EVALUATING TRIP CHARACTERISTICS FROM CALL DETAIL RECORD USING MACHINE LEARNING

携帯電話記録(CDR)を活用した機械学習による人流特性の把握

ダス ルベル*・中村 茂**・ジャイン***
Rubel DAS, Shigeru NAKAMURA and KYAING

急速な都市化は、都市交通インフラへの過剰な負担を招いている。携帯電話事業者が保持する Call Detail Record (CDR)（通話時刻、通話時間、通話に使用した基地局などが含まれる）は、このインフラの使用状況を把握するものとして期待されているビッグデータである。近年、そこから有用な情報を抽出する研究が進んでいるが、トリップチェーンを考察した研究は未だ少ない。トリップチェーンとは人の活動軌跡であり、トリップの開始地点とその後の経由地点及び目的地の時系列である。本研究では CDR を用いて、機械学習の手法として用いられるマルコフ連鎖モデルによりトリップチェーンを解析するモデルを開発した。対象地域であるヤンゴンを 17 の地域に区分し地域間の移動を示すトリップチェーンを生成したところ、開発したモデルは実態を良く再現出来ることが分かった。

Keywords ： *CDR、隠れマルコフモデル、トリップ特性、ミャンマー、データマイニング*

1. INTRODUCTION:

The fast development of cities today makes it difficult for urban planners to handle all the emerging demands. For instance, the population of a town can be highly dynamic. The island of Manhattan, in New York City, is home to about 1.6 million inhabitants. This number, though, may roughly double during any workday. Such conditions have been causing traffic congestion to grow increasingly at temporal and spatial scales. Investing in transportation infrastructure seems to be an obvious solution to ease congestion. However, these investments did not evidently improve the performance of transportation networks, as expected, because new traffic facilities resulted in increased traffic demand. Since the budget limitation for infrastructure development is a concerning issue, promoting low-cost technology like public transportation (e.g., buses) has become essential. Not to mention that public transportation is comparatively more environment-friendly and contributes to the achievement of sustainable development goals (SDG).

Bus routing and operational attributes like frequency are planned based on Household Travel Surveys (HTS) or Personal Travel Surveys (PTS). HTS/PTSs were generally undertaken by employing retrospective or diary surveys, in which respondents either report their travel information in face-to-face/telephone interviews or record their travel information on paper- or internet-based questionnaires. For this reason, respondents need to remember the details of trips like start and end times, origin and destination, and travel modes. Consequently, there is a probability of underreporting trips that have a short distance or duration. To overcome this issue, in the late 1990s transportation planning researchers drew on Global Positioning System (GPS) to carry out the surveys and it rapidly became quite popular. However, a travel survey with GPS devices has the following limitations: (1) purchasing the devices can be expensive, (2) the collected data is usually incomplete because respondents often forget to take with them the devices and (3) devices must be distributed and then recovered for as long as a respondent participates in the survey, and (4) the sample size is limited by the number of available GPS devices. Since this survey method is expensive and challenging to implement, the data that was already being collected by increasingly more equipped mobile phones drew the attention of traffic researchers. During the last decades, there has been a surge in the amount of location data that

* コンサルティング事業統括本部 中央研究所 先端研究センター
** コンサルティング事業統括本部 中央研究所
*** Yangon Technological University, Department of civil eng.

has been generated by mobile big data (MBD). There are three main reasons for the surge. First, recent hardware development made the positional devices smaller, require less electric power, and consequently easier to embed in mobile phones. Second, The US government's "Enhanced 911" requirement has mandated US wireless carriers to provide accurate location estimates of emergency calls, accelerating the adoption of positioning technologies by mobile phone manufacturers. Third, widespread adoption of positioning hardware has catalyzed the development of location-aware software. For these reasons, it is no surprise how much nowadays transportation planning depends on MBD.

The application of Call Detail Record (CDR) for mobility planning is not a new practice since mobile positional data contains spatiotemporal stamps and can be generated without additional interventions. The challenge is how to handle such big data or how to understand people's behavior from this data. Another ongoing challenge is how to modify the traditional traffic planning models to make them able to handle big data. Asakura and Hato (2004) utilized mobile positional data for movement tracking and stay extraction while Bar-Gera (2007) estimated traffic speed. Candia et al. (2008) utilized CDR data to identify the detection of emergencies. Gonzalez et al. (2008) explored individual travel patterns from mobile positional data. The results of these studies showed that people have a high degree of temporal and spatial regularity; everyone is characterized by a travel distance with time-independent characteristics and a significant probability of return to a few highly frequented locations. Calabrese et al. (2013) validated the applicability of mobile phone data in transportation planning. They compared mobile phone data with vehicle odometer readings. Toole et al. (2015) estimated the travel demand from CDR data. Alexander et al. (2015) utilized mobile positional data for estimating origin-destination trips by purpose and time of day. The above-mentioned studies, however, did not consider trip chains explicitly. The exploration of trip chain is essential for various transportation planning for example bus route planning and bus frequency.

MBD enables the identification of the activity chain of people. Most of these movements are carried out in known places such as the home, the workplace, a relative's house, a favorite cinema, or a shopping mall. Hence, patterns of these movements can be studied, and a prediction system can be designed by storing all the destination places and anticipating the arrival. A variety of machine learning approaches (Kapicioglu, 2013) are used in existing literature, including decision trees, Bayesian networks, naïve Bayesian classifiers, fuzzy logic, hierarchical conditional random fields, discriminant function analysis, and support vector machine. Machine learning creates a hypothesis based on a given algorithm with a training dataset, and this hypothesis is set aside to be applied to future data. For example, a person can be detected in a place that was not a real destination but rather a stop in a trip. Therefore, an algorithm for identifying significant locations (hereafter referred to as the point of interest, "POI") is essential. In this study, we propose a new hidden Markov model (HMM) approach. As a classifier to cope with a sequence of data, an HMM is an appropriate approach for detecting trip chains. We utilized the same Myanmar CDR data as Lwin et al. (2018), who estimated the hourly link population, utilized. The raw CDR data was collected by the largest cell operator in Yangon. We received the raw data from Yangon Technological University which is our joint research partner related to Big Data analysis.

Our proposed HMM model can predict the probability of choosing the next POI given that the current location is known. Here, we like to emphasize that in a dimension of transportation planning and policymaking, it is meaningless to estimate the exact trip purpose of a specific individual. Rather, it is sufficient to know the probability that a traveler has a certain purpose under certain conditions. Moreover, people go from one place to another before returning to their origin. The probability could then be used to synthesize the sequence of activities for a trip chain. This model outcome could be a good input for the activity-based demand-forecasting model (Bowman and Ben-Akiva, 2001).

Here we propose an HMM with millions of geotagged data points as an input that performs many aspects of trip characteristics. We begin by outlining the system architecture in Section 2. In section 3 we explain our methods of extracting, cleaning, and clustering method of data. We discuss the results of our model in section 4. We summarize our study in section 5.

2. HIDDEN MARKOV MODEL (HMM):

Hidden Markov models (HMMs), named after the Russian mathematician Andrey Andreyevich Markov, who developed much of relevant statistical theory, is

introduced and studied in the early 1970s. Recently, HMM has become popular as a tool in machine learning. These models have been used successfully for different fields like speech recognition, character recognition, and mobile communication techniques. HMMs are statistical models to capture hidden information from observable sequential symbols. The CDR datasets comprise a long sequence of directional measurements that are observed by fixed location towers when the user makes calls or accesses a mobile network. The trip-chain system being modeled is assumed to be a Markov process with unknown parameters. Han and Sohn (2016) proposed a hidden Markov model for exploring trip-chain by using a smart card. Figure 1 shows the structure of HMM where $i, j, and\ k$ are traffic analysis zones (TAZs).

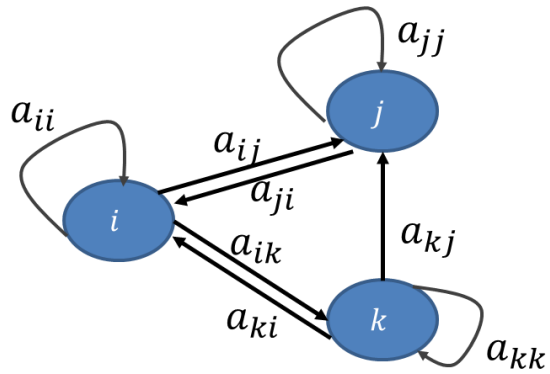

Fig. 1 Architecture of HMM

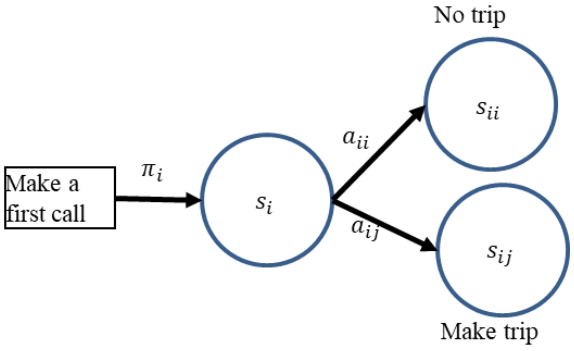

Fig. 2 Flowchart of trip pattern analysis

Here, a_{ij} represents the probability of moving from TAZ i to TAZ j. Note that a_{ji} does not need to be equal to a_{ij}. Besides, a_{ii} represents the probability that the person will not change the TAZ. The elements that formally define an HMM are (S, π, A) which must be adapted to our purpose. Figure 2 illustrates the linear formation of Figure 1. The model assigns each people to one of all possible locations from where trips originate depending on the time of the day.

- S represents the state of individuals. It contains either a single index (i.e., i) or more than one index (i.e, ij etc). Here $i, and\ j$ are the members of the set of N distinct TAZ in the model. Here, TAZs are the potential locations of a user.
- It is a logical supposition that a person may stay in the same location after making a call. Here, an individual's origin is presented by s_i. The person might decide one from two situations: (a) the person stay in same cluster and (b) the person moved to a different cluster. For the latter, the person may select a TAZ with a given probability. If a TAZ is not accessible to the person, the probability of that TAZ is zero.
- The initial state distribution is $\pi = \{\pi\}_{i=0}^{N}$. In our model, π_i denotes the probability that the individual at location s_i , when a trip is beginning. The most probable state is usually the home in the morning and the office in the evening.
- The transition model is the core part of the model. The state transition probability distribution is $A = \{a_{ij}\}$. The matrix contains the transition probabilities between each pair of states. The transition probabilities are computed from CDR data. However, these probabilities can also be presented through spatial properties of the TAZs environment (i.e., number of restaurants, distance, accessibility) that create attractions for people. For example, people will prefer the nearest shopping mall. We adopt the Gibbs distribution (Tripuraneni et al., 2015) as the transition model because it provides a means to incorporate spatial features without making any independent assumptions and generalizes the discrete version of simpler random walk models.

3. DATA MINING:

Telecommunication company collects CDR record mainly for billing purpose. To utilize such big data, we need to process it. Figure 3 shows the framework for data clustering and filtering of the CDR database.

(1) Filtering: We have utilized one week of data of CDR. Each day of the database contains about 10,970,000 call records before applying any filter technique. These large quantity records were produced by 1,517,300 users, known as “PID”. Many of them have accessed telecommunication towers only for a single time in a whole day. Since those records cannot be utilized to predict trips, we have removed those records. After removing those single recorded PIDs, the useful unique users become 1,216,696 users. With this step, 19.8% of users were removed from the original database. Further filtering technique was implemented based on model types (i.e.: morning HMM and evening HMM). Office going people move to office from home in the morning while from office to home in the evening. Therefore, people’s flows are in opposite directions during the morning and evening periods. Therefore, we have analyzed the HMM for two filtered data: one for morning HMM (6 am to 11.59 am) and another for evening HMM (2 pm to 7 pm). It allows us to observe the changes in the trip pattern of the day the time. Total unique users are 882,363 in the morning filtered data, while total unique users are 988,144 in the evening filtered data. For each model type, we have utilized 75% data for model building and the remaining data for model validation.

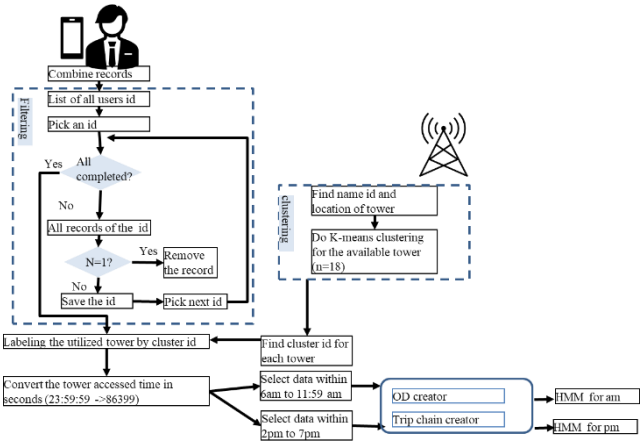

Fig. 3 Filtering and clustering of data

(2) Clustering the access points: The CDR data has 32 base stations. Some base stations serve very large areas where telecommunication network barriers like tall buildings are quite a few. Some base stations are closely positioned. Those places have higher population density or business activities. One feature of closely positioned base stations is that the connection of a mobile phone may be switched from one base station to another depending on the signal strength even though the mobile phone user may not travel a significant distance. Now, this telecommunication connectivity feature makes it difficult for computing the geographical location of the user. Moreover, such base station switching creates fictitious trips. To overcome such trips, we have clustered closely situated base stations. Here, we have utilized the k-means clustering algorithm. After doing several trial-and-error methods, we have selected 18 cluster points as shown in Figure 4. The maximum cluster diameter is 2.5 km. In our trip analysis, we utilize those cluster points as the traffic analysis zone (TAZ). The maximum and minimum values of the cluster id are 17 and 0 respectively. Figure 4 shows the original base station locations and the k-means center and label of each cluster point. The colors of base stations represent the membership of cluster zones.

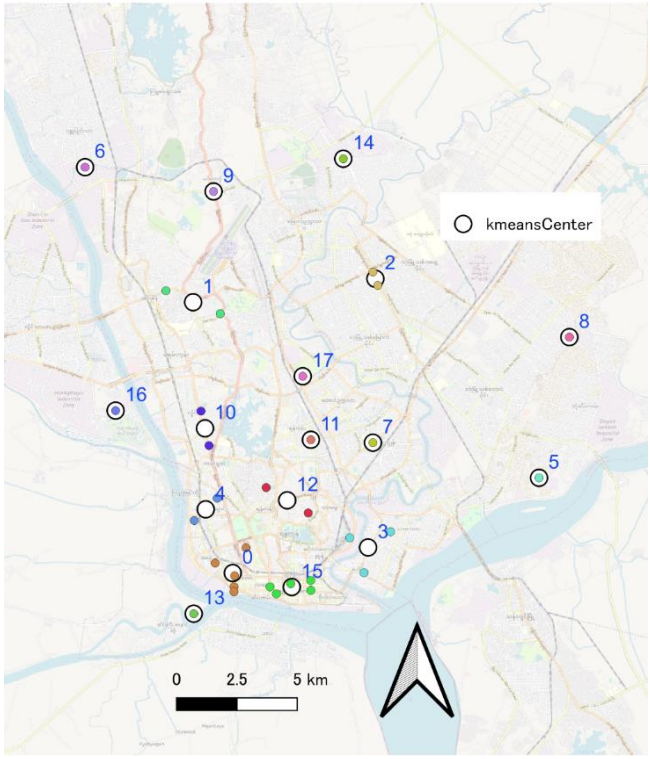

Fig. 4 clustering of base stations

(3) Converting access time: The CDR database has a base station access time format is “20151201143243”. Here the format structure is YYYYMMDDHHMMSS which is “year: month: day: hour: minute: second”. We have

converted the " hour : minute: second" portion into a 24-h seconds format where 23:59:59 becomes 86399. This time format conversion helps to do sorting or filtering of the whole database based on the access time. Thus, the identification of the sequence of geographic locations of a user becomes easier.

(4) OD creator and Trip chain creator: After filtering and converting access time, CDR data were rearranged for creating origin and destination (OD). In addition, intermediate points in between origin and destitation were also computed. Thus, trip chains were created. Finally, OD and trip chain were combined for making HMM for two distincs periods.

4. RESULTS:

First, we validated our model by using testing data set. Although the proposed methodology adopted an unsupervised machine learning tool that requires no calibration process based on labeled data, the result should be validated in a qualitative manner. We validated our model by comparing the number of potential trips from each zone. Though am model (Figure 5) shows variations, both am, and pm models (Figures 5 and 6) illustrate a similar trend to the original data. we can say that HMM of am and pm could predict trip origin with reasonable confidence.

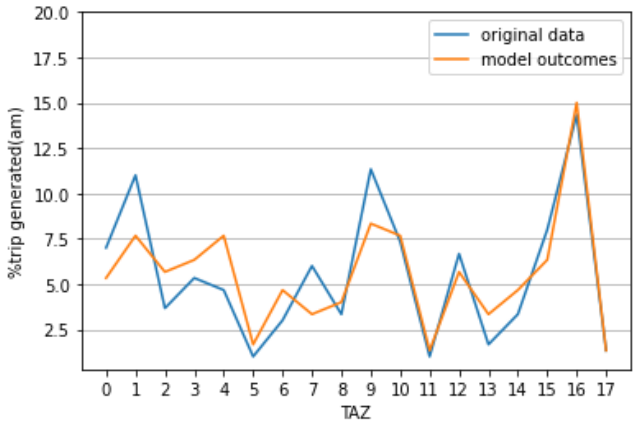

Fig. 5 Comparison of trip generation in the am model

(1) Probability of origin zone: The probability of the first call in the morning and evening periods is presented in Figure 7. People may make the first call from home before going to work or from a place after reaching there. Those locations were considered the origin of the trip. As expected, commercial regions (TAZ 15,0,3) in the pm period have a higher probability and residential zones (TAZ 13,5) have a lower probability of behaving as a trip origin. The implication of these findings is that TAZ 15,0 and 3 will create more trips in the evening. On the hand, TAZ 13 and 5 create a smaller number of trips in the evening period. Some zones (i.e. TAZ 1, 10, and 7) do not show a significant difference in the evening and morning periods. Those zones can be considered mixed zone.

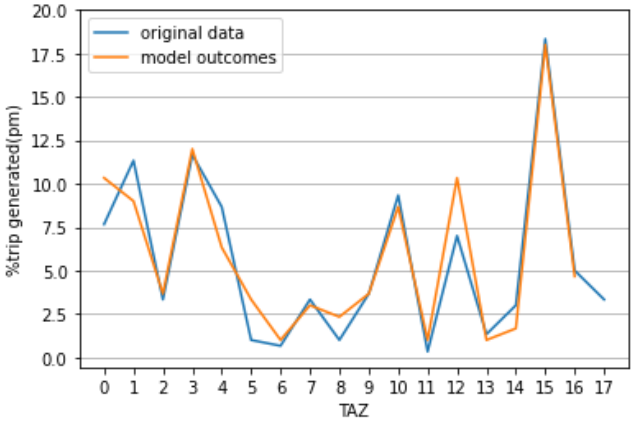

Fig. 6 Comparison of trip generation in the pm model

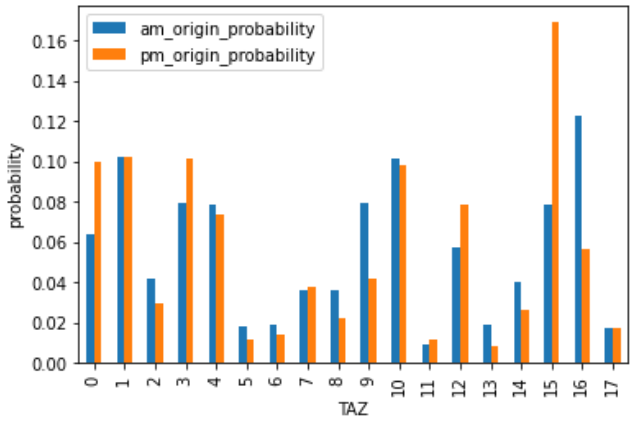

Fig. 7 Probability of making first call in certain TAZ

(2) Probability of staying in the same location: Figure 8 shows the probability of a person staying in the same TAZ within the time frame. The minimum value among all zones in the am model is 0.48 while in the pm is 0.35. Here, we like to highlight the properties of TAZ 0 which is in the central business district (CBD) where people create many calls. The a_{00} is 0.57 in the am model and 0.4 in the pm model. People making the first call in TAZ 0 have a higher probability to stay in the same zone in am model than in the pm model.

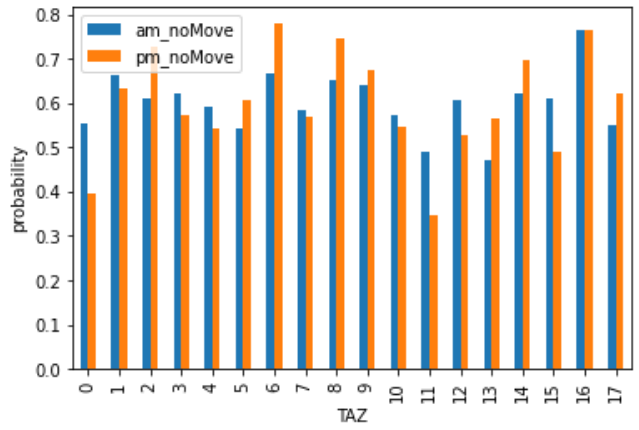

Fig. 8 a_{ii} for am and pm HMM mode

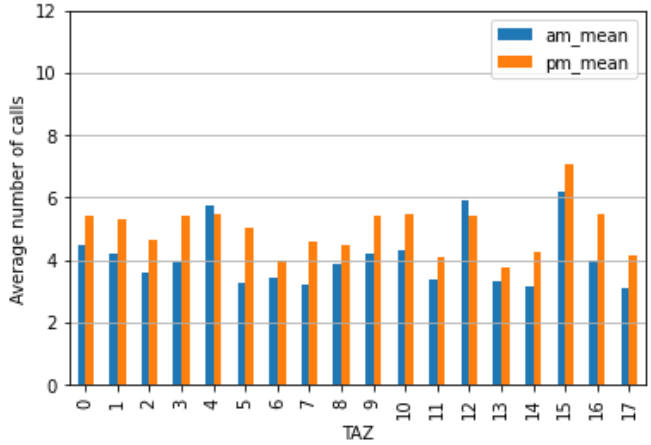

Fig. 9 Mean number of calls in each zone

(3) Differentiating CBD zones and non-CBD zones: The central business district (CBD) is the commercial and business center of a city. The needs and features of a CBD are quite different from other zones. For example, CBD areas need higher car parking capacity. Therefore, identifying CBD and non-CBD zones in a city can be beneficial. The average number of calls and standard deviation of the number of calls in each TAZ for each distinct period are presented in Figure 9 and 10. We can see that the mean call is larger in TAZ 0,4,12, and 15 in the morning period. These zones also have a higher standard deviation meaning people gather in those zones has higher variation considering the necessity of making calls. By considering the geographic features of Yangon city, we can see that TAZ 0, 4, 12, and 15 are situated in CBD areas. Therefore, we can conclude that the CBD area has a higher mean and standard deviation of calls in the morning period. In other words, we can distinguish CBD areas and non-CBD areas by observing the mean and standard deviation of the number of calls in the morning period. The calls in the evening period do not have any obvious trend. For example, TAZ 16 has a higher mean of calls and a relatively lower standard deviation. It represents that people in TAZ 16 do not have a large variety of characteristics and it can be a residential zone. However, we can say confidently that TAZ 15 is a CBD region. Another TAZs has no definite features. Therefore, we can conclude that Yangon is a city of mixed zones.

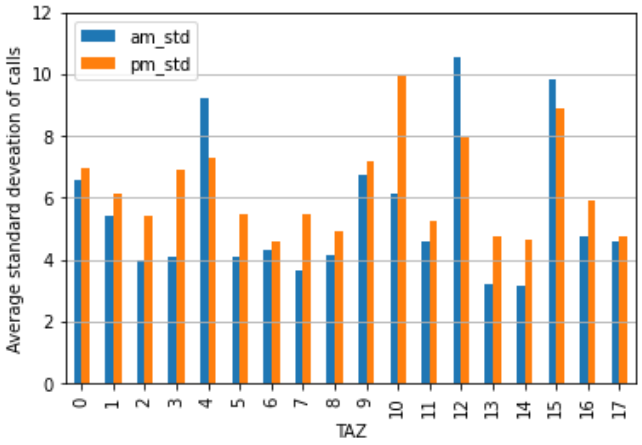

Fig. 10 Standard deviation of calls in each zone

(4) Trip distribution: The distribution graphs show that people often prefer closer TAZ. Figures 11 and 12 illustrate the trip distribution from three TAZ 14, 16, and 17. These three zones were selected randomly. The top two destinations from TAZ 14 are TAZ 1 and 9. The top two destinations for trip generator TAZ 16, are TAZ 1 and 10. TAZ 17 shows a bit different result. The top two destinations are TAZ 7 and 1 in the morning while TAZ 2 and 7 in the evening. Some people return to start TAZ after visiting different TAZ. For example, 20.2% of trips return to TAZ 14 after visiting different TAZ within the same period.

The trip distribution shows a different pattern in the morning and evening (comparing Figures 11 and 12). 17% of trips originated from TAZ 14 go to TAZ 1 in am while 26% of trips go in pm. 6% of trips during the morning period go to TAZ 10 but no trips in the pm go to TAZ 10. A significant variation can be observed in the destination TAZ 2 from TAZ 17. In the morning TAZ 2 attracts only 10% of trips while 27% of trips go to TAZ 2 during the evening. TAZ 7 and 11 are

geographically closer to TAZ 17 and do not have significant variations in am and pm models.

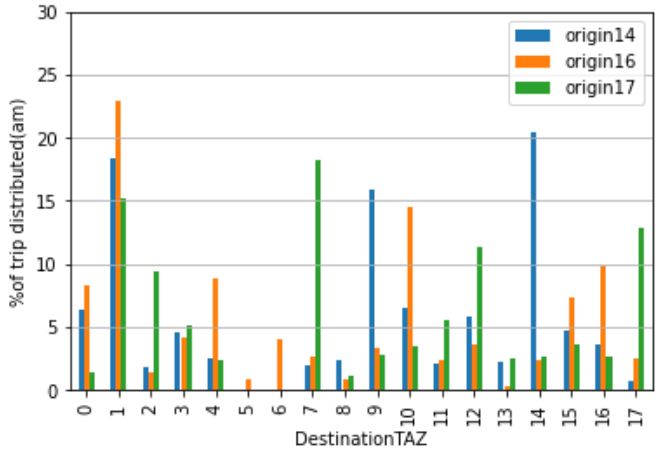

Fig. 11 Trip distribution for each origin in am model

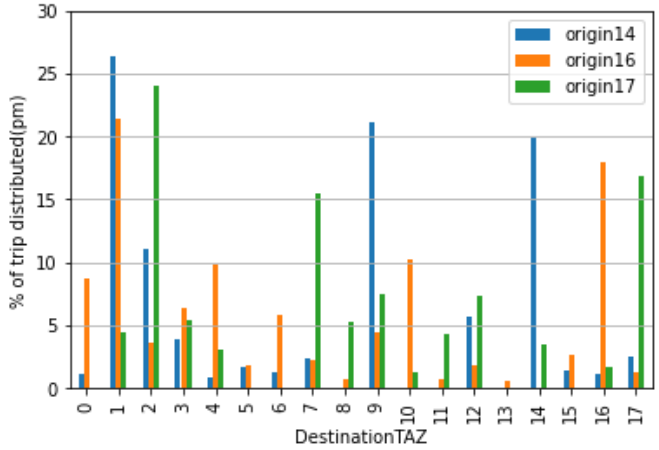

Fig. 12 Trip distribution from each origin in pm model

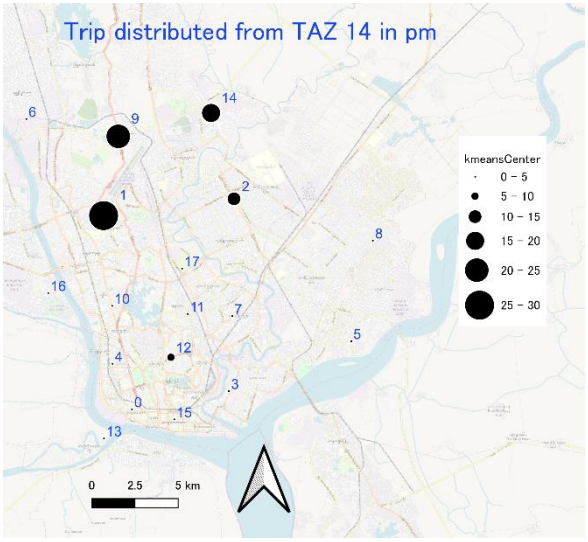

Fig. 13 Trip distribution from TAZ 14 in am

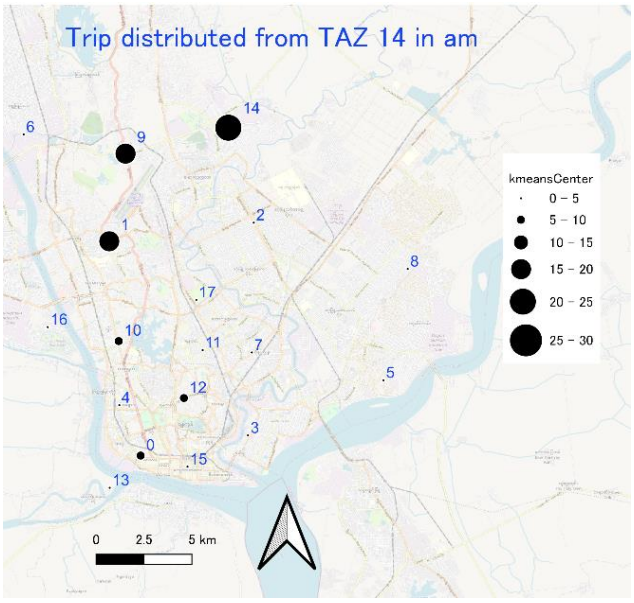

Fig. 14 Trip distribution from TAZ 14 in pm

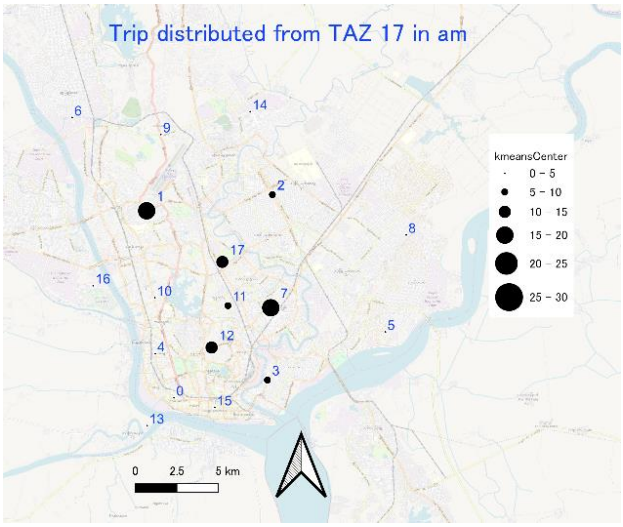

Fig. 15 Trip distribution from TAZ 17 in am

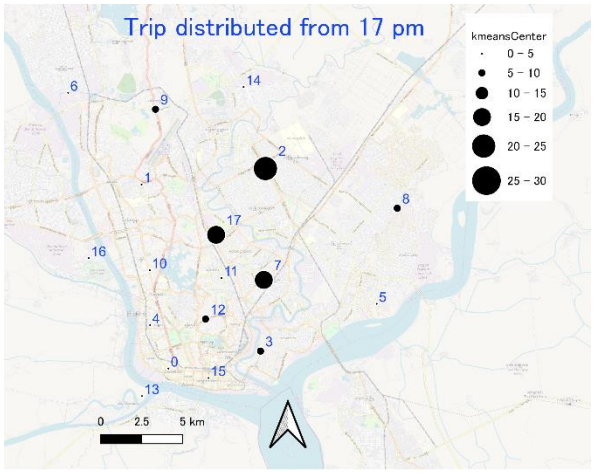

Fig. 16 Trip distribution from TAZ 17 in pm

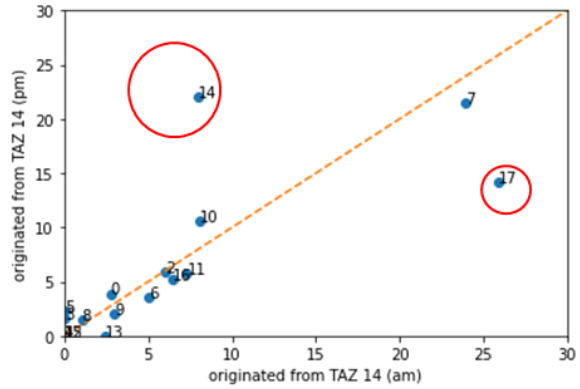

Fig. 17 all trips generated from TAZ14. Red circles points attract more trips in evening (the unit of axes are % trip)

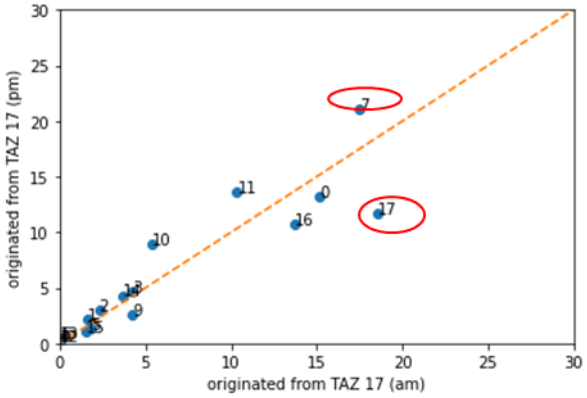

Fig.18 all trips generated from TAZ17. Red circled points have larger variation (the unit of axes are % trip)

Figures 13~16 show the trip distribution on the GIS map. Those figures show clearly that geographical distance or accessibility has a clear influence on making trips.

Figure 17 shows that 3 out of 18 TAZs attract more trips from TAZ 14 in the evening period. Note that the dotted line in the Figure represents the 45^0 line. Therefore, the point above the dotted line attracted more trips in the evening. The red circles recognize the points which attract more trips in the evening. A large variation of trip distribution can be found for TAZ 17 (Figure 18). The red circles in the Figures show the points where large variations could be observed. The red circles recognize the points having large variations.

(5) Trip chain: we explored the trip chain from CDR data. For brevity, only trip chains generated from TAZ 4 were provided in table 1. In the morning period, 2.4% of trips from TAZ 4 reach TAZ 1 through TAZ 10. Similarly, in the evening period, 2.55% of trips from TAZ 4 reach 16 through TAZ 10. Many trips do not create a trip chain in a particular period. As our study focused on two distinct periods (i.e., am and pm), we ignored the trip-chain in a day. One finding is that it is important to consider a longer period for exploring trip chain properties.

Table 1 Trip chain from TAZ 4 in different periods

Model am		Model pm	
Trip chain	% trip	trip chain	%trip
4_10	13.95	4_10	13.70
4_0	12.65	4_0	8.80
4_16	6.90	4_10_4	8.05
4_1	5.90	4_12	3.75
4_12	4.90	4_0_4	3.30
4_10_4	4.60	4_0_15	2.70
4_15	4.45	4_10_16	2.55
4_0_4	4.15	4_15	2.50
4_9	2.90		
4_10_1	2.40		

*Trip percentage less than 2% are ignored

5. CONCLUSIONS:

Transportation planning needs a clear picture of people's movement within cities. Though traditional survey methods can capture many attributes, they fail to collect unbiased data. Moreover, the number of samples collected through the traditional survey is limited particularly due to budget limitations. The rise of mobile positional data has generated a wealth of new data on human mobility, but new tools must be developed to integrate these data and insights into traditional transportation modeling approaches. The strength of this big data is that data are automatically generated, and the trip information is real.

Here, we have proposed a model for evaluating trip chain characteristics in cities like Yangon, Myanmar, where updated survey data is not available. The model was validated by comparing with trip origins from our testing data. The ratio of trips from each zone was predicted correctly by our proposed model. Zonal variations were observed in whether people will go to different zones or stay in the same zones. Some zones show variation for the time of the day. Depending on the telecommunication tower access properties, we have classified the zones in Yangon. It shows that Yangon consists of mixed zones. In other words, residential zones are not only for residing but also have commercial activities. While thinking of smart city planning, the outcomes of zonal characteristics should be considered.

The massive CDR data allow us to explore the trip distributions within the city. We can see the clear influence of distance or accessibility. Therefore, trips generated from zone 17 terminate mostly in nearby zones. Besides, we have found that many zones have a higher probability to produce trips in the morning than in the evening.

We have shown that the ratio of trips for each destination varies in the different time frames. Besides, accessibility in terms of distance is a significant feature in trip distribution.

REFERENCE:

1) Asakura, Y. Hato, E. (2004) Tracking survey for individual travel behaviour using mobile communication instruments, Transportation Research Part C: Emerging Technologies Vol 12, Issues 3-4: 273-291
2) Alexander, L., Jiang, S., Murga, M., and González, M. C. (2015) Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies, 58, Part B:240–250.
3) Bar-Gera, H. (2007) Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: A case study from Israel, Transportation Research Part C: Emerging Technologies Vol 15, Issues 6: 380-391.
4) Bowman, J. L. and Ben-Akiva, M. E. (2001) Activity-based disaggregate travel demand model system with activity schedules, Transportation Research Part A: Policy and Practice, vol. 35(1), pages 1-28.
5) Calabrese F, Diao M, Di Lorenzo G, Ferreira, J. and Ratti, C. (2013) Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. Transportation Research Part C: Emerging Technologies 26: 301–313.
6) Candia, J., González, M.C., Wang, P., Schoenhar, T.,Madey, G., and Barabási, A.L. (2008) Uncovering individual and collective human dynamics from mobile phone records, Journal of Physics A: Mathematical and Theoretical 41(22), 224015.
7) Han, G., and Sohn, K. (2016) Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model, Transportation Research Part B: Methodological , vol 83, pp 121-135
8) Kapicioglu, B. (2013) Applications of machine learning to location data, PhD dissertation, Princeton University
9) Lwin, K.K., Sekimoto, Y., Takeuchi, W. (2018). Estimation of Hourly Link Population and Flow Directions from Mobile CDR. ISPRS-International Journal of Geo-information, 7, 449.
10) Gonzalez, M., Hidalgo, C., Barabasi, A. (2008) Understanding individual human mobility patterns. Nature, vol. 453, no. 7196, pp. 779– 782.
11) Tripuraneni,N., Gu,S., Ge, H., Ghahramani, Z. (2015) Particle Gibbs for Infinite Hidden Markov Models, In NIPS, pages 2386–2394. Curran Associates, Inc., 2015.
12) Toole, J.L., Colak S., Sturt, B., Alexander, L.P., Evsukoff, A., González, M.C. (2015) The path most traveled: Travel demand estimation using big data resources, Transportation Research Part C