การพัฒนาระบบจัดประเภทพื้นที่จากข้อมูลการใช้งานโทรศัพท์เคลื่อนที่

นางสาวณฤทัย ทองผาสุข

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2560

Developing an area classification system from mobile phone usage data

Miss Naruethai Thongphasook

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Engineering Program in Computer Engineering
Department of Computer Engineering
Faculty of Engineering
Chulalongkorn University
Academic Year 2017

Thesis Title	Developing an area classification system from mobile phone usage data
By	Miss Naruethai Thongphasook
Field of Study	Computer Engineering
Thesis Advisor	Assistant Professor Dr. Veera Muangsin

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

....Dean of the Faculty of Engineering

(Associate Professor Dr. Supot Teachavorasinskun)

THESIS COMMITTEE

....Chairman

(Assistant Professor Dr. Krerk Piromsopa)

....Thesis Advisor

(Assistant Professor Dr. Veera Muangsin)

....Examiner

(Associate Professor Dr. Twittie Senivongse)

....External Examiner

(Associate Professor Dr. Worasait Suwannik)

ณฤทัย ทองผาสุข : การพัฒนาระบบจัดประเภทพื้นที่จากข้อมูลการใช้งานโทรศัพท์เคลื่อนที่ (Developing an area classification system from mobile phone usage data) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. วีระ เหมืองสิน, หน้า.

เนื่องจากกิจกรรมของมนุษย์มีความหลากหลาย แตกต่างกันไปไปตามเวลาและสถานที่ รวมถึงประเภทการใช้งานพื้นที่ในเมือง ซึ่งข้อมูลเหล่านี้สามารถเป็นประโยชน์ต่อการวางผังเมือง และเนื่องจากในปัจจุบันผู้คนมักพกพาโทรศัพท์มือถือเพื่อการใช้งาน จึงทำให้เราเกิดความคิดที่จะจำแนกประเภทพื้นที่จากพฤติกรรมการใช้งานโทรศัพท์เคลื่อนที่ ข้อมูลซีดีอาร์ถูกใช้เพื่อบอกพฤติกรรมการใช้งานโทรศัพท์ โดยแบ่งเป็นช่วงเวลาละหนึ่งชั่วโมง นับจาก 1:00 ถึง 24:00 และแต่ละวันในสัปดาห์ นับจากวันจันทร์ถึงวันอาทิตย์ จากนั้นจึงนำมาจับกลุ่มเพื่อหารูปแบบพฤติกรรมการใช้งานเสาสัญญาณ และได้คำนวณสัดส่วนปริมาณของพฤติกรรมการใช้งานแต่ละแบบ เพื่อนำมาใช้ในการจัดแบ่งประเภทพื้นที่ นอกจากนี้ ผู้วิจัยยังได้วิเคราะห์ข้อมูลที่จัดแบ่งประเภทไม่ตรง และพื้นที่ที่อาจสามารถแบ่งแยกเป็นกลุ่มย่อยได้อีก

ภาควิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อนิสิต ....

สาขาวิชา วิศวกรรมคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

ปีการศึกษา 2560

5770156821 : MAJOR COMPUTER ENGINEERING

KEYWORDS: CDR / PATTERN ANALYSIS

NARUETHAI THONGPHASOOK: Developing an area classification system from mobile phone usage data. ADVISOR: ASST. PROF. DR. VEERA MUANGSIN, pp.

Since human activities are vary by time and place, there have been many attempts to extract social behavior in spatial studies which included area use in the city. This information helps gaining advantages in city and facilities planning. Nowadays, many people carry mobile phone with them for communication purpose. This motivates us to analyze mobile phone usage in different area types. This thesis proposes method of an analysis for area use classification from mobile phone usage pattern. CDR data was used to define mobile phone usage pattern by hour from 1:00 to 24:00 and day of week from Monday to Sunday. And then processed clustering algorithm to find optimal groups of usage pattern from cell towers. Ratio of usage from each pattern is used to predict area type and facilities. Besides, the researcher also revealed the examples of mismatch location and minor groups of area use in this thesis.

Department:	Computer Engineering	Student's Signature	....
Field of Study:	Computer Engineering	Advisor's Signature	....
Academic Year:	2017		

ACKNOWLEDGEMENTS

I would like to express my deep gratitude to all the person for their contribution to this thesis. Especially, Dr. Veera Muangsin, my advisor, who gave me advise and continued support throughout this study.

I would also greatly appreciate comments from all my thesis committees, Asst. Prof. Krerk Piromsopa, Ph.D., Assc.Prof. Dr. Twitte Senivongse, Ph.D., Assoc. Prof. Atiwong Suchato, Ph.D. and Assoc. Prof. Worasait Suwannik, Ph.D.

I would like to thank Advanced Info Service Public Company Limited (AIS) for supporting resources in this research.

I would also like to thank Faculty of Engineering, Chulalongkorn university, especially, Scientific Parallel Computer Engineering Laboratory (SPACE lab).

Finally, I must express my gratitude to my parents and friends for providing me support and encouragement throughout my years of study.

CONTENTS

Page

Page

Page

List of Figures

List of Tables

Chapter 1

Introduction

1.1. Introduction

Mobile phones have become an essential tool for communication purpose. A survey from National Statistical Office of Thailand [1] reported that 93.5% of people older than 6 years old in Bangkok use mobile phone in their daily life. Advanced Info Service (AIS), one of the largest telecommunications operators in Thailand have more than 40 million subscribers which is approximately 60% of Thai population. And since most people are carry mobile phone with them when they go outside, data of mobile phone data from million users is useful for human behavior analysis.

In order to work, mobile phone must connect with a cell tower. At the same time, the service operator creates Call Detail Records (CDR) data for billing purpose. This data contains data such as timestamp and location when the call happened. For this reason, mobile phone data can be useful for human behavior analysis.

Since human activities are vary by time and place, usage behavior of mobile phone at different locations tend to be diverse as well. And because human activities are related to types of place and facilities in the areas. For example, school should be very active in the evening after school time. And same area types located in different location should have similar usage pattern. Thus, it can be assumed that area use and facilities can be detected by mobile phone usage pattern.

1.2. Objective

This work aims to develop a method in to distinguish area types from mobile phone usage and discover similarity in same types of places located in different areas.

1.3. Scope of Study

- This research studied mobile usage pattern obtained from different area types.
- This research analyzed mobile usage pattern from CDR which collected between 1st August – 31st October 2016

1.4. Research Process

1. Study research relating to mobile phone pattern.

2. Study detail of received data and explore initial data.

3. Filter defect data.

4. Implement K-means clustering algorithm to group cell towers with similar usage pattern.

5. Set up list of area types and facilities to predict.

6. Apply principal component analysis to reduce variable.

7. Visualize the result.

8. Design model to improve location classification.

9. Compare results using different variables.

10. Evaluate classification model using K-fold validation.

11. Results and write thesis

1.5. Expected Benefits

- Be able to distinguish area types from large mobile phone usage data automatically
- Can be applied in marketing and urban planning

1.6. Publications

A part of this thesis had been published as follows:

Naruethai Thongphasook and Veera Muangsin, " Analysis of Mobile Usage Pattern for Area Classification", The 8th International Workshop on Computer Science and Engineering (WCSE 2018), Bangkok, Thailand, June 28-30, 2018.

Chapter 2

Background Theory and Related Works

2.1. Statistics

2.1.1. Principal Component Analysis

Principal component analysis is a statistical technique that uses an orthogonal factor to transform original variables into linear combination by giving more weight to variables with larger variance.

Principal Component Analysis has 4 steps following:

1. Find covariance matrix
   - Covariance is a value that imply relationship between two data. Positive covariance implies a direct relationship, negative covariance implies an inverse relationship between variables, while zero covariance means variables are not related. Covariance can be calculated from

     $$cov(x, y) = \frac{\sum_{i=1}^{n}(X - \bar{X})(Y - \bar{Y})}{n - 1}$$

     Where $\bar{X}$ and $\bar{Y}$ are means of variables

     n is number of data is the data set
   - Covariance matrix is a matrix that indicate relationship between n data
2. Calculate eigenvalues and eigenvectors of covariance matrix
   - Eigenvalues λ of a given matrix A are set of special scalars that

     $$A - \lambda I = 0$$

     Where I is identity matrix

- Eigenvector v of a given matrix A are vectors that

  $$A \cdot v = \lambda v$$

  Where λ is eigenvalue of matrix A

3. Eigenvectors with high eigenvalue are principal components
4. NewData = RowFeatureVector x RowData Adjust
   - RowFeatureVector is matrix of eigenvectors ordered by eigenvalues from high to low
   - RowDataAdjust is transformed matrix of data which each row refers to each variable

2.1.2. Elbow Method

The elbow method is a method to determine the optimal number of clusters in a dataset. The elbow method looks at the percentage of variance (WSS) explained a function of the number of clusters.

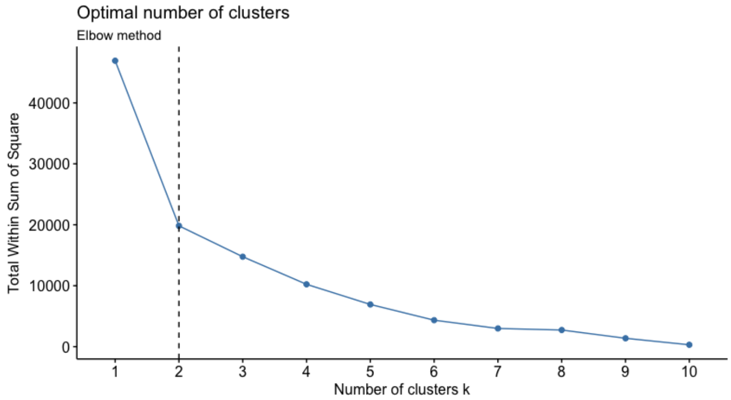


Figure 1 Sample of Elbow Plot shows a crucial point at k=2

2.1.3. Gap Statistic Method

The gap statistic is a method that compare $log(W_k)$ with value under a null reference distribution of the data. The optimal cluster will be the value that $log(W_k)$ maximize the gap statistic.

$$Gap(k) = E_n^*\{log(W_k)\} - log(W_k)$$

Where k is number of clusters

W_k is within-cluster dispersion

2.2. Data

2.2.1. Mobile Phone Data

Call Detail Record (CDR) is the data created when a mobile phone connects to a cell tower for an activity such as making a call, receiving a call or sending an SMS. It is normally used by a service operator for billing purpose. Each CDR 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.

Currently, many researches relating to spatial and urban study, use CDR or social network data to analyze city behavior. S. Hassan et. al [2] uses check-in data to study probability of activity by timeline and reveal relationship between check-in places and visit frequency. In this case, researcher uses word category to identify activities. S. Phithakkitnukoon et al. [3] uses Python APIs for Yahoo! search to profile activity for spatial study. The advantage of CDR is that it covers a large number of users. However, service area of cell tower is uncertain. It depends on usage amount in the area. Density of cell towers varies in line with usage amount. If there is high usage number in that area, density of cell towers will be high as well.

2.3. Machine Learning

Machine learning is a method of data analysis that provides systems the ability to automatically learn. Types of machine learning are separated into two categories:

2.3.1. Unsupervised Learning

Unsupervised learning is the learning process that all input data are unlabeled, and the algorithms learn to discover structure from training data set.

- K-Means Clustering is an iterative algorithm that aims to separate observations into k clusters by reassigning observations to the new closest centroid until all centroids are stable. The procedure starts with:
    i. Select initial centroids of K clusters at c1, c2,...ck
    ii. Assigns each observation to the cluster which has the least Euclidean distance
    iii. Recalculate new centroid of each cluster
    iv. Iterates step 2-3 until it no longer observations reassigns observations to another clusters. The result of cluster analysis aims to group observations by similarities.

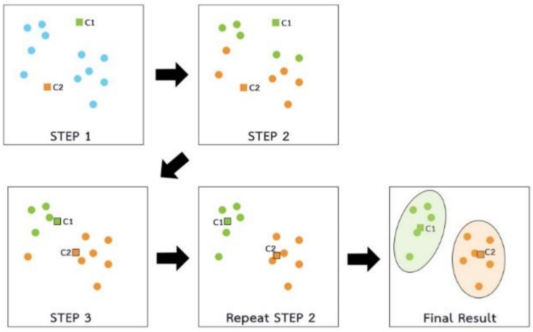


Figure 2 Sample of K-means Clustering Step by Step

- Hierarchical Clustering is an algorithm that try to identify two closest data and merge them together as one. The result can be illustrated in a dendrogram tree. The process starts with:

i. Find two closest observations and merge them together

ii. Iterates until all clusters are merged.

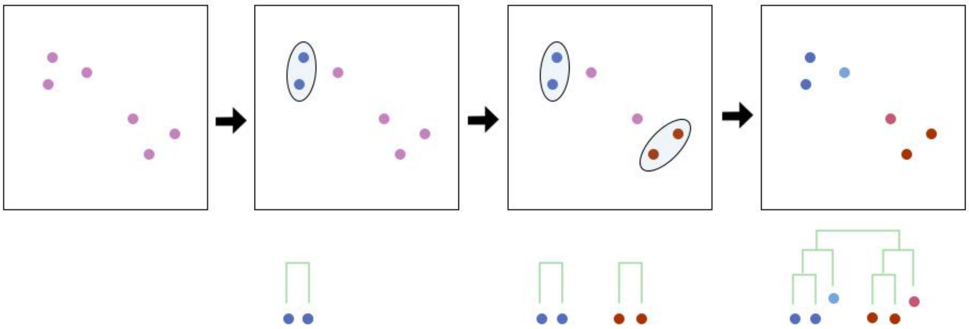


Figure 3 Sample of Hierarchical Clustering Step By Step

2.3.2. Supervised Learning

Supervised learning is the process that all input data are labeled, and the algorithms learns to predict output from the training data set.

Support Vector Machine (SVM) is a supervised learning algorithm which can be used for classification. SVM define class of data by finding the optimal hyperplane that have maximum margin between classes.

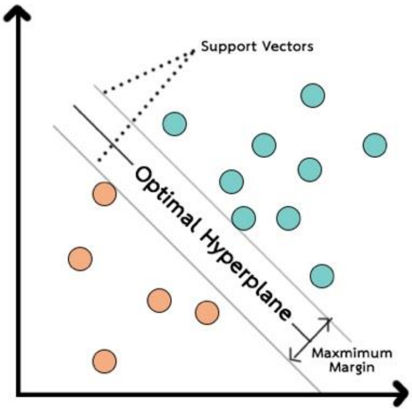


Figure 4 Optimal Hyperplane from Support Vector Machine

There are many kernel methods that used in SVM in order to map decision boundary. Some well-known kernels are:

- Linear Kernel - is the basic kernel function that suitable for linearly separable data.

$$K(X,Y) = X^T Y$$

- Radial Basis Function Kernel (RBF) – is a popular kernel function that commonly used in SVM classification. The value depends on distance between separation boundary and origin.

$$K(X,Y) = \exp(-\frac{\left||X-Y|\right|^2}{2\sigma^2})$$

Where σ is a free parameter that define strictness of separation

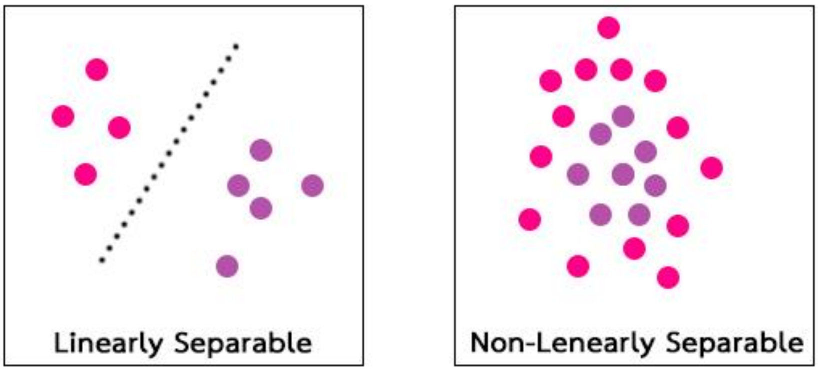


Figure 5 Sample pf Linearly – Non-Linearly Separable Data

Chapter3

Design and Implementation

This research designs to design a model that helps improving area use classification from CDR. This study uses CDR, obtained from cell towers within study locations which consist of various area types, to cluster cell towers. The result shows clusters that contain cell towers with similar usage pattern together and the usage pattern of each cluster. Later then, we use this result to points out behavior structure of study location.

3.1. Data Description

3.1.1. Call Detail Record

CDR is mobile phone usage data that are collected by telecommunications operator between 1st August and 31st October 2016 from 24 areas within Bangkok and surrounding provinces. Data contains:

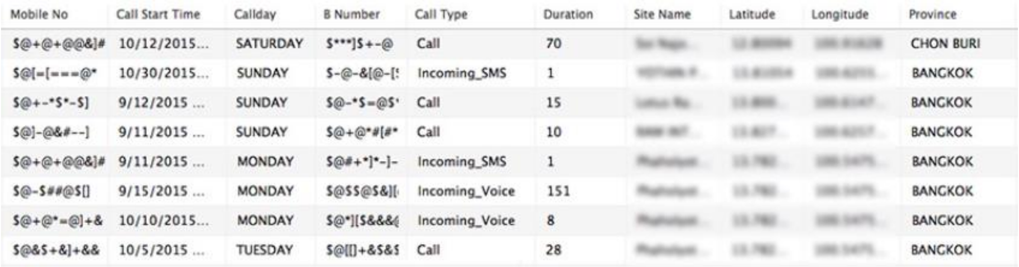
moblile_no is encrypted mobile phone number that connected to cell tower

call_start_time is time that mobile phone starts communicating to cell tower

call_type is type of usage that are separated into 4 types: Incoming/Outgoing Voice and Incoming/Outgoing SMS

duration is duration of call, appears to be a constant 1 for every SMS data

latitude is latitude of cell tower that communicate to the phone when the call initiated longitude is longitude of cell tower that communicate to the phone when the call initiated



Mobile No	Call Start Time	Callday	B Number	Call Type	Duration	Site Name	Latitude	Longitude	Province
$@+@+@@&]#	10/12/2015...	SATURDAY	$***]$+-@	Call	70	[redacted]	[redacted]	[redacted]	CHON BURI
$@[=[===@*	10/30/2015...	SUNDAY	$-@-&[@-[!	Incoming_SMS	1	[redacted]	[redacted]	[redacted]	BANGKOK
$@+-*$*-$]	9/12/2015 ...	SUNDAY	$@-*$=@$'	Call	15	[redacted]	[redacted]	[redacted]	BANGKOK
$@]-@&#--]	9/11/2015 ...	SUNDAY	$@+@*#[#*	Call	10	[redacted]	[redacted]	[redacted]	BANGKOK
$@+@+@@&]#	9/11/2015 ...	MONDAY	$@#+*]*-]-	Incoming_SMS	1	[redacted]	[redacted]	[redacted]	BANGKOK
$@-$##@$[]	9/15/2015 ...	MONDAY	$@$$@$&][	Incoming_Voice	151	[redacted]	[redacted]	[redacted]	BANGKOK
$@+@*=@]+&	10/10/2015...	MONDAY	$@*][$&&&(	Incoming_Voice	8	[redacted]	[redacted]	[redacted]	BANGKOK
$@&$+&]+&&	10/5/2015 ...	TUESDAY	$@[[]+&$&$	Call	28	[redacted]	[redacted]	[redacted]	BANGKOK

Figure 6 Sample of CDR Data

3.2. Study Area

The study area contains with 24 locations in Bangkok Metropolitan including Suvarnabhumi airport. Each location is a 600 square meters grid. Each location may contain from 1 up to more than 20 cell towers. And after filtering null data, there are 136 active cell towers used in this study.

Location 1-4 are transportation hubs such as airports, location 5-7 are government bureaus, location 8-12 are schools and location 13-24 are shopping sites from outskirts and downtown area.

Some locations consist of rapid transit station which refers to BTS Sky Train and MRT station. Airport Rail Link is not included in this study.

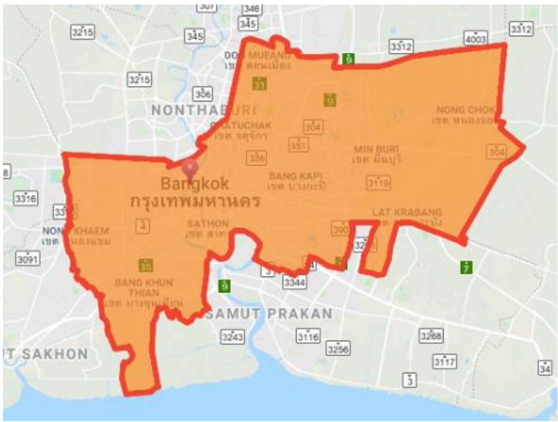


Figure 7 Boundary of Bangkok Metropolitan where 24 study locations located in.

3.3. Data Exploration

3.3.1. Number of Records

From collected data, there are 129,585,962 records gathered during 1st August – 31st October 2016. Average (Mean) amount of overall records from 24 locations is 5,399,415. Standard deviation (SD) of overall records from 24 locations is 4,235,880. Coefficient of variation (CV) which is $\frac{SD}{MEAN} = 0.78$.

This indicates that number of records of locations are quite spread out from the mean, refers to difference of usage amount between different locations.

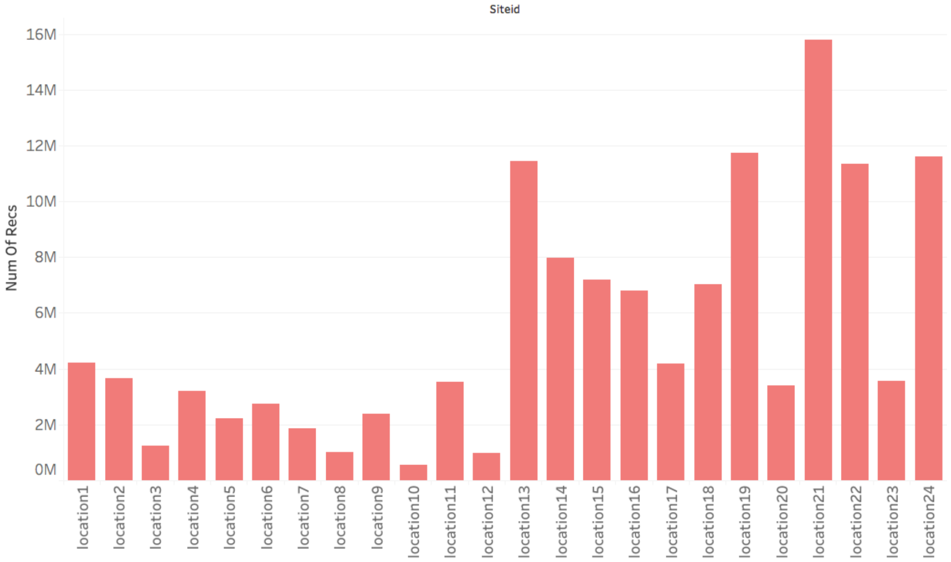


Figure 8 Graph of mobile usage represents number of records obtained from location 1 to location 24 (X-axis)

Figure 8 displays amount of number of records by day of week, shows that usage amount at most location tend to decrease in the weekend. However, some locations such as cluster 23, the number of records increase in the weekend. The lower dash line highlights number of records from Monday to Thursday of location 23, while the upper dash line highlights number of records in the weekend. From picture, usage amount at location 23 tend to increase on Friday and Saturday. Then, number of records is slightly drop on Sunday but still more than weekdays.

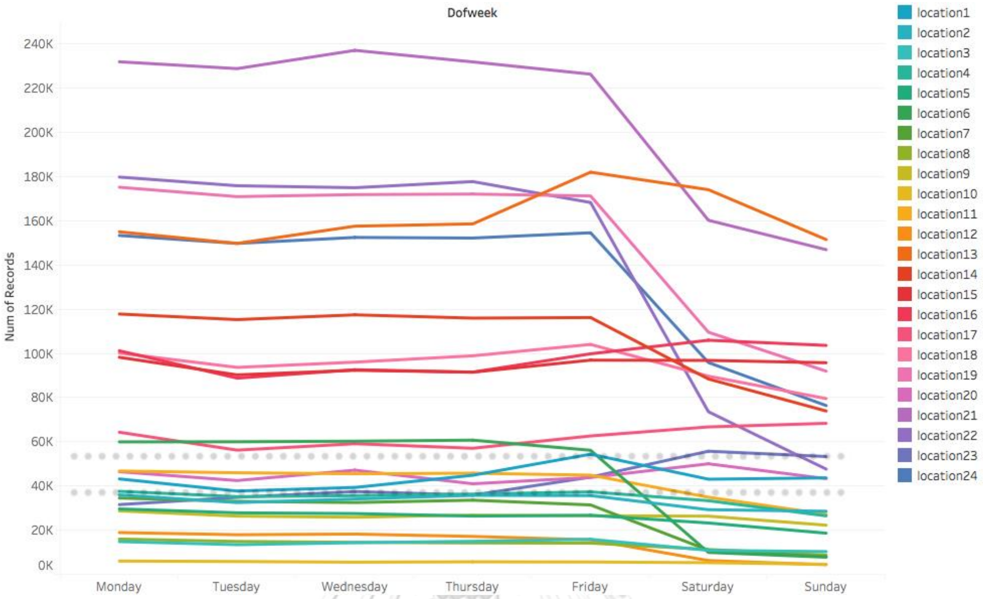


Figure 9 Number of records in each location from Monday to Sunday.

3.3.2. Distinct Phone Numbers

From collected data, there are 2,602,401 distinct phone numbers gathered during 1st August – 31st October 2016. Figure 10 displays amount of distinct numbers from location 1 to 24. Average (Mean) amount of distinct phone numbers from 24 locations is 108433. Standard deviation (SD) of distinct phone numbers from 24 locations is 80626. Coefficient of variation (CV) which is $\frac{SD}{MEAN} = 0.74$.

This indicates that number of records of locations are slightly spread out from the mean less than number of records. However, these two values are very close, both refers to difference of usage amount between different locations.

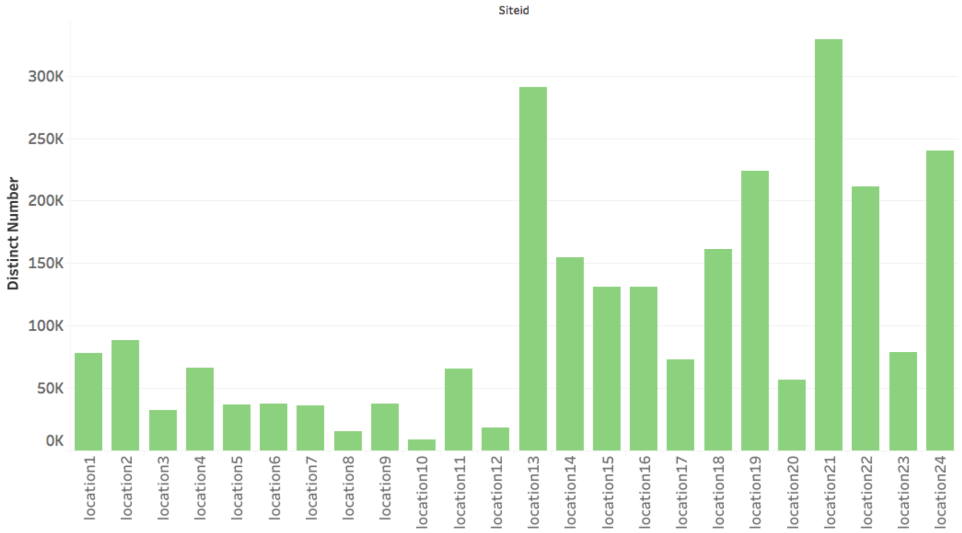


Figure 10 Distinct phone numbers obtained from location 1 to location 24 (X-axis)

Figure 11 displays amount of distinct phone numbers by day of week, shows that distinct phone numbers tend to be constant from Monday to Thursday. Distinct phone numbers at some locations increase in the weekend, while others decrease. However, most locations have peak distinct numbers on Friday.

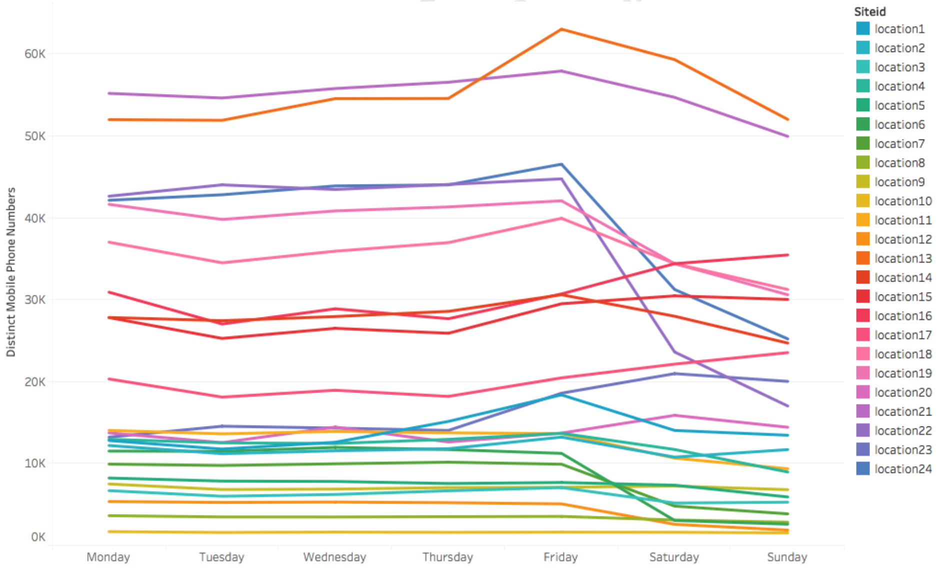


Figure 11 Graph shows amount of distinct phone numbers in each location from Monday to Sunday

3.3.3. Call Duration

Call duration is average duration of phone calls happened in a space of time-location. Due to the fact that call duration of SMS is constant 1, call duration is calculated from voice call only. Figure 12 displays average call duration from location 1 to location 24. Average call duration of overall data is 94.92 seconds.

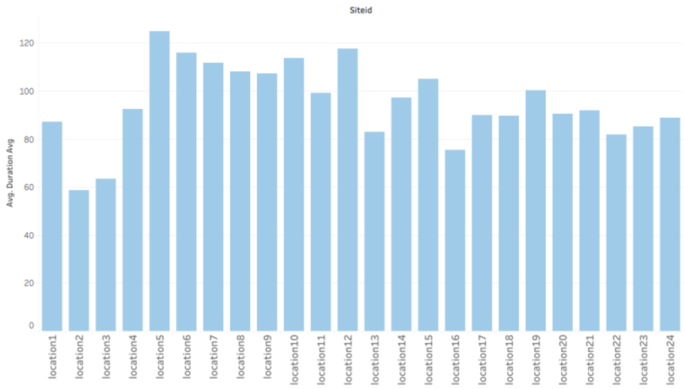


Figure 12 Mobile usage represents average duration obtained from location 1 to location 24 (X-axis)

Call duration is very swing. Most locations have high call duration on Tuesday and Friday. Some locations are different. However, there are also some locations that call duration is very stable such as location 2.

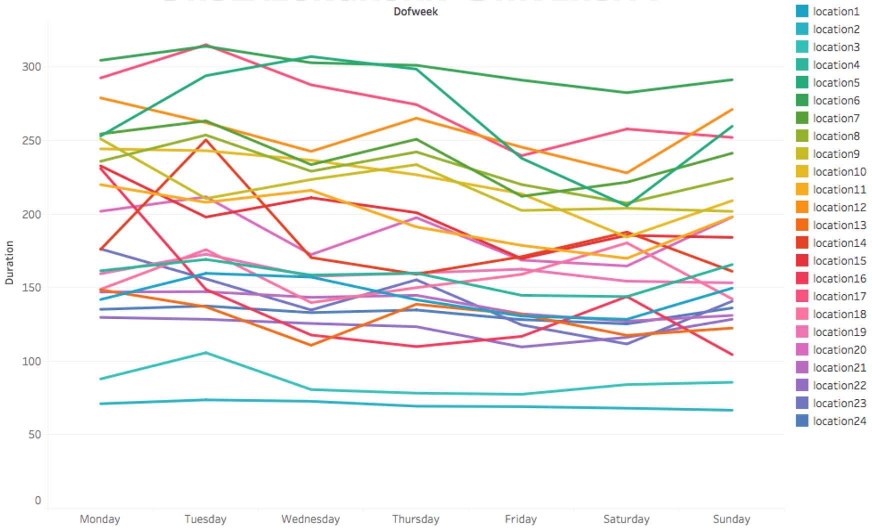


Figure 13 Average duration in each location from Monday to Sunday

3.4. Mobile Usage Pattern

We put usage record into coarse period of hour and day of week. The 2-dimension matrix contributes from 24 hour of day - and 7 days of week = 24x7 giving an outcome of 168 values. Each value represents usage amount during a period.

We contribute 3 types of matrix from 3 measurements - number of records, distinct phone numbers and call duration.

For a better visualization, we construct matrices from normalized value of 3 measurements which are number of records, distinct phone number and average duration.

Figure 14 displays sample of heat map from mobile usage pattern from number of records. Y-axis refers to day of week from Monday to Friday. X-axis refers to hour of a day, each block means a period of an hour. Green color means low value of measurement while stronger red means higher value of measurement. In this study, we visualize each measurement separately.

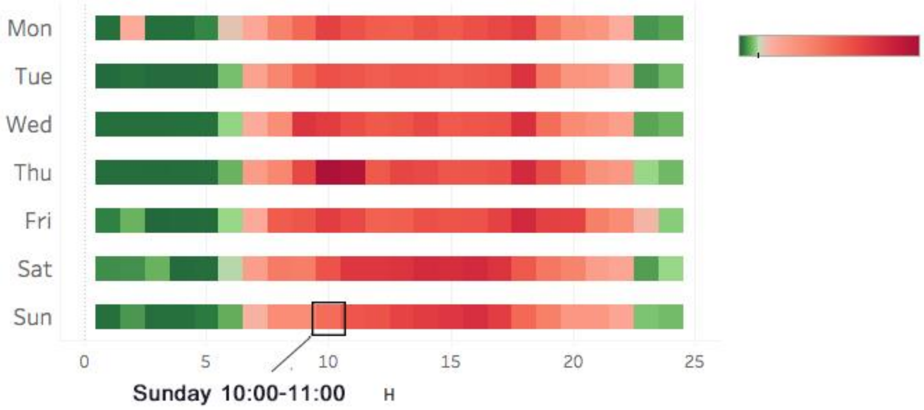


Figure 14 Sample of Mobile Usage Pattern that displays in Heat Map. Each block refers to an hour period in day of week.

3.5. Dimensionality Reduction

After constructed matrices from 3 measurements for every single cell tower, each measurement contains 24x7 = 168 variables refers to usage amount in a period of an hour-day of week. Overall of usage pattern which is combined from three measurements has many variables. It is possible that some variables might be correlated.

In order to reduce the number of variables, overall of mobile usage is combined from three measurements and then processed through Principal Component Analysis (PCA) method.

PCA is used to transform original variable into linear combination and calculate new variable called principle component or PC. These new variables are used for further process instead of original variables.

The number of PC is determined by cumulative variance and cumulative plot from PCA. From Figure 15, Y-axis means cumulative proportion of variance while x-axis means number of PC. The value of cumulative variance refers to percentage of data explained by number of PC.

This means approximately 3 PCs can represent 70% of data and 10 PCs can represent approximately 80% of data. For this reason, 5 PCs are then used for further process instead of original variables.

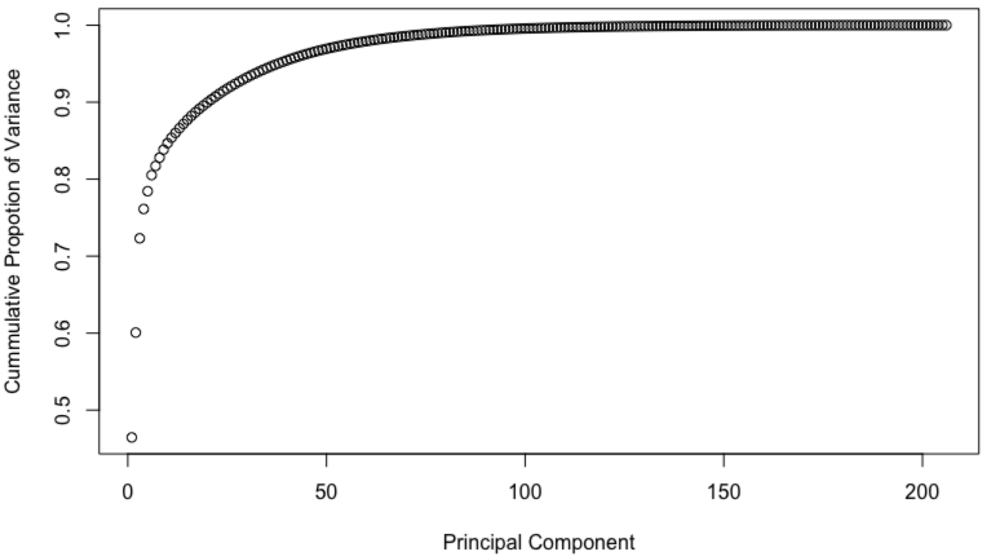


Figure 15 Cumulative plot from PCA means percentage of data that can be referred by number of PC. This reflects that 3PCs can refer around 70% of data and 10 PCs can refer approximately 80% of data.

3.6. Pattern Clustering

Considering every cell tower, k-means clustering algorithm is applied to group cell towers with similar mobile usage pattern together.

Tree diagram is difficult to interpret in this case and the elbow plot also gives ambiguous crucial point. Thus, the appropriate number of cluster is determined by gap statistic method. The result from gap statistic method gives the optimal number of cluster when k is 9.

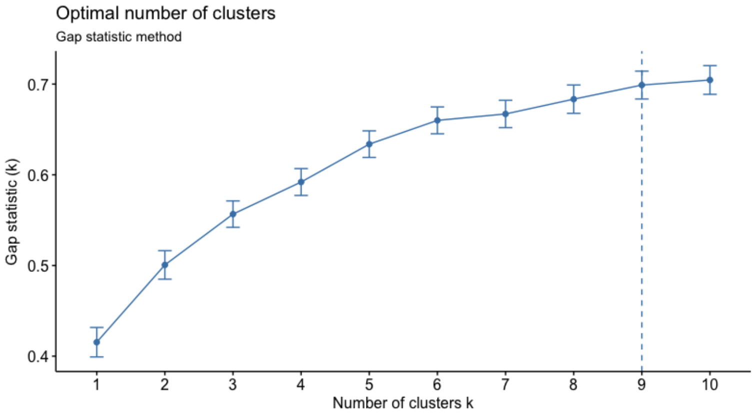


Figure 16 Graph from gap statistic method gives an optimal cluster at k=9

3.7. Cell Tower Prediction Model

After finding the optimal number of clusters and clustering train data set, we then create SVM prediction model for cell tower itself from clustering result. This model will use to particularly predict cluster of cell tower from mobile usage pattern.

In order to predict type of cell tower, we define cluster as type of cell towers and transform raw data into observations of usage pattern by hour and day of week.

Table 1 shows prediction model by using usage pattern in time period as observations in or to predict type (cluster) of cell tower

	Monday 0:00-1:00	Monday 1:00-2:00	Monday 2:00-3:00	Monday 3:00-4:00	Monday 4:00-5:00	———	Cluster (Type of Cell Tower)
Cell 1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$	$X_{1,5}$	$X_{1,k}$	Y_1
Cell 2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$	$X_{2,5}$	$X_{2,k}$	Y_2
Cell 3	$X_{3,1}$	$X_{3,2}$	$X_{3,3}$	$X_{3,4}$	$X_{3,5}$	$X_{3,k}$	Y_3
Cell 4	$X_{4,1}$	$X_{4,2}$	$X_{4,3}$	$X_{4,4}$	$X_{4,5}$	$X_{4,k}$	Y_4
Cell n	$X_{n,1}$	$X_{n,2}$	$X_{n,3}$	$X_{n,4}$	$X_{n,5}$	$X_{n,k}$	Y_n

3.8. Corresponding Ratio of Cluster

After gathering cell towers with similar mobile usage pattern together by K-means clustering method, the result gives cluster number of cell towers and shows 9 cluster that represent different usage pattern. However, each study location consists of different clusters. Some locations contain cell towers within a single cluster while others contain cell towers from multiple clusters. And amount of usage within each cluster may also vary in area use of locations.

Table 2 Sample of mobile phone records after receive cluster result

Study Site	Cell Tower ID	Cluster	Number of Records
Location 1	001	5	40000
Location 1	003	5	20000
Location 1	004	6	30000
Location 1	005	7	10000

Location 2	006	8	4000
Location 2	007	8	1000

Table 3 Data after group Table 2. by clusters

Study Site	Cluster	Number of Records
Location 1	5	60000
Location 1	6	30000
Location 1	7	10000
Location 2	8	5000

Table 4 shows sample overall numbers of mobile phone in locations

Site	Number of Records
Location 1	100000
Location 2	5000

For this reason, we calculate number of records of each cluster and calculate into ratio compare to overall number of records in each location. Corresponding Ratio (Number of records). We also do the same thing with distinct numbers.

Table 5 result of corresponding ratio calculated from usage in each cluster comparing to overall usage of location

Study Site	Cluster	Corresponding Ratio (Number of records)
Location 1	5	60000/100000 = 0.6
Location 1	6	30000/100000 = 0.3
Location 1	7	10000/100000 = 0.1
Location 2	8	5000/5000 = 1

3.9. Probability Vector of Location

Some locations may consist of cell towers from multiple clusters. At the same time, active periods of clusters are different. Periodically, numbers of records from a cluster compare to overall in a location, show corresponding amount of cluster in a period.

For example, location 5 may have many cell towers from two clusters. In table 6 shows sample that there are two clusters in location 5. Cluster 1 and cluster 2 have different active hour. Cluster 1 is more active during 9:00-10:00 while cluster 2 is more active in the evening during 16:00-17:00.

Table 6 Sample data shows number of records of location 5 in each cluster by time period

		Hour	Number of Records
Location 5	Cluster 1	9:00-10.00	90
Location 5	Cluster 2	9:00-10:00	10
Location 5	Cluster 1	16:00-17:00	3000
Location 5	Cluster 2	16:00-17:00	7000

Table 7 shows sample overall numbers of record in a location by time period

	Hour	Number of Records
Location 5	9:00-10:00	100
Location 5	16:00-17:00	10000

↓

Table 8 Result of probability in each cluster by time period

	Probability		
	Time period	Cluster 1	Cluster2
Location 5	9:00-10:00	90/100 = 0.9	10/100 = 0.1
Location 5	16:00-17:00	3000/10000 = 0.3	7000/10000 = 0.7

Periodically, probability of visit at a location should add up to one. For this reason, the probability that a person at location $\boldsymbol{A}$ is related to cluster $\boldsymbol{i}$ within time period $\boldsymbol{t}$ is

$$P_t(i, A) = \frac{Number\ of\ records\ in\ location\ which\ are\ in\ cluster\ i\ in\ time\ period\ t}{Total\ number\ of\ records\ in\ location\ in\ time\ period\ t}$$

The probability vector $\boldsymbol{p}$ of location $\boldsymbol{A}$ constructed of probability $\boldsymbol{P_t(i, A)}$ from $\boldsymbol{i = 1\ to\ k}$ where k is an optimal number of cell tower cluster, explains possibility that people would be related to each cluster within a location at the same time.

$$p_t(A) = \begin{bmatrix} P_t(1) & \dots & P_t(9) \end{bmatrix}$$

For example, the result in Table 8 shows that people seems to be related to cluster 1 during 9:00-10:00. But during 16:00-17:00, people tend to be related to cluster 2.

If k is 4, the probability vector for location 5 between 9:00-10:00 is

[0.9 0.1 0 0 0 0 0 0 0]

3.10. Area Classification

From 24 study locations, they can be roughly assigned into 4 different area types. And we also define 4 types of rapid transit availability in the location which are BTS (sky train), MRT (subway), both and none.

In order to build a classifier, we construct cluster composition from cluster ratio, we then perform SVM to build classifiers for area type and transportation facility from corresponding ratio model because support vector machine (SVM) can work with imbalance and non-linearly separable data.

Table 9 shows sample of classification model from cluster composition. Value of corresponding ratios are from Table 5.

Table 9 result of corresponding ratio calculated from usage

	Observations of Cluster Composition									Features	
Location	Clust 1	Clust 2	Clust 3	Clust 4	Clust 5	Clust 6	Clust 7	Clust 8	Clust 9	Transport	Area Use
1	0	0	0	0	0.6	0.3	0.1	0	0	BTS	A
2	0	0	0	0	0	0	0	1	0	MRT	B

3.11 Implementation Diagram

From a location, amount of usage data from cell towers within location are separated into each hour-day of week. These values are transformed into set of variables and then processed to reduce variable dimension. The outcome is Principal Component (PC), which is used instead of original variables, to find groups of cell towers with similar usage pattern.

Amount of usage from cell towers within a location are summarized by cluster types and calculated ratio of usage amount of each type compare to overall. These values are used to construct a model like in Table 9 to classify area type of a location.

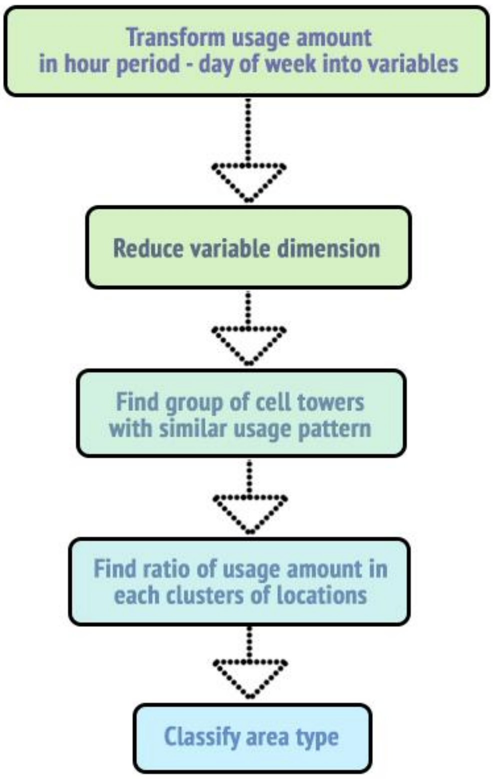

Transform usage amount in hour period - day of week into variables
Reduce variable dimension
Find group of cell towers with similar usage pattern
Find ratio of usage amount in each clusters of locations
Classify area type

Figure 17 Diagram displays implementation steps

Chapter 4

Result and Evaluation

4.1. Research Environment

Since telecommunications operator must collect large amount of CDR. development tools in this research which connected to database are also designed for big data. We use SQL to connect with database and then process statistical analysis with R language.

- Impala is an open source distributed massively parallel processing (MPP) analytic database. Data are stored in Hadoop Distributed File System (HDFS). User can connect with HDFS using SQL faster than other database engine.
- csv (comma-separated values) is a file format used to store tabular data. Each line is a data record while comma separate fields of data into columns.
- Tableau is a visual analytic tool that helps to create interactive visualization for data analysis.

4.1.1. R Packages

We select R language because there are several libraries that are necessary for data analysis. Most packages are related to statistical computing. These are some additional packages:

factoextra - is an R package used to process and visualize multivariate data analysis.

limma - is a library that originally build for microarray data analysis. It contains a lot of effective statistic functions for data analysis.

e1071 - is a R package used for machine learning process.

Plotly - is a package for interactive data visualization.

4.2. Visualization

4.2.1. Cluster Result

Study Site/Cluster	Clust1	Clust2	Clust3	Clust4	Clust5	Clust6	Clust7	Clust8	Clust9
1	-	✓	-	-	-	-	✓	-	-
2	-	-	-	-	-	✓	-	-	-
3	-	✓	-	-	-	-	-	-	-
4	✓	✓	-	-	-	-	-	-	-
5	-	✓	-	-	-	-	-	-	-
6	-	-	-	✓	✓	-	-	-	-
7	-	-	-	✓	-	-	-	-	-
8	-	✓	-	-	-	-	-	-	-
9	-	✓	-	-	-	-	-	-	-
10	-	✓	-	-	-	-	-	-	-
11	-	✓	-	-	-	-	-	-	-
12	-	-	-	✓	-	-	-	-	-
13	-	-	✓	-	-	-	✓	✓	-
14	✓	✓	✓	-	✓	-	✓	-	-
15	✓	✓	✓	-	-	-	✓	✓	-
16	-	-	-	-	-	-	-	✓	-
17	-	-	-	-	-	-	-	✓	-
18	✓	-	✓	-	✓	-	✓	✓	-

19	✓	-	✓	✓	-	✓	-	✓	-
20	-	-	-	-	-	-	✓	✓	-
21	✓	-	✓	✓	-	-	-	✓	-
22	✓	-	-	✓	✓	-	-	-	-
23	✓	-	-	✓	-	-	✓	-	✓
24	✓	✓	✓	-	✓	-	✓	✓	-

From gap statistic method and observation, mobile phone usage pattern represents the best result at k=9. Table below shows composite clusters of 24 study locations.

Table 10 displays composite clusters of 24 study locations

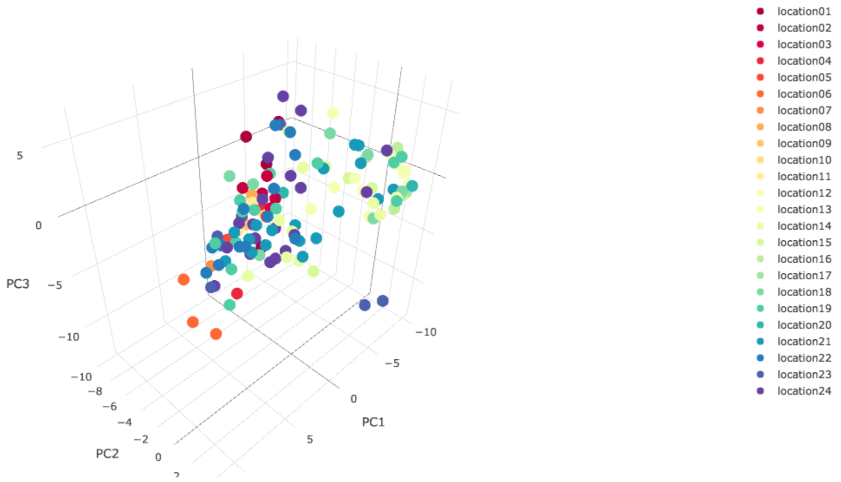


Figure 18 Three-dimension scatter plot shows dispersion of clusters from locations

This 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 mean that they are obtained from same location. Most color are located nearby. However, some locations have distributed cell towers

Some clusters such as cluster 8 seem likely to separate cell towers into two groups. In fact, we also use cluster ratio that calculated from corresponding number of records and distinct numbers, to determine cluster result.

From Heat map of usage pattern, distinct numbers seem obviously correspond to number of records. Meanwhile we did not find call duration is in the other direction. In most clusters, duration of call drop during the day and increase at night. However, cluster 6 is an only cluster that have noticeably high duration of call in daytime.

Number of records and distinct number which relatively refer to amount of people, show that people are likely to be active from 7am to about 8 or 9pm.

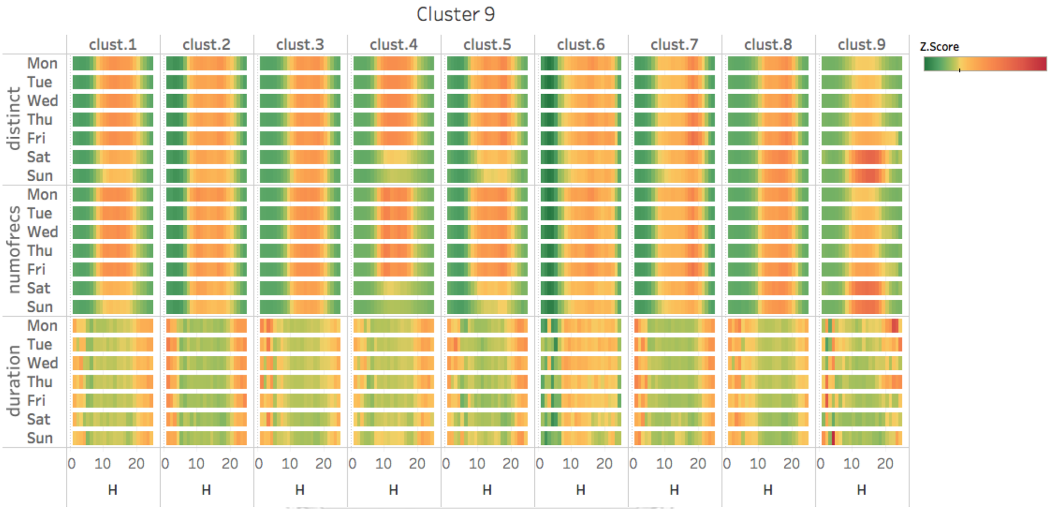


Figure 19 Heat map displays mobile phone usage pattern from cluster 1 to cluster 9

Cluster 1 contains with cell towers from location 4, 14, 15, 18, 19, 21, 22, 23 and 24. Heat maps of usage behavior in this cluster look like squares with faded tail because usage amount tends to decrease on Saturday and Sunday. During weekdays, people are likely to be active in this area from 9am to 7pm. But the start hour tends to be more late than usual in the weekend.

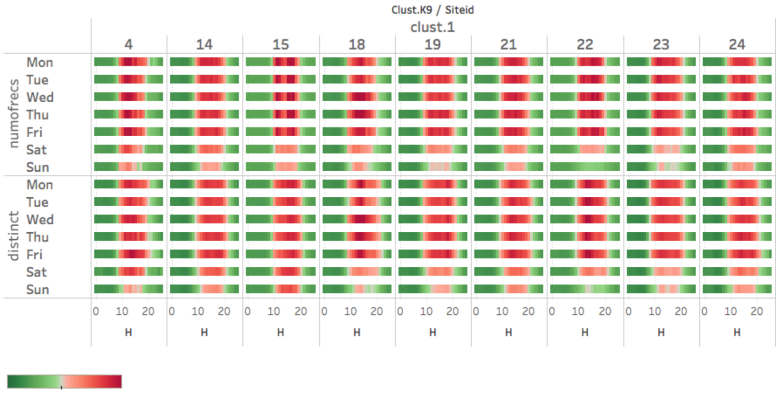


Figure 20 shows heat map of usage pattern from cluster 1

Mobile usage patterns from cluster 2 is a square with more obvious tail than cluster 1 since usage amounts in the weekend do not reduce much. In this cluster, usage amount on weekday is highly active around 11am and 7pm. These two periods create an afternoon gap in usage pattern.

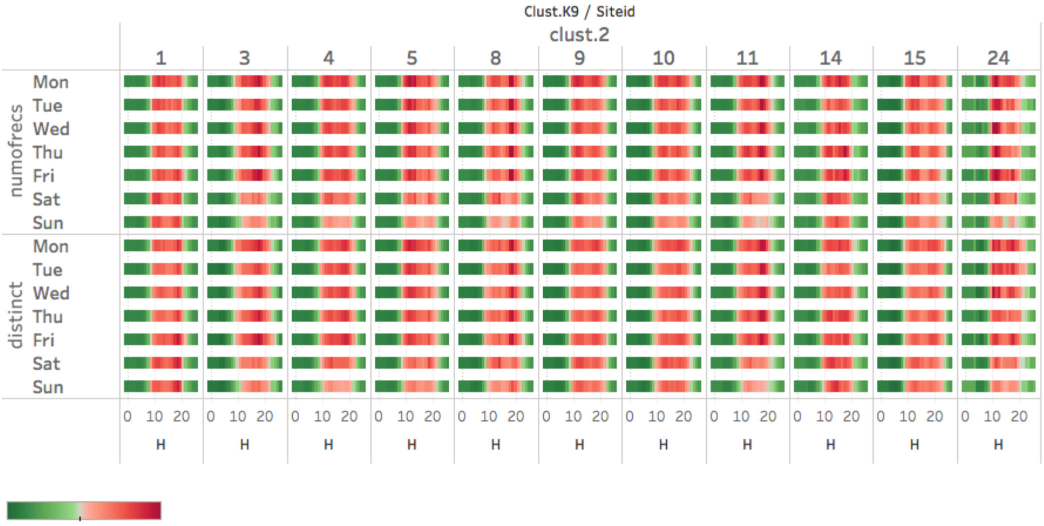


Figure 21 shows heat map of usage pattern from cluster 2

In cluster 3, usage amount in the weekend seem to be higher than cluster 1 and 2. And usage amount during daytime is smoother as well. There is no obvious afternoon gap like cluster 2.

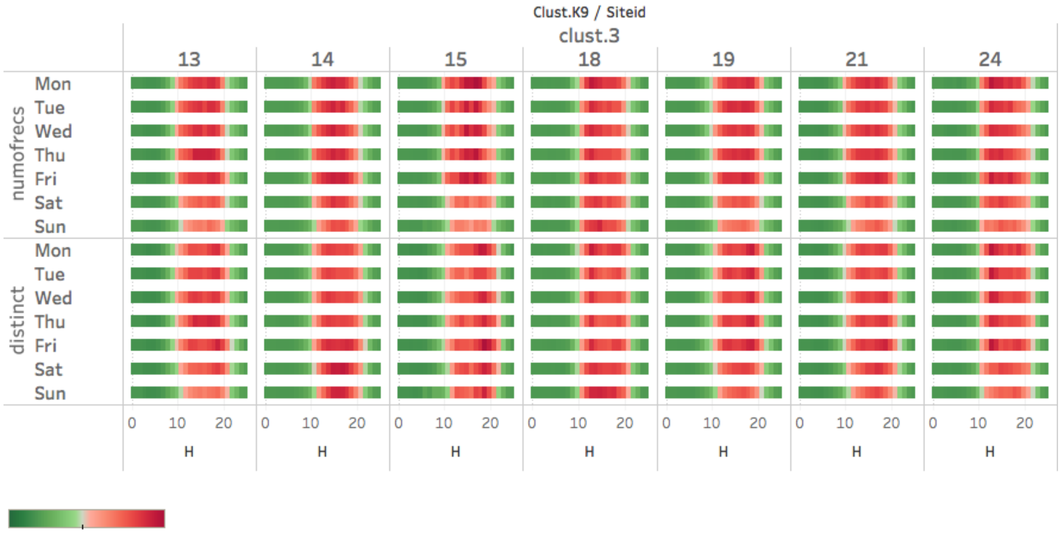


Figure 22 shows heat map of usage pattern from cluster 3

Usage patterns from cluster 4 shows obvious square since usage behavior seems to be barely active in the weekend. There is also gap period at noon because usage amount in this cluster is highly active around 11am and 3pm.

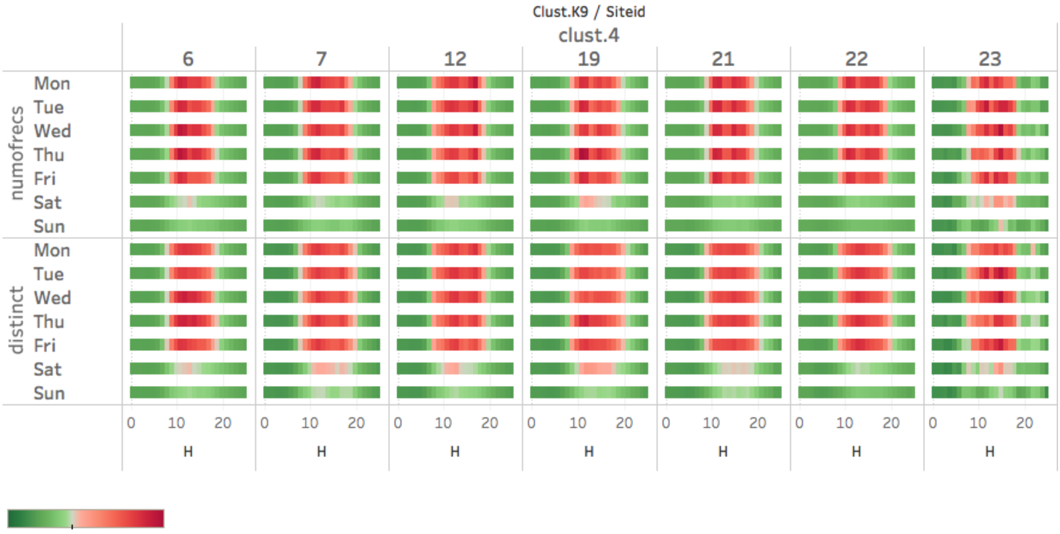


Figure 23 shows heat map of usage pattern from cluster 4

In cluster 5, usage pattern is also obvious square like in cluster 4. However, the gap period is not distinctive as in cluster 4 and usage behavior seems to be more active in evening of weekdays.

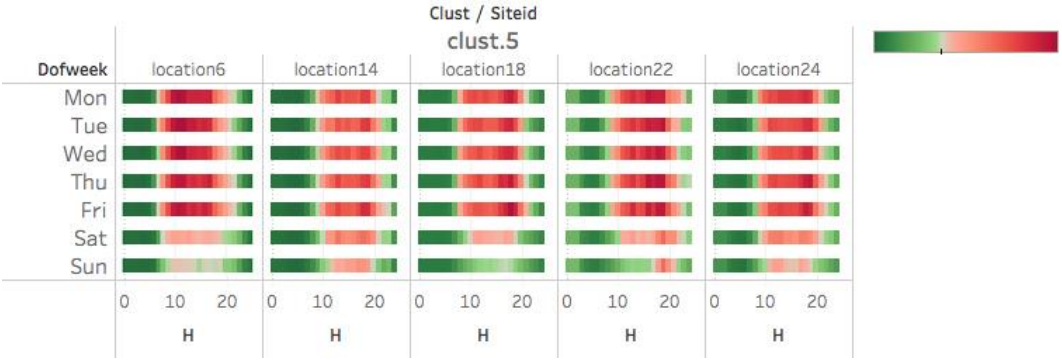


Figure 24 shows heat map of usage pattern from cluster 5

Cluster 6 consists of location 2 and 19. Usage pattern in this cluster seem to have wide spread active period from approximately 7am to 11pm. However, peak time is quite scatter. In location 2, the highest usage amount seems is at 4pm. And the second highest is at 9pm. At the same time, usage behavior in location 9 seem to be active the most at 12pm and 8pm. Considering from peak time, it might say that cluster 6 also has a low active gap. Though it appears in different time at different location.

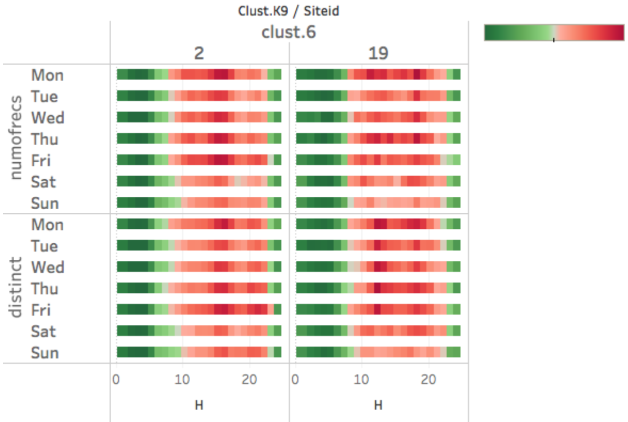


Figure 25 shows heat map of usage pattern from cluster 6

Usage pattern in cluster 7 also looks like square with faded tail like in cluster 1. However, the most active hour appears to be in the evening approximately at 7pm.

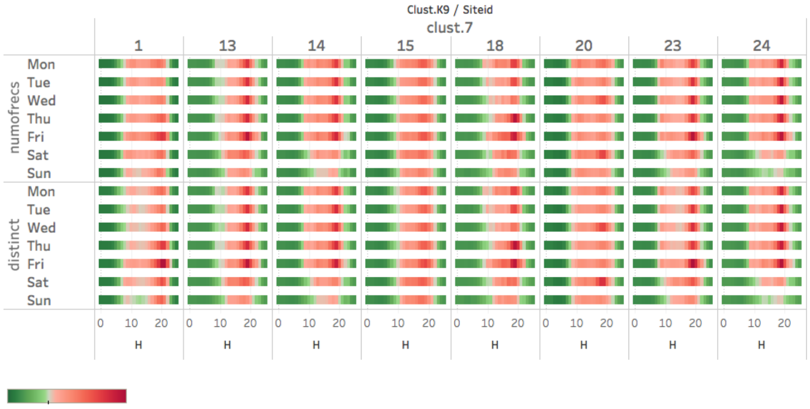


Figure 26 shows heat map of usage pattern from cluster 7

In cluster 8, usage pattern shows square with lucid tail since active amount in the weekend is higher than weekday. While in cluster 9, usage pattern is also higher in the week end. But it also has noticeably low usage amount during weekdays.

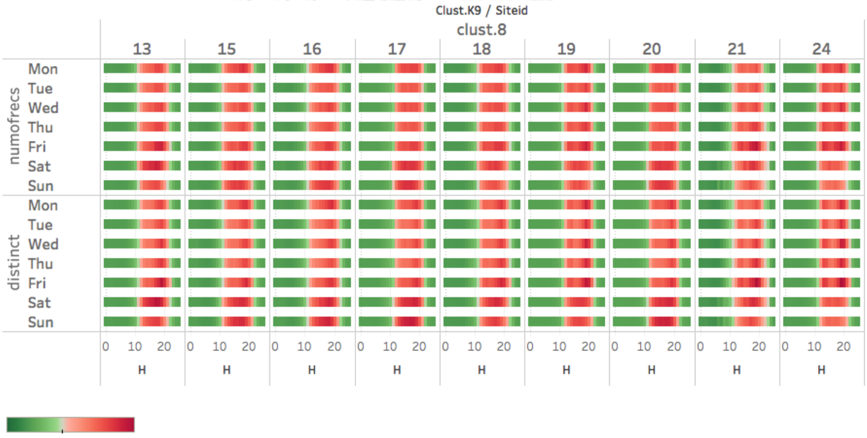


Figure 27 shows heat map of usage pattern from cluster 8

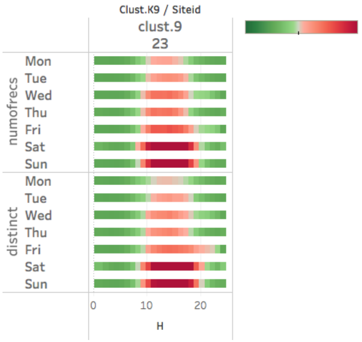


Figure 28 shows heat map of usage pattern from cluster 9

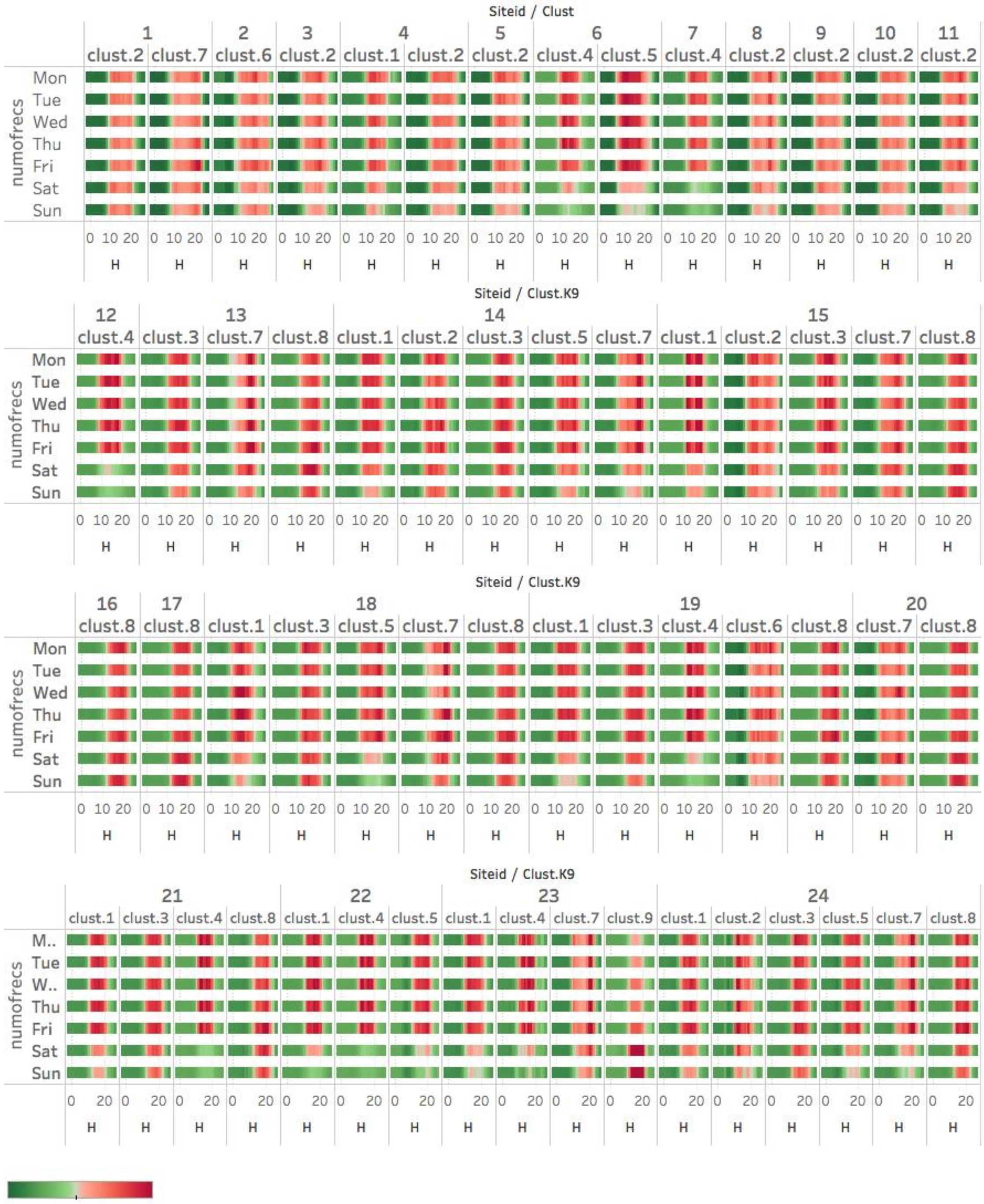


Figure 29 Heat map displays mobile phone usage pattern from each cluster in 24 locations

4.2.2. Active Curve from Probability Vector of Location by Time Period

In some locations that consists of multiple clusters, there is a dominate cluster that is much more active than others. From Figure 24, There are two clusters in location1. Cluster 2 and cluster 7, however, cluster 7 seems to be much more active during day time. This means that usage happens in location 1 during day time tend to be related to cluster 7.

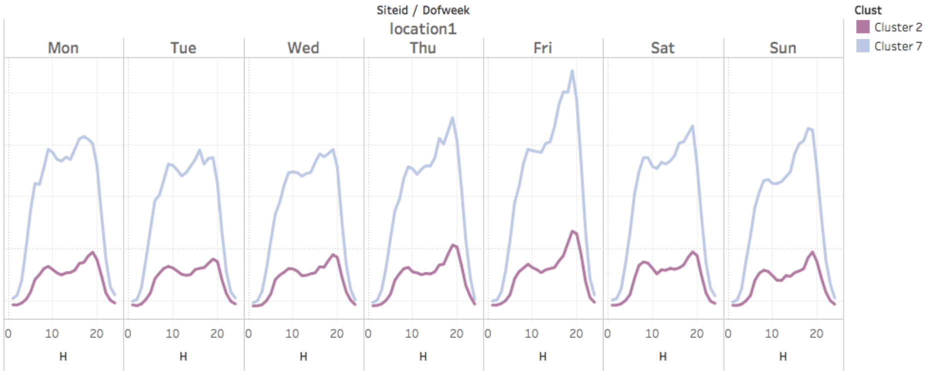


Figure 30 Graph shows active curve from usage pattern within location 1 by comparison.

Another active curve from location 20 shows noticeably different active period of two cell towers. In the morning, especially, before 10am, usage during this time tend to be related to cluster 8. While usage in the afternoon, people are likely to active within cluster 7.

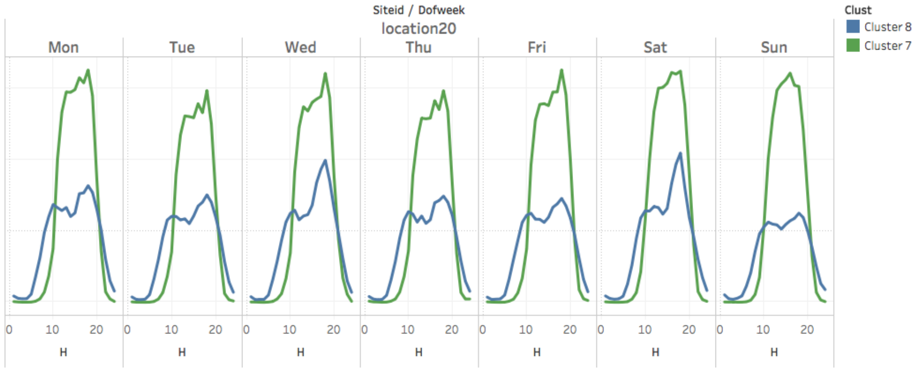


Figure 31 Active curve shows usage amount within location 20 by comparison between cluster 7 and cluster 8

4.3. Evaluation

4.3.1. Result from Feature Classification

We use cluster composition as observations to perform SVM in order to build a prediction model. In this case, we calculate both number of records and distinct numbers to construct cluster composition. We also create another cluster composition combined ratio from number of records and distinct phone numbers which measure amount of usage. Later then, we build a model to classify area use and transport facility such as rapid transit station.

In this case, we roughly label study locations by 4 types of area use which are inter-city transportation hub, bureau, school and shopping mall. We also label transportation facility into 4 types which are BTS (sky train), MRT (subway), both and none. Table 11 displays confusion matrix from the prediction model. Most shopping sites and schools can be correctly predicted. However, some schools and government bureau seem to have a misclassification between each other.

Table 11 Confusion matrix from prediction model

	Actual			
Predict	Bureau	School	Shopping	Trans-hub
Bureau	0	1	0	0
School	2	4	0	1
Shopping	1	0	10	0
Trans-hub	0	0	0	2

Total accuracies of prediction models from cluster compositions and raw data that are evaluated by 5-fold cross validation represent in the table below.

Table 12 Percentage of total accuracies from prediction model using mobile phone usage pattern to predict area use and transport facility

	Percentage of Total Accuracy	
	Area Use	Transport Facility
Cluster Composition (By Number of Records Ratio)	66.7	62.5
Cluster Composition (By Distinct Numbers Ratio)	75	62.5
Cluster Composition (Combine both two ratio)	75	62.5
Raw Number of Records	50	62.5
Raw Distinct Numbers	50	62.5

For transportation facility insights, we try separating into 4 different features by combining each rapid transit with both. We also try to predict some other answers in case locations have only single rapid transit station. And the last feature is yes/no answer for rapid transit station.

Table 13 Total accuracies of prediction model to predict transportation facility availability in different cases

	Percentage of Total Accuracy			
	BTS or Both / MRT	BTS/MRT or Both	Single/Both	Yes/No
Cluster Composition (By Number of Records Ratio)	62.5	58.3	70.1	83

Cluster Composition (By Distinct Numbers Ratio)	66.7	62.5	66.7	75
Cluster Composition (Combine both two ratio)	79.2	62.5	75	95.9
Raw Number of Records	62.5	62.5	62.5	62.5
Raw Distinct Numbers	62.5	62.5	62.5	62.5

4.4. Example of Mismatch Location

4.4.1 Location Traceback

To find out the sources of errors, we generate a dendrogram plot from cluster composition. Tree diagram shows that location 7 and location 12 are in the same clusters. Actually, Label of these two locations are different since location 7 is bureau but location 12 is school.

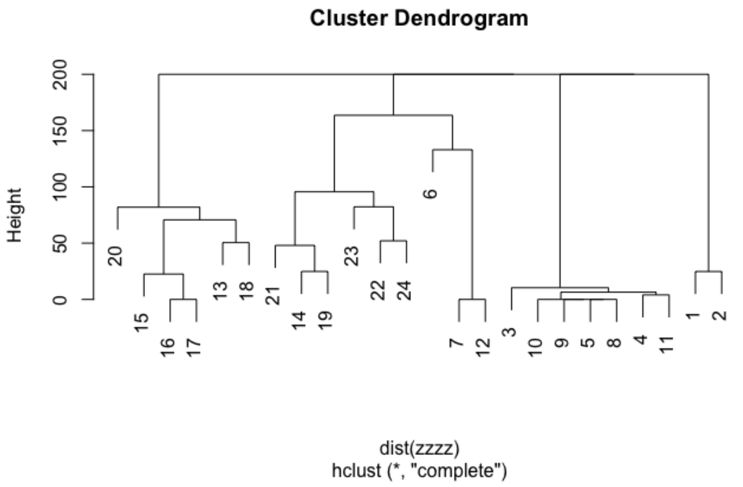


Figure 32 Tree diagram shows linkage between locations that have similarity between them

However, when we look into heatmap and cluster composition. Both also contains cell towers within cluster 4 only. The heat map also shows similar pattern between location 7 and location 12. While location 8, 9 and 11 which are school, have significant difference in pattern. This means that there are some locations which do not completely match with others in a same type.

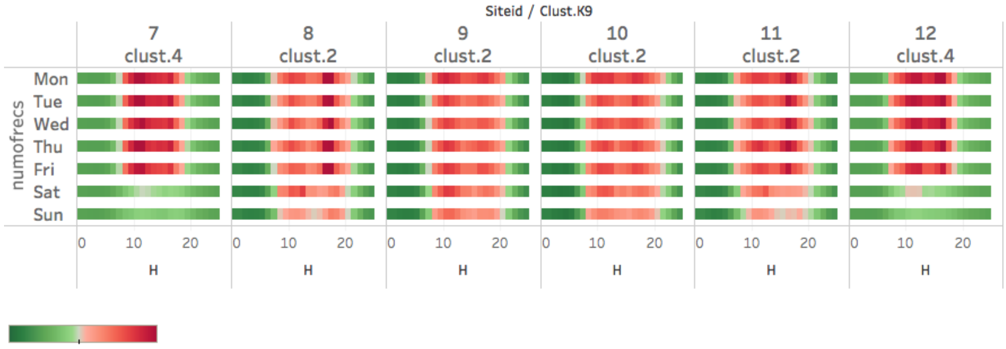


Figure 33 Heat map shows clusters and usage pattern of location 7, 8, 9 10 and 12. Although location is a school, usage pattern from this location is similar to location 7 which is a bureau.

4.5. Area Use Insights

Some type of locations such as large shopping mall gather many visit points together and offers convenience in many ways. Most of locations in this type also contain cell towers in multiple clusters. From tree diagram, location 13-24 which are labelled as shopping mall, can be separated into large 2 types. Right branch of this diagram consists of location 14, 19, 21, 22, 23 and 24. All of these locations have rapid transit station. However, location 13, 18 and 20 in the left branch also have rapid transit station. Others are not. Considering the transportation availability, optimal number of cluster at 6 can properly split shopping malls into minor groups.

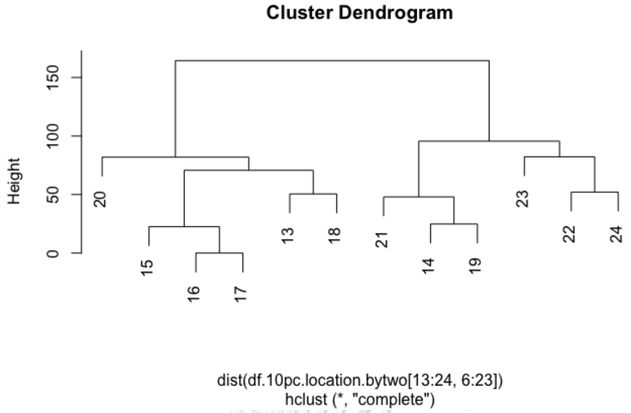


Figure 34 Dendrogram shows linkage of shopping locations that have similarity. From tree diagram, the closest locations have the most similar usage pattern.

Chapter 5

Conclusion and Future Works

5.1. Conclusion

This research use CDR to study mobile phone usage pattern in different area types and implement a model to classify area use and facility availability. To build a model, we find out cluster ratio within a location. And then transform cluster ratio into observations to form a prediction model.

This model is useful for telecommunication operator in order to classify area types from overall country. This also can be applied in marketing and urban planning.

5.2. Future Works

It is undoubted that single-use area type can be distinguished from mobile usage pattern. In fact, there are plenty of mixed-use locations. These locations generally consist of many cell towers with different patterns and difficult to classify in detail. Moreover, it can be assumed that some cell towers have mixed-use characteristic as well.

Besides, we also find out that some locations do not have similar usage pattern to others in a same type. This problem might be able to solve by finding cross possibility between each area type from big data set. However, we use limited number of locations that be able to survey surroundings in this research. There would be a limitation in accuracy

REFERENCES

1. *National Statistical Office*. 2017; Available from: http://statbbi.nso.go.th/.
2. Hasan, S., X. Zhan, and S.V. Ukkusuri, *Understanding urban human activity and mobility patterns using large-scale location-based data from online social media*, in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. 2013, ACM: Chicago, Illinois. p. 1-8.
3. Phithakkitnukoon, S., et al. *Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data*. 2010. Berlin, Heidelberg: Springer Berlin Heidelberg.
4. Hsu, C.W., C.C. Chang, and C.J. Lin, *A Practical Guide to Support Vector Classification*. Vol. 101. 2003. 1396-1400.
5. Isaacman, S., et al., *A tale of two cities*, in *Proceedings of the Eleventh Workshop on Mobile Computing Systems &#38; Applications*. 2010, ACM: Annapolis, Maryland. p. 19-24.
6. Järv, O., R. Ahas, and F. Witlox, *Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records.* Transportation Research Part C: Emerging Technologies, 2014. **38**: p. 122-135.
7. Jiang, S., J. Ferreira, and M.C. González, *Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore.* IEEE Transactions on Big Data, 2017. **3**: p. 208-219.
8. Jiang, S., J. Ferreira, and M.C. González, *Clustering daily patterns of human activities in the city.* Data Mining and Knowledge Discovery, 2012. **25**(3): p. 478-510.
9. Jiang, S., et al., *A review of urban computing for mobile phone traces: current methods, challenges and opportunities*, in *Proceedings of the 2nd ACM*

SIGKDD International Workshop on Urban Computing. 2013, ACM: Chicago, Illinois. p. 1-9.

10. Robert, T., W. Guenther, and H. Trevor, *Estimating the number of clusters in a data set via the gap statistic.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2001. **63**(2): p. 411-423.

11. Schneider, C.M., et al., *Unravelling daily human mobility motifs.* Journal of The Royal Society Interface, 2013. **10**(84).

12. Tostes, A.I.J., et al., *From data to knowledge: city-wide traffic flows analysis and prediction using bing maps*, in *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*. 2013, ACM: Chicago, Illinois. p. 1-8.

13. Yang, P., et al. *Identifying Significant Places Using Multi-day Call Detail Records*. in *2014 IEEE 26th International Conference on Tools with Artificial Intelligence*. 2014.

14. Yin, M., et al., *A Generative Model of Urban Activities from Cellular Data.* IEEE Transactions on Intelligent Transportation Systems, 2018. **19**(6): p. 1682-1696.

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

16. Alhasoun, F., et al., *The City Browser: Utilizing Massive Call Data to Infer City Mobility Dynamics*. 2014.

17. Axhausen, K.W., *Activity Spaces, Biographies, Social Networks and their Welfare Gains and Externalities: Some Hypotheses and Empirical Results.* Mobilities, 2007. **2**(1): p. 15-36.

18. James, G., et al., *Statistical Learning*, in *Springer Texts in Statistics*. 2013, Springer New York. p. 15-57.

19. Kang, C., et al., *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, ACM: Chicago, Illinois. p. 1-8.

20. Widhalm, P., et al., *Discovering urban activity patterns in cell phone data.* Transportation, 2015. **42**(4): p. 597-623.

21. Zhang, D., et al., *Exploring human mobility with multi-source data at extremely large metropolitan scales*, in *Proceedings of the 20th annual international conference on Mobile computing and networking*. 2014, ACM: Maui, Hawaii, USA. p. 201-212.

22. Zheng, Y., et al., *Urban Computing: Concepts, Methodologies, and Applications.* ACM Trans. Intell. Syst. Technol., 2014. **5**(3): p. 1-55.

23. Zheng, Y., et al., *Understanding mobility based on GPS data*. 2008. 312-321.

APPENDIX

VITA

Naruethai Thongphasook was born on 13 January 1992 in Bangkok, Thailand. She received Degree in Bachelor of Engineering (Computer Engineering) from Chulalongkorn University. As part of her master’s studies, she focuses on data analytics and urban computing.