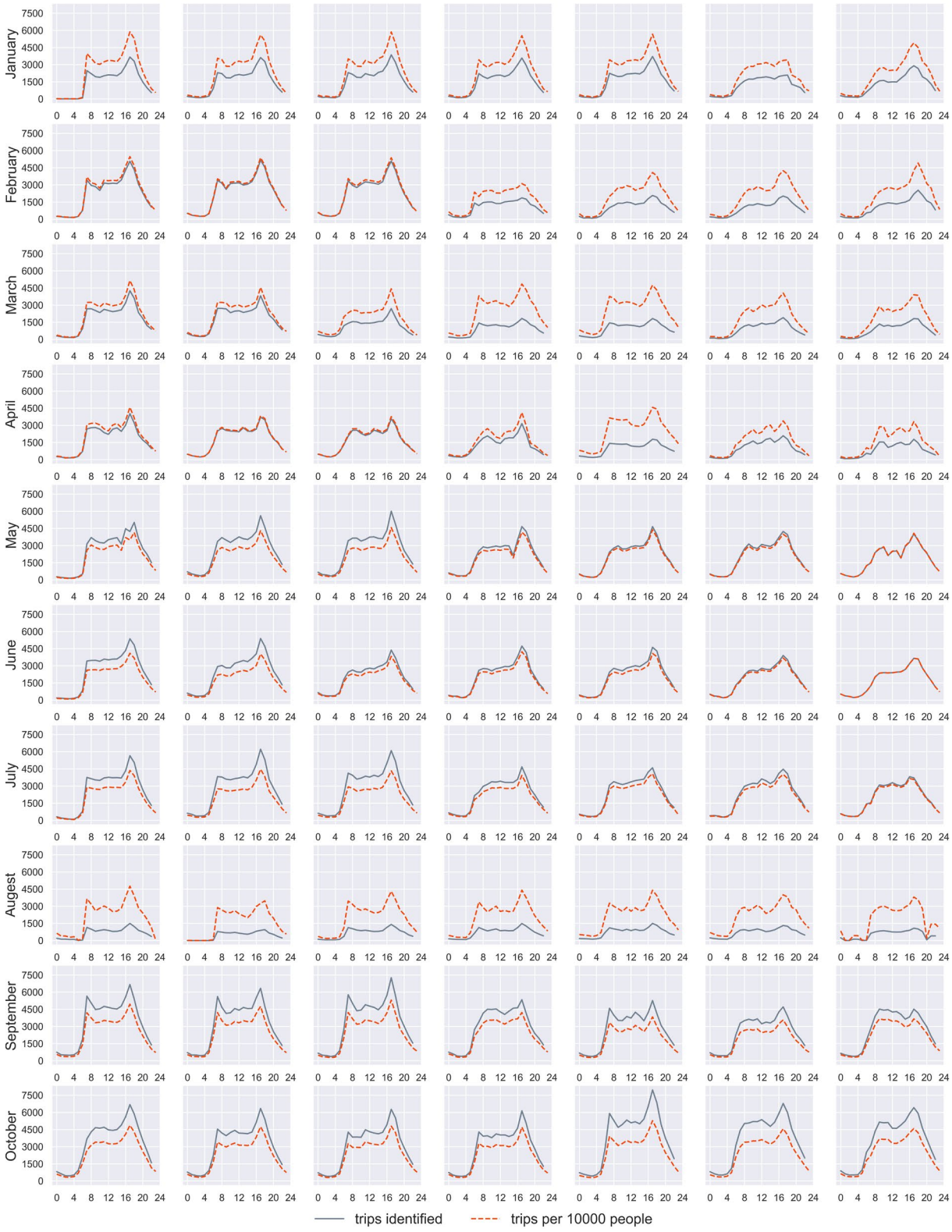

Figure 3 Temporal distribution of trips

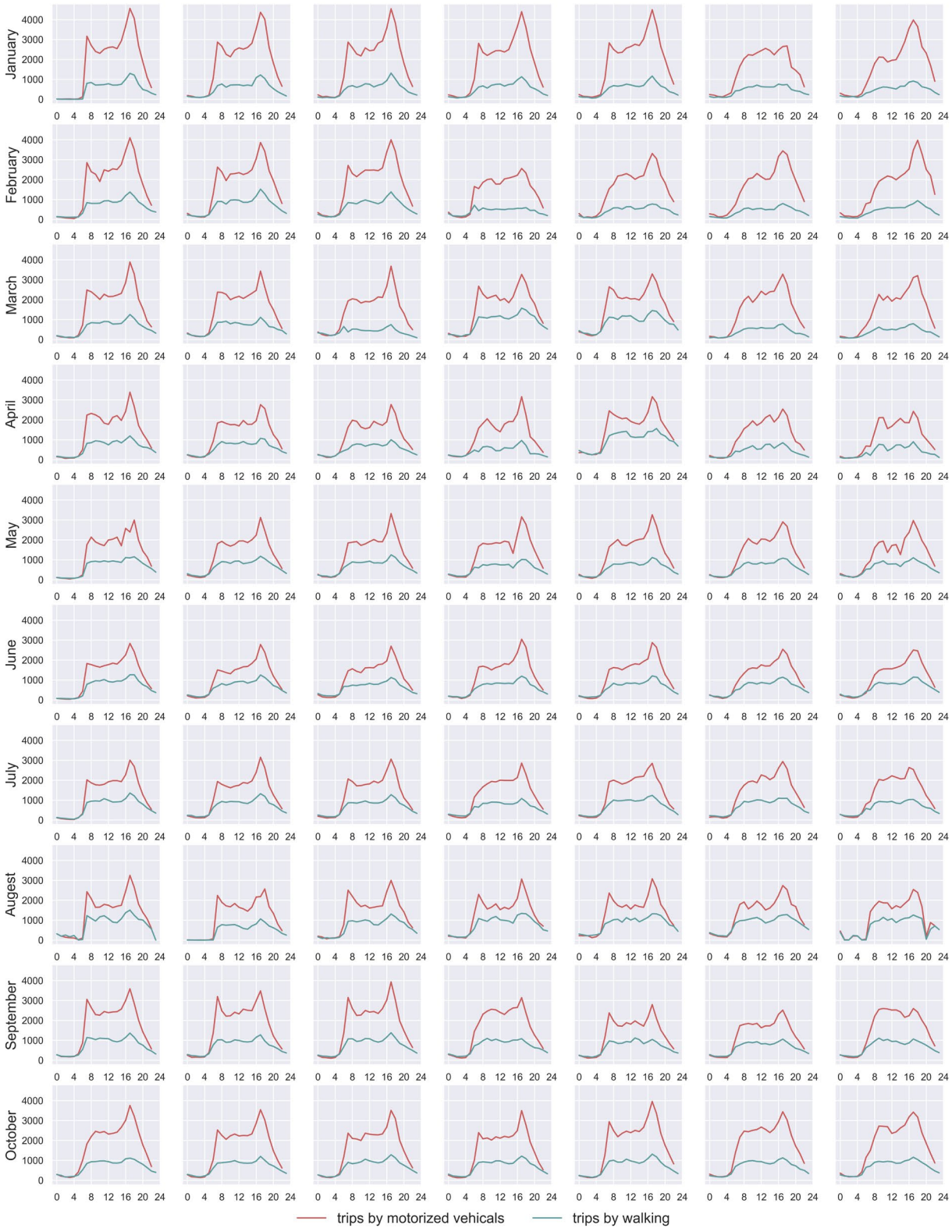

Figure 4 Temporal distribution of trips by different modes

3 Spatial distribution of trips

3.1 Traffic analysis zone

72 TAZs (Traffic Analysis Zones) in Phnom Penh are defined as shown in Figure 4. TAZs boundaries are defined based on the Sangkat (commune) boundaries. There are 105 Sangkat in Phnom Penh, and those in suburban area are much larger than those in city centre. Thus, to better understand how

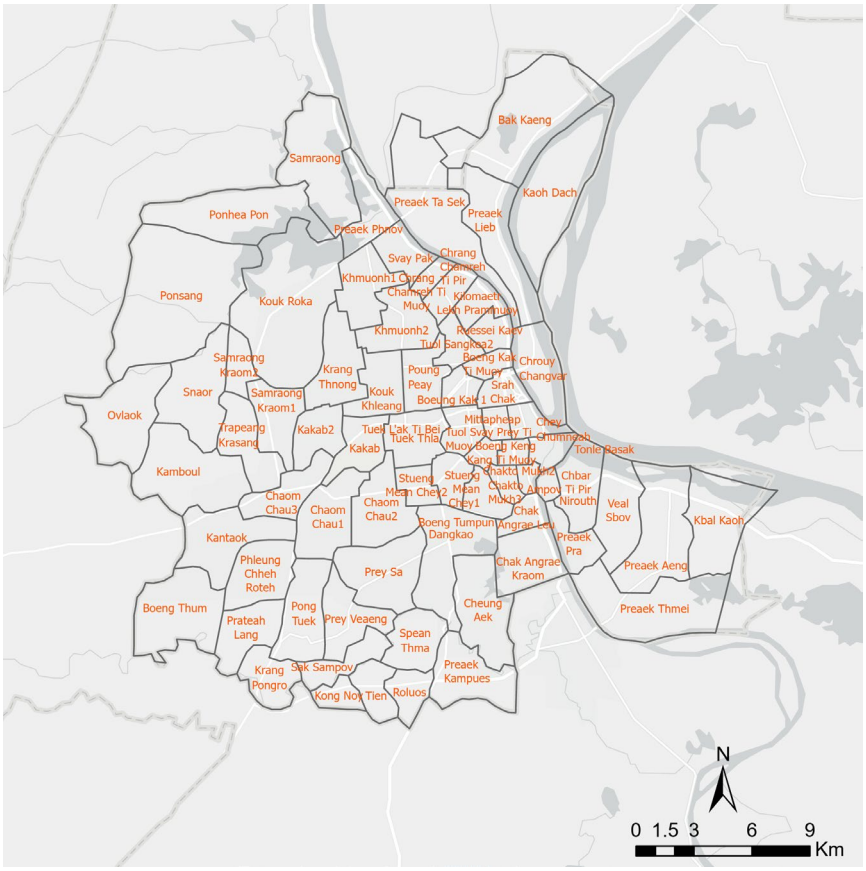

Figure 5 Traffic analysis zone based on commune.

3.2 OD of trips identified

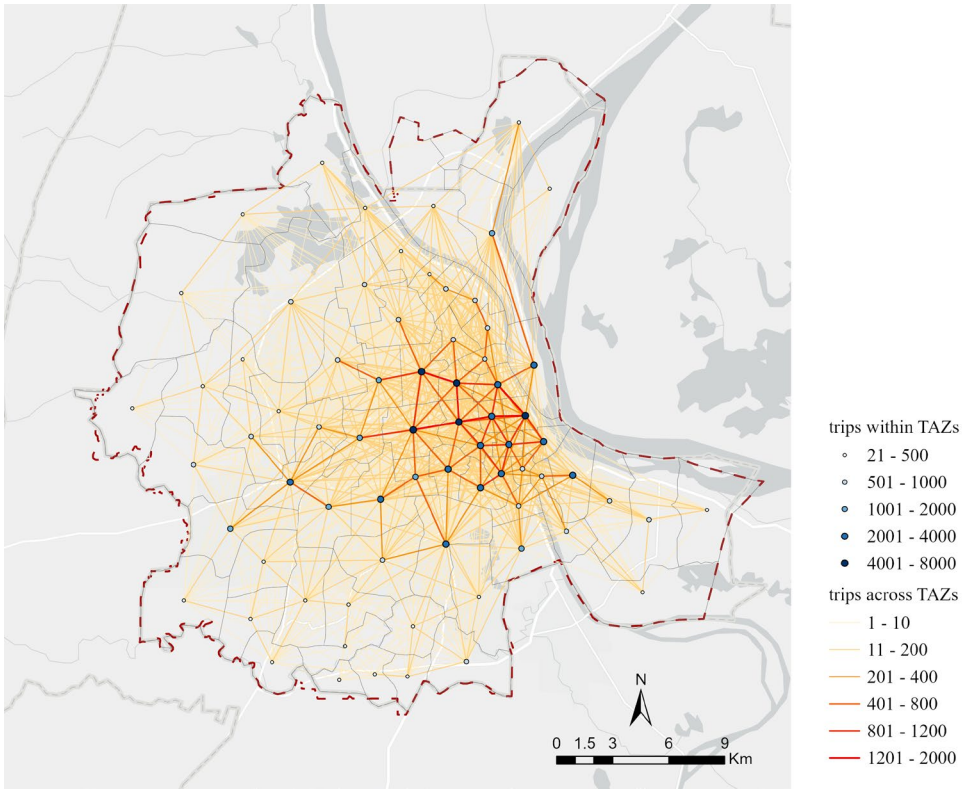

Figure 6 OD of trips identified in January.

3.3 Hotpot of OD

3.4 Street volume

Figure 7 Street Volume of motorized vehicles in July

Figure 8 Street Volume of walking in July

Reference:

Fang, Z., Jian-yu, L., Jin-jun, T., Xiao, W. & Fei, G. (2018) 'Identifying activities and trips with GPS data', *IET Intelligent Transport Systems*, 12, 884–890.