ISBN 978-981-11-7861-0
Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering (WCSE 2018)
Bangkok, 28-30 June, 2018, pp. 342-350
doi: 10.18178/wcse.2018.06.061

Analysis Mobile Usage Pattern from CDR in Different Area Types

Naruethai Thongphasook [1+] and Veera Muangsin [2]

[1] Dept. of Computer Engineering, Faculty of Engineering, Chulalongkorn University Bangkok, Thailand
[2] CU Big Data Analytics and IoT Center (CUBIC), Dept. of Computer Engineering, Faculty of Engineering, Chulalongkorn University Bangkok, Thailand

Abstract: Since mobile phone has become one of the most popular communication method. In order to find different characteristics of each cell towers and locations from various type of data collecting within CDRs. We explore CDRs to find amount of people the city in a period of time, to analyze highly active period and inactive hour in day of weeks. Behavior of mobile phone usage. And implement clustering algorithm to find a proper number which gather distinctive usage patterns from several cell towers within 24 study locations, which are located in Bangkok and surroundings area. In this study, we also discovered some patterns that can acutely describe area use.

Keywords: call detail records, pattern analysis.

1. Introduction

Human geo-demographic and mobility behaviour are strongly associate with geographic area. People activities also vary from time to time and place to place. [1]-[5] For examples, A business area is filled up with company branches and office employees during working hours. Students and teachers stay at school during the day, while parents may only drop off their children in the morning and pick them up in the evening. At night, nightlife areas are full of party lovers.

In recent years, most people carry mobile phones with them since it offers ease of communication. In Thailand, mobile phone penetration is nearly 150% of population, with about 98 million subscribers. The largest mobile service operator, Advanced Info Service (AIS) has more than 40 million subscribers. For billing and other purposes, a mobile operator collects information about every phone call and text message (SMS). Such information is called Call Detail Record (CDR). Each record contains data of a single call or SMS, including the source phone number, the destination phone number, call types (incoming or outgoing), call start time, call duration, and cell site ID. The cell site in a CDR record is the cell tower that communicate to the phone when the call is initiated. Therefore, the geo-location of the cell site indicates that the mobile phone user is inside the coverage area of the cell tower while making the phone call. Therefore, CDR can be useful for activity analysis in terms of user behaviour [6]-[8] and patterns of urban community [9], [10].

When we consider about cell towers in places that have their own specific importance, there is strong possibility that same types of places in different locations, such as school, tend to have something similar in usage pattern.

Moreover, some places also have specific distinctions that are influenced by date and time. For instances, people come to exhibition hall have various motives owing to events, students and parents are liable to visit school at different time, etc.

+ Corresponding author. Tel.: +66-2-218-6981; fax: +66-2-218-6955
E-mail address: veera.m@chula.ac.th

We are motivated to explore mobile usage patterns of service sites from CDRs within a grid from different areas. It is an interesting question whether we can gather some cell sites together or distinguish area types from mobile usage data.

This study will be useful for area classification from mobile usage data and the result will imply facility usage and population density for capacity planning of mobile infrastructure.

2. Data Description and Study Area

We collected data between 1st August and 31st October 2016 from 24 areas within Bangkok and surrounding provinces. The study areas are selected from 600x600 meter grids so that they are different in term of land use and each of them has at least one cell tower. All phone numbers were encrypted before provided to researchers due to privacy issues. From 136 cell sites within study areas, we obtained 129,585,962 records, both voice calls and SMS from 2,602,401 distinct users.

3. Methodology

3.1. Data Exploration

This study focuses on pattern of mobile phone usage from various locations. Two types of call both can evaluate amounts of mobile phone usage into 1) average duration of call (Figure 1.)

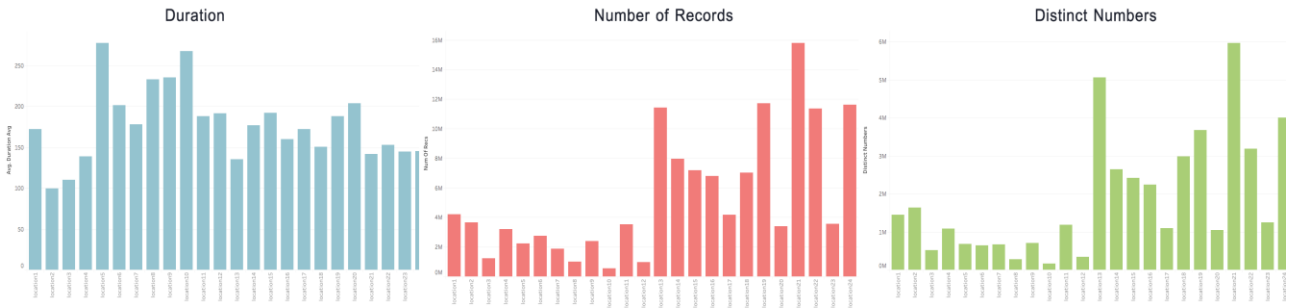

Fig. 1: Graph of mobile usage represents average duration (Left) number of records (Mid) and distinct numbers obtained from location 1 to location 24 (X-axis).

Other two quantity values 1) Number of records means summary of records created within a period of time 2) Distinct numbers refers to number of unique mobile phone appeared within a period of time. (Figure 1.) However, SMS records have constant duration and are therefore ignored (Figure 3.)

Investigating minor elements of study locations, each location contains number of different cell sites depend on area and service provider. Figure 2. implies that each location basically has only few number of dominant cell towers.

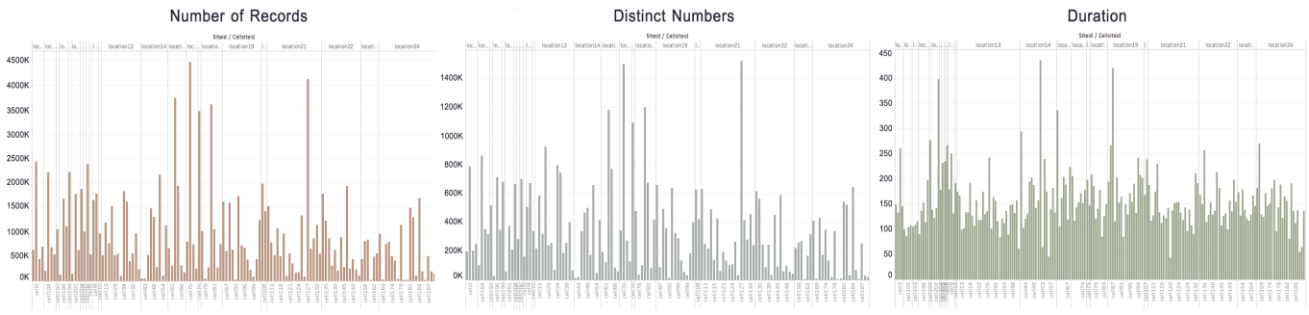

Fig. 2: From left to right - number of records, distinct numbers and duration displays separately by cell tower within each location.

3.2. Normalization

Because of huge difference of usage amount between each study location. It is necessary to normalize data before doing further analysis.

$$z_i(t_k) = \frac{x_i - \mu_k}{\sigma_k} \quad (1)$$

Since we aimed to discover mobile usage pattern of each cell tower, we normalized value z_i from original variable x_i by mean and variance of each cell tower t_k as in (1). In this paper, we separately calculated within 4 sub groups - incoming voice call, outgoing, voice call, summary voice call and SMS.

3.3. Principle Component Analysis

In order to reduce variable dimension, we applied PCAs separately for each sub group of data. With a large number of variables, some of these might be correlated. We used PCA to transform original variables into linear combination.

For i=1,2,..,n observation units, PCA transforms original x_i variables into new variables w_i called principle component as in (2).

$$\sum_{i=1}^{n} \sum_{p=1}^{m} w_{i,k} = c_{pk} x_{ip} \quad (2)$$

Coefficient value c_{pk} informs how significant of each original variable x_{ip} contributes to linear combination in order to calculate new variable $w_{i,k}$.

The result of cumulative plots from principle components (Figure 3.) reflects how cumulative proportion of each evaluation value (Y-Axis) represents at number of principle components (X-Axis). Graph plot by Distinct numbers is slightly steeper than Number of records. Meanwhile the outcome calculated from average duration is obviously the most inclined slope. This implies that Distinct numbers can refer higher amount of data within fewer numbers of PC. Nevertheless, sub groups of data – incoming, outgoing, summary of both or numbers of all records - do not give a significant difference.

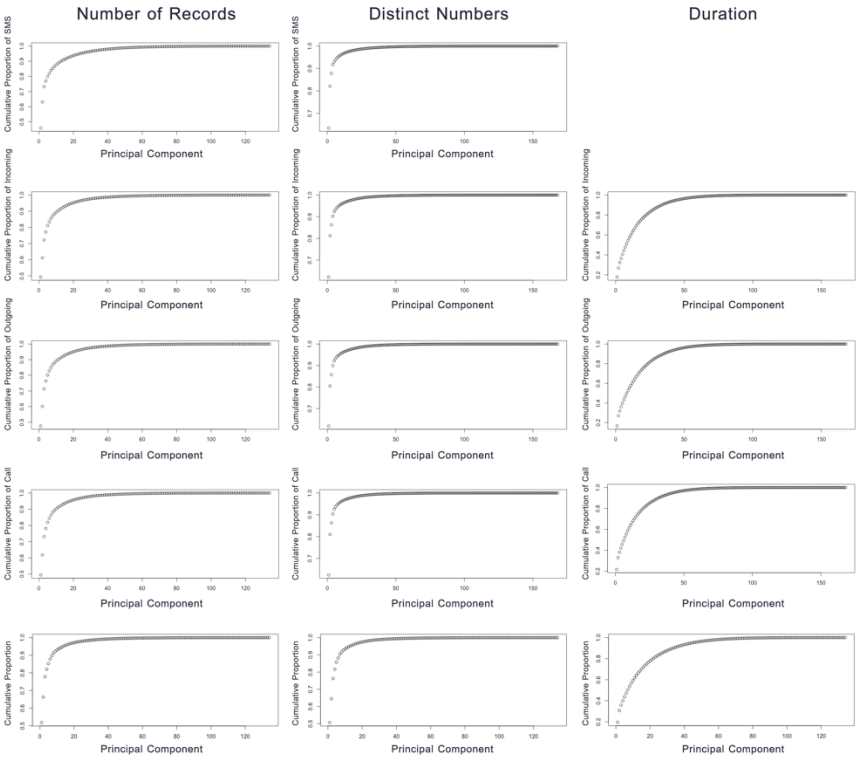

Fig. 3: Cumulative plot from PCA of each value. PCA from distinct numbers (Middle) gives slightly steeper cumulative than number of records (Left). Meanwhile duration of call is the most inclined. From top to bottom, SMS, incoming voice call, outgoing voice call, summary voice call and summary of CDRs.

3.4. Mobile Phone Usage Pattern

Since our study locations are diverse. Each location also contains different number of cell towers within a grid area. Moreover, 600x600 m^2 is the least coverage distance that every grid definitely contains a cell tower, it is still a large area in which many activity spots is located. For this reason, it is challenging to investigate mobile phone usage in particular cell tower and explore whether can classify location types from usage pattern or not.

In order to describe mobile phone usage pattern, we aggregated value into a block representing an amount in hour and day of week (24x7) giving an outcome of 168 values in particular time period. 2-dimension matrices construct of data representing pattern S =$\{s_{i,j}\}$ for number of records, T=$\{t_{i,j}\}$ for distinct numbers and U=$\{u_{i,j}\}$ for duration.

Since cumulative graph plot in Figure 3. shows barely noticeable difference. We chose summary of both number of records and distinct number to represent data in order to find usage pattern.

3.5. K-means Clustering

K-means clustering is an iterative algorithm that aims to separate observations into k clusters. The procedure starts with select initial centers of K clusters. Next, it assigns each observation to the cluster which has the least Euclidean distance, and then recalculate new centroid of each cluster. K-means iterates these steps, assigning observation to the nearest centroid until it no longer reassigns observations to another clusters. The result of cluster analysis aims to group observations by similarities.

4. Experiment and Result

We combined two matrices, S and T together into large matrices of 2x24x7, processed principal components calculated. Number of PCs using in k-means clustering algorithm is 4, which explain more than 80% of data.

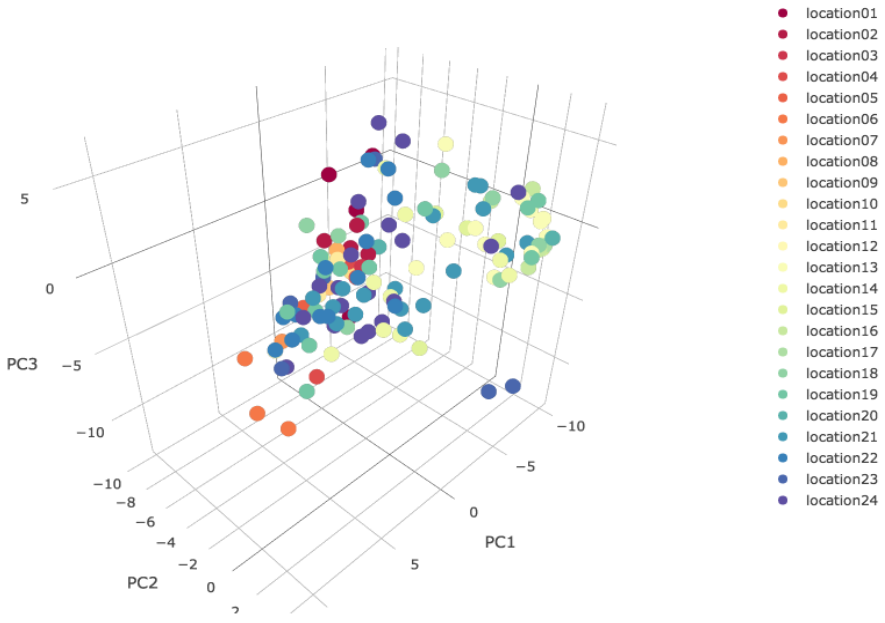

Fig. 4: Three-dimension scatter plot show how cell towers from each location displays coordinates representing by 3 PCs which refer approximately 70% of data. Same color points are obtained from same location.

From our observation, mobile phone usage pattern represents the best result at k=9. (Figure 5.) Usage pattern of distinct numbers seems obviously correspond to number of records. Meanwhile we did not find a significant relationship between call duration and other information.

Fig. 5: Mobile phone usage pattern, displays into 9 groups of clusters grouped by k-means algorithm considering two units - number of records and distinct users. From top, distinct users, number of records and average duration.

Most of usage patterns illustrate in a square shape with tail. This is reasonable since people generally do activity during daytime. However, there are noticeable differences in peak usage hour. Especially in cluster 9, it is clearly distinguished that cell towers within this cluster is highly active in the weekend. We looked into area use and found out a large weekend market where consists of more than 8,000 trading stalls.

And in addition to cluster 9 breakdown, both two cell towers within cluster 9 are also obtained from a same grid.

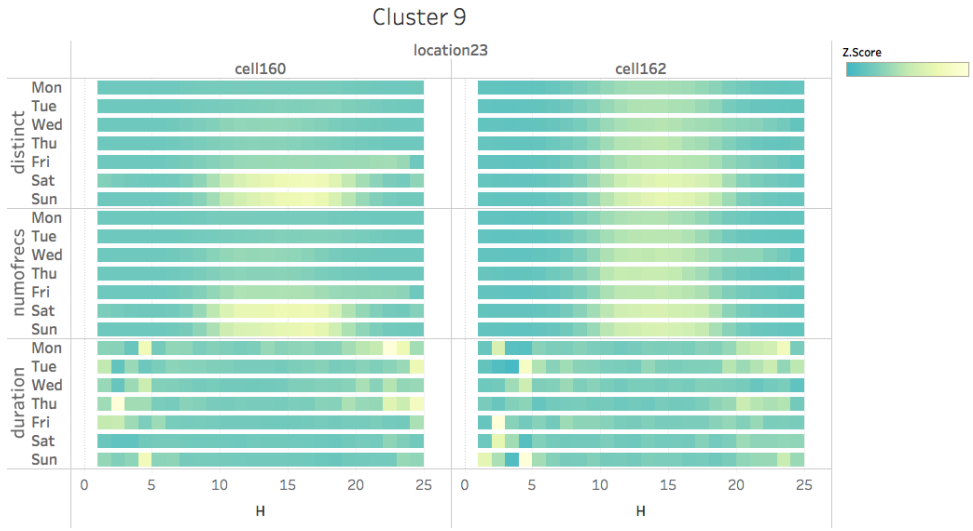

Fig. 6: Mobile phone usage pattern reflects unique characteristic of cell towers in cluster 9, both highly actives in the weekend.

For further investigation, we broke down interesting locations into cell tower and explored how infrastructure can be gathered together.

First of all, both location 1 and 2 are quite unique places. Mobile usage pattern of these two locations are also interesting. Considering Figure 7., cell towers in cluster 7 tend to active to most on Friday 8pm. While peak time of cluster 6 is likely to be earlier in the afternoon. Both two cluster seems to be inactive during daytime. On the other hand, cell tower no.3 which also located in location 1 actives in working hour, is separated in another cluster.

Our survey inform that the most notable area use of these two locations are inter-city transportation hub.

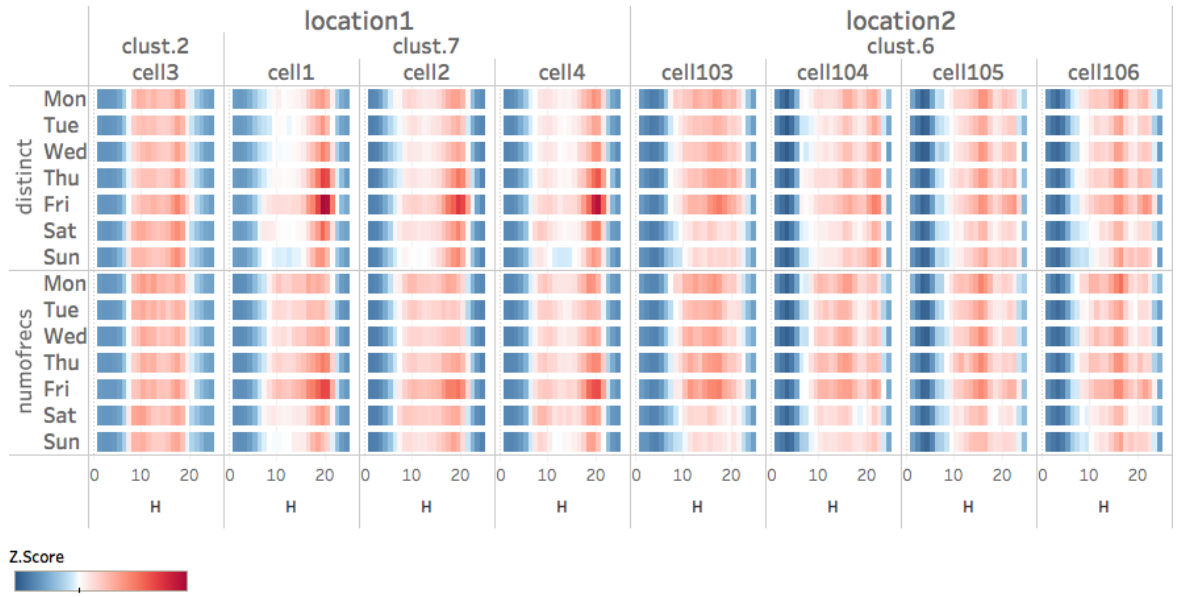

Fig. 7: Shows mobile usage pattern of location 1 and location 2 discretely into particular cell towers. (Top) Pattern of distinct numbers and (Bottom) Pattern of number of records. Pattern from both types of data shows high active period on Friday evening.

All cell towers within location 6 and 7, where government agency located, produce explicitly square shape of pattern. It is normally active during work hour approximately from 9am to 6pm. However, mobile pattern shows hardly active usage in the weekend. Location 8 to 10 – where contains schools and small merchandises have the same active period of time but seem to be a lot more active in the weekend.

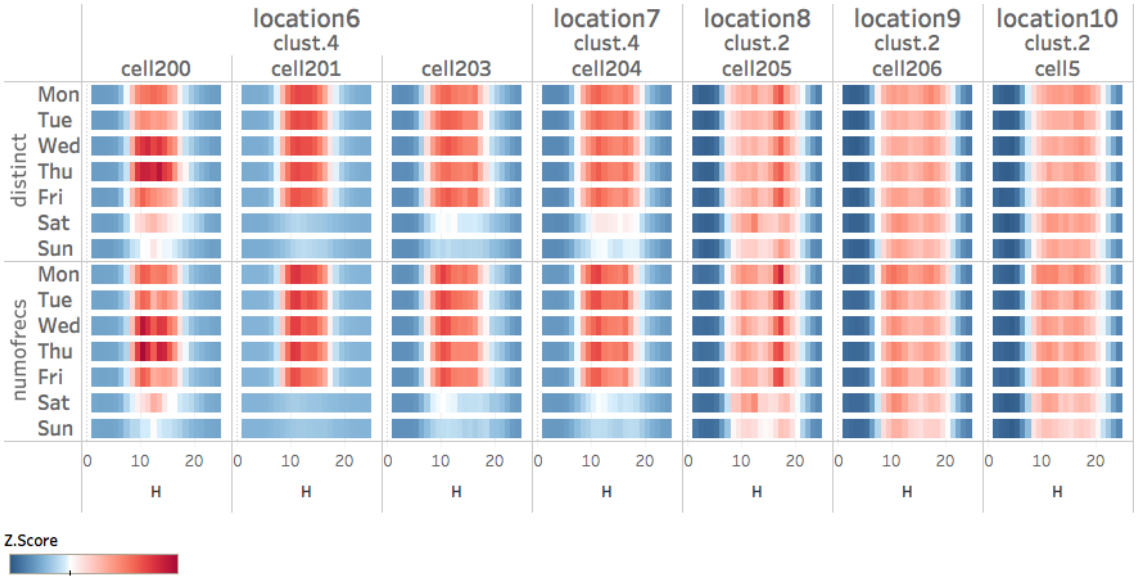

Fig. 8: Shows mobile usage pattern of location 6-10 discretely into particular cell towers. (Top) Pattern of distinct numbers and (Bottom) Pattern of number of records. Usage pattern shows highly active during working hour. Meanwhile, location 6 and 7 are silent in the weekend, location 8 9 and 10 tends to be more active.

It became more difficult to describe location type when location is highly active. Upper row of Figure 9. represents all cell towers within location 13. A highly active location may contain numbers of cell tower up to 20. However, some cell towers created tiny number of data. To help this issue, we defined a value k representing percentage of data for every single cell tower compare to summary of data within a location.

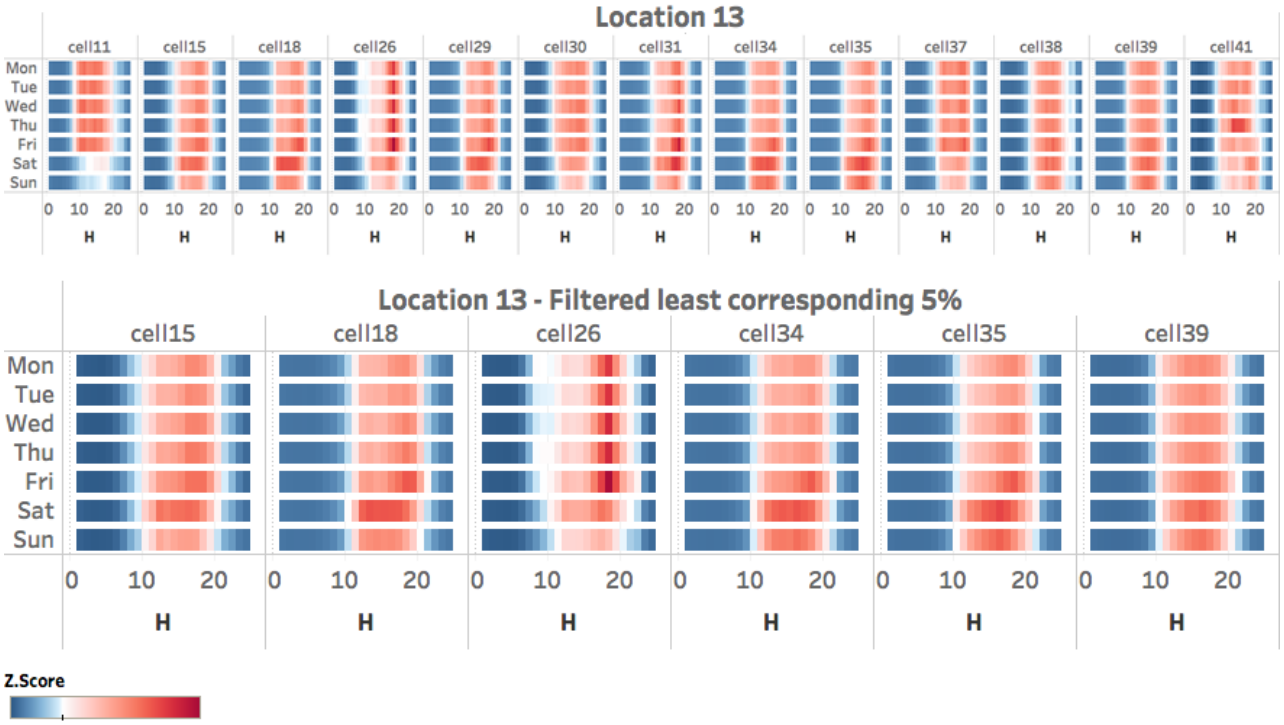

Fig. 9: Shows mobile usage pattern of location 13 discretely into particular cell towers. (Top) Pattern of distinct numbers and (Bottom) Pattern of number of records.

After filtering cell tower at least corresponding 5 % of data within each location, remaining cell towers seemingly have more correlation in their pattern. Cumulative proportion calculating from selected cell towers approach 80% of data at PC3, faster than before filtering. However, appearance of usage pattern by every single location still remain the same characteristic. This also implies strong influence of high activity cell tower.

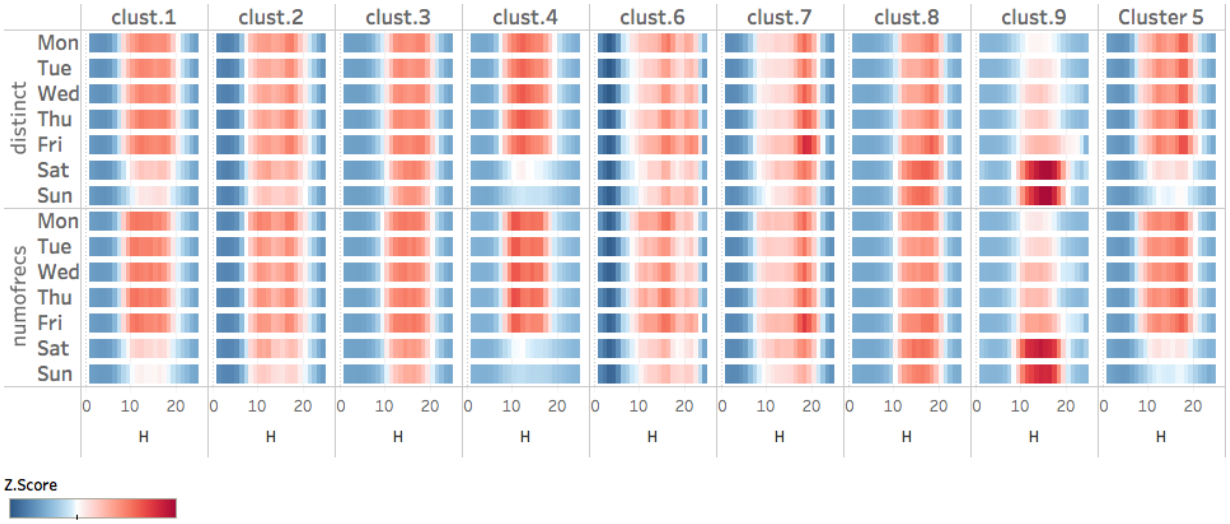

Fig. 10: Shows mobile usage pattern after filtering low usage cell towers. Meanwhile usage patterns are sustained the same characteristics.

While dominant cell towers can represent overcall characteristic of an area. Others may explain detail components of the area. In Figure 11. shows a part of usage patterns corresponding to cell towers within location 21. Rectangular patterns plotted in this area may indicate routine activities during daytime such as government agency or workplace. At the same time cell tower no.127 is highly active in the weekend. We discovered this kind of characteristic before in Figure 6. representing a popular weekend market.

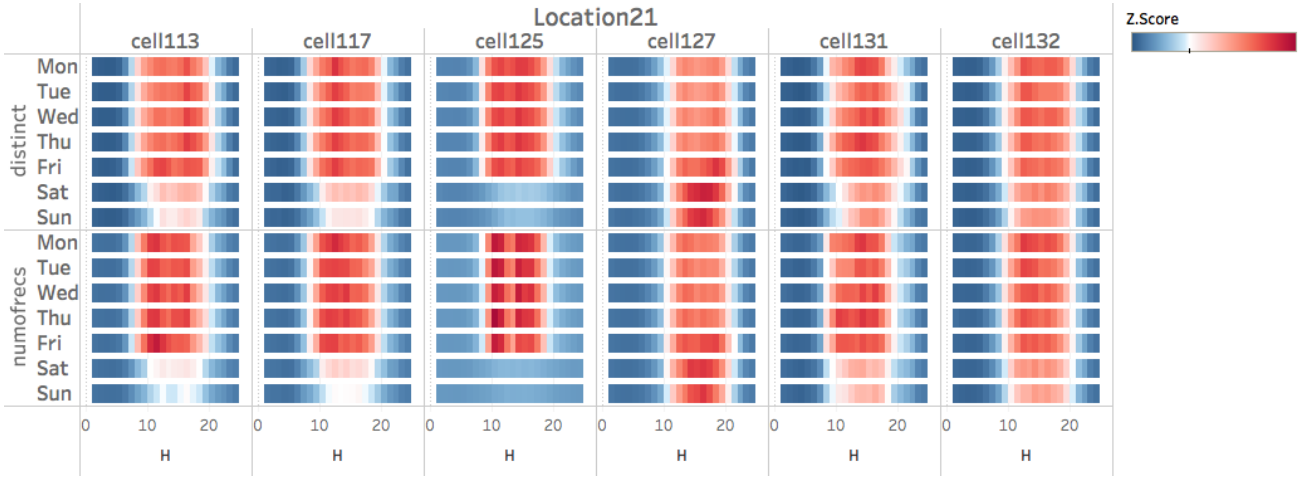

Fig. 11: Shows mobile usage pattern of location 21 discretely into particular cell towers. (Top) Pattern of distinct numbers and (bottom) pattern of number of records.

From cluster result, cell sites with similar pattern from different locations were also gathered together. For examples, Figure 12. displays cell sites in cluster 5, which were obtained from 4 different locations. However, whole locations are shopping mall with rapid transit station.

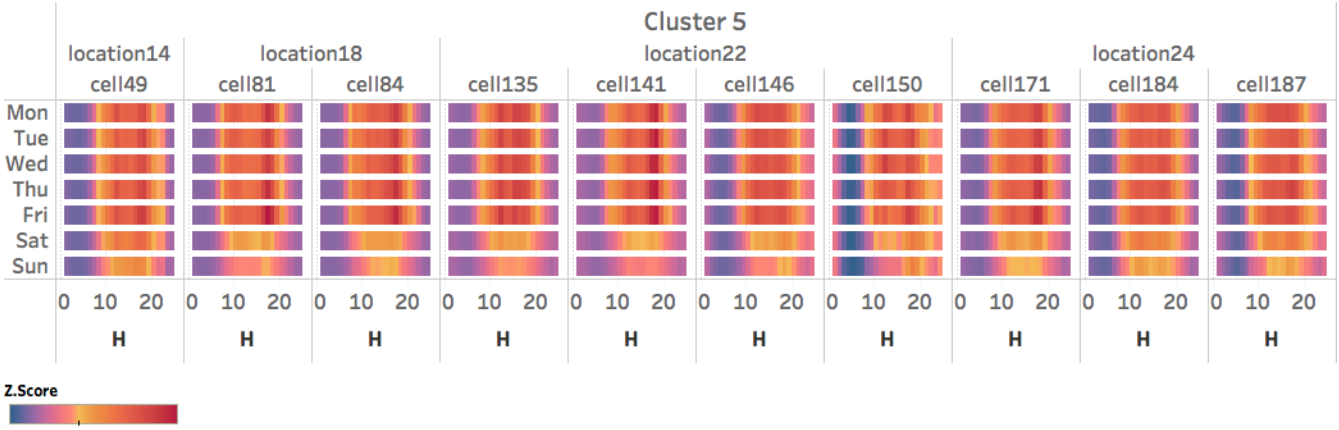

Fig. 12: Shows mobile usage pattern of cell towers in cluster 5, which were obtain from location 14, 18, 22, and 24. All locations are shopping mall. Moreover, there is sky train station in location 14, 22 and 24 meanwhile location 18, 22 and 24 have underground train station.

5. Conclusion

This paper presents a method for area classification from mobile phone usage. By implementing PCA, 4 components from PCA can cover approximately 80% of dataset.

We processed to find out the fact that most places have few numbers of dominant cell towers whereas others are slightly to be inactive. Therefore, mobile usage from dominant cell towers can indicate area types.

Some locations have single usage purpose which reflect obvious characteristics. These locations can be effectively detected. However, some locations have multi-purpose usage. For examples, shopping mall and office building located in a same grid, etc. These locations are so complicated that area type identification can be difficult.

Eventually, we examined consistency of cluster result from each variable by implementing confusion matrix. From this experiment, accuracy of cluster results calculated from same type – distinct, number of records and duration – tend to give close result to each other. On the other hand, accuracy of distinct number voice call comparing to number of records voice call can slightly drop to only 10% (Figure.13)

Although number of records and distinct numbers are likely to have similar pattern of usage, cluster results are different. Among these 3 types of data in attention, distinct numbers seem to be the most consistency. Especially, when corresponding to same type of data.

Table 1: Shows consistency result computing by confusion matrix

	NUMBER OF RECORDS CALL IN	DISTINCT CALL IN	DURATION CALL IN
NUMBER OF RECORDS CALL OUT	0.5063	0.8608	0.2025
DISTINCT CALL OUT	0.481	0.8734	0.2152
DURATION CALL OUT	0.2152	0.1772	0.1519

6. Acknowledgement

This research is supported by Advanced Info Service Public Company Limited (AIS) and Chulalongkorn Academic Advancement into 2nd Century Project, Chulalongkorn University.

7. References

[1] Axhausen, K.W., Activity spaces, biographies, social networks and their welfare gains and externalities: some hypotheses and empirical results. Mobilities 2, 2007.

[2] S. Hassan, X. Zhan, and S. Ukkusuri. Understanding Urban Human Activity and Mobility Patterns Using Large-scale Location-based Data from Online Social Media. Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013.

[3] S. Jiang , G. Fiore , Y. Yang , J. Ferreira, Jr. , E. Frazzoli , M. Gonz ́alez, A review of urban computing for mobile phone traces: current methods, challenges and opportunities, Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, August 11-11, 2013

[4] S. Jiang, J. Ferreira, and M. González, “Clustering daily patterns of human activities in the city,” Data Min. Knowl. Discov., vol. 25, no. 3, pp. 478–510, Apr. 2012.

[5] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data. In: Salah A.A., Gevers T., Sebe N., Vinciarelli A. (eds) Human Behavior Understanding. HBU 2010.

[6] C. M. Schneider, V. Belik, T. Couronn ́e, Z. Smoreda, and M.C. Gonza ́lez. Unravelling daily human mobility motifs. Journal of The Royal Society Interface, 10(84), 2013.

[7] M. C. Gonza ́lez, C. A. Hidalgo, A.-L. Baraba ́si, Understanding individual human mobility patterns, Nature 458 (2009) 238–238. doi:10.1038/nature07850.

[8] C. Kang, S. Sobolevsky, Y. Liu, C. Ratti “Exploring Human Movements in Singapore: A Comparative Analysis Based on Mobile Phone and Taxicab Usages” in Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, 2013.

[9] Data for Development (D4D) Challenge. [Online] Available: http://d4d.orange.com/en/Accueil.

[10] Alhasoun, F., Almaatouq, A., Greco, K., Campari, R., Alfaris, A. and Ratti, C., 2014. The City Browser: Utilizing Massive Call Data to Infer City Mobility Dynamics. In the Proceedings of the 3rd International Workshop on Urban Computing (UrbComp), 2014