Clustering Analysis of Urban Fabric Detection Based on Mobile Traffic Data

Chuangfei Liu\(^1\ast\), Chang Liu\(^2\ast\), Fuqiang Liu\(^{1,3}\) and Jianyao Hu\(^4\)

\(^1\)College of Electronics and Information Engineering, Tongji University, Shanghai, 201800, China
\(^2\)Department of Urban Planning and Design, The University of Hong Kong, Hong Kong, China
\(^3\)College of Design and Innovation, Tongji University, Shanghai, 201800, China
\(^4\)China CEPREI Laboratory, Guangzhou, 510610, China

\ast Corresponding author’s e-mail: liuchuangfei@foxmail.com, Chloeliu@hku.hk

Abstract. The rapid development of city makes it complicated to analyse urban structure. It’s difficult to learn city ecosystem using traditional methods including interview and survey. A city generates lots of data every day, which could tell city’s dynamic fabric. The moving patterns of citizens could be illustrated through analysing traffic data on their mobile phones, since those who carry a mobile phone possess a large percentage of the urban population. Thus, we can get the designated urban area’s functions through the moving pattern analysis. In this paper, we explore the fabric of city using cluster analysis based on deep learning with mobile phone communication data. We get the inspiration from image processing and build communication snapshot map to represent each region. After extracting features using deep learning method, we use unsupervised learning to find similar regions of the city. The clustering analysis result is examined by the ground truth data.

1. Introduction

Modern city is a complicated system, which consists of residential buildings, commercial facilities, sports facilities and public green spaces. It is the result of the interaction between human social activities and urban physical spaces. With the development of the city and its increasingly intricate structure, various spatial functions change dynamically. Traditional urban structure exploration methods such as interviews and questionnaires are not only time-consuming and laborious, but also difficult to reflect the dynamic changes in urban structure.

The distribution of various functional areas in cities is closely related to the pattern of human activities. Studying the distribution and patterns of people activities is the basis for understanding all the problems and the structure of a city. With the development of mobile communication technology, smart phones have been widely used in urban area. In the process of using mobile phones, a large amount of CDR (Call Detail Records) data is generated. This kind of data is not only large in sample size, stable and reliable, but also closely related to the travel behaviour of citizens. By analysing the mobile communication data, we can understand the trajectory of their time and space activities. And then we can explore the urban structure. At present, relevant research has paid little attention to the value of mobile phone CDR data in urban planning. Studies have shown that citizens’ behaviour patterns are
closely related to the function of the city in some regions [1]. Therefore, we can obtain the functional
distribution of the city by studying the communication behaviour of residents [2].

In this paper, we focus on the analysis of urban structures using mobile traffic data. Through the
cluster analysis of people's communication behaviours, the functionally similar areas in the city are
found. Some researchers have tried to extract some communication data for clustering directly to study
the functions of different regions [3]. However, this type of processing will lose some information.

We use the Telecom Italia Big Data Challenge dataset for our research [4]. This data set divides
Milan into 10,000 squares and provides two months of communication data (including text messages,
calls, and network traffic). We take inspiration from image processing to process communication data
[5]. For each square, we build a snapshot using its traffic data. We use a week’s traffic data change graph
to represent the activity patterns in their region. Then we use the deep learning method to extract the
features of each flow snapshot map, and finally perform cluster analysis. In this way, areas where
communication traffic data changes similarly are divided into the same type. Finally, we use the ground
truth data to analyse and evaluate our clustering results, and compare it with the traditional urban
structure clustering analysis method.

Figure 1. The number of calls between two
squares in one week.

Figure 2. Internet mobile traffic heat map of
Milan in one week.

The following sections of the paper is structured as below. Section 2 introduces the mobile
communication dataset and ground truth dataset used in this paper. Section 3 introduces the construction
of flow snapshot map and the process of feature extraction using deep learning method. Section 4
introduces the process of clustering. Section 5 combines the ground truth data to analyse the results.
Section 6 is the summary.

2. Data Sources and Collection
The dataset is mainly from the Telecom Italia Big Data Challenge dataset. This data set is provided by
the 2014 Italian Big Data Challenge. After the competition, it was opened for researchers. The Call
Detail Records data we used are provided by the Semantics and Knowledge Innovation Lab (SKIL) of
Telecom Italia. In order to verify and analyse our clustering results, we also collected spatial planning
of Milan city [5].

2.1. Mobile phone data
When people use mobile phones to communicate, they interact with the communication base station. So,
the telecommunications providers can obtain a large amount of mobile communication data [6]. The
data used in this article is provided by Telecom Italia and contains telephone. We get SMS and mobile
traffic data for Milan city in Italy. And the data lasts from November to December 2013. Raw data was
processed for research convenience and privacy protection. First, Milan could be divided into 10,000
square blocks, each of which has a size of 235*235 meters. All data is described in terms of such blocks.
In each square block, we get the number of received and sent text messages, the number of made and
received calls and the number of internet connections. All data is recorded every 10 minutes. Figure 1
shows the number of calls in a week for two blocks. Figure 2 is a heat map of communication behaviour based on this dataset. People's behaviour patterns could be found from these data. For privacy reasons, this communication data is carried out in accordance with specific normalized way.

2.2. Ground truth data
The ground truth data is derived from the Milan City Government's public data set. The public dataset provides a distribution of facilities including residential areas, commercial areas, sports venues and public green spaces in Milan. In order to verify the clustering results, we refer to the method of dealing with the ground truth data to the same type with the mobile phone data. At last the ground truth data is also counted in blocks of 235*235 meters. The number of commercial facilities, sports fields, bus stops and universities in each block was finally obtained. In addition, we also obtained the ratio of population to green space in each region. These data help to verify the results of our cluster analysis. There are some limitations to this approach. First, only about 30% of the regions have these field data, which may decrease the representativeness of our study. Second, some facilities cover different areas, which will require repeated statistics on the number of facilities. Compared with other people's verification of the results through interviews, we still have obvious advantages in using field data for analysis.

3. Feature engineering
The CDR data used in this paper belongs to one type of time series data, which cannot be directly clustered, require feature extraction [7]. There are three main types of traditional methods of feature extraction. The first one is based on basic statistical methods for feature extraction [8]. Based on the extracted statistical features, the feature vectors are constructed. The features constructed can be explained strongly, but a large number of experiments are needed to select features. The second way is based on the model for feature extraction, which usually requires the time series to have a smooth autoregressive nature to achieve better results. The third is doing feature extraction by transforming. The most common method is the Fourier transform (FFT), which selects the appropriate features in the frequency domain [9]. The results of these feature extractions need to be manually selected and have strong uncertainty. The development of deep learning provides a new idea for time series feature extraction. This paper transforms time series data into flow snapshot matrix, and then uses auto encoder based on deep learning to extract features. The obtained feature vector can represent the original data better using this way in most cases.

Figure 3. A snapshot of the communication data of a certain area, the red indicates that the communication behaviour is high.  

Figure 4. The input and output of the auto encoder are basically the same, which proves that it has a better feature extraction effect.

3.1. Snapchat
We use $o_i(t_m)$ to represent the value of the communication data in the block. We use $i$ to represent the block number, and $m$ for the record point number. We select the traffic data for the week to generate
the traffic snapshot matrix. In each block, we can use the following matrix to represent its communication behaviour within one week.

\[
\begin{bmatrix}
\mathbf{o}_i(t_1) & \cdots & \mathbf{o}_i(t_{144}) \\
\vdots & \ddots & \vdots \\
\mathbf{o}_i(t_{1009}) & \cdots & \mathbf{o}_i(t_m)
\end{bmatrix}
\] (1)

Since the user's communication behaviour has periodicity in the granularity of day and week, we use data as a row of the matrix. Then we use the data of one week in one square form a matrix to represent the communication behaviour of this area this week. Using this way, we build a snapshot of the traffic based on the matrix, and visualize a snapshot of a portion of the area. It can be clearly seen that since the data is constructed in this way, and the usage of mobile traffic in different areas is significantly different. Figure 3 shows a snapshot of traffic of 20 regions. It can be seen that there are differences in communication behaviour between different regions.

3.2. Auto encoder

After each region is represented by a flow snapshot matrix, we cluster the traffic snapshots from the image clustering algorithm so that similar regions can be found. In this way, similar regions can be found. Here, we chose to use the auto encoder based on deep learning for feature extraction. The auto encoder is a neural network that uses a backpropagation algorithm to make the output value equal to the input value. It first compresses the input into a low dimensional space and then reconstructs the output by this characterization. The auto encoder consists of an encoder and a decoder. The encoder compresses the input into potential spatial features, which can be represented by the encoding function \( h = f(x) \). The decoder reconstructs the input from the potential space, represented by the decoding function \( r = g(x) \). This paper uses a convolutional auto encoder for feature extraction. Figure 5 shows the structure of the autoencoder we used.

![Figure 5. The structure of auto encoder](image)

We added a convolutional layer in front of the encoder to assist in feature extraction, which helps the neural network learn the flow relationship between different dates. As can be seen from the figure 4, the data at the input and output have very similar trends, which shows that our self-encoder has achieved good results. So, we can get better feature extraction in the hidden layer of the neural network.

4. Clustering

After feature extraction using a convolutional auto encoder, we cluster the feature vectors of each block. We assume that regions with similar communication behaviours belong to similar functional regions in a city, and some studies have confirmed this hypothesis [10]. Therefore, through the cluster analysis of people's communication behaviour, we can understand the functionally similar areas in the city. We chose to use k-means which is an unsupervised machine learning algorithm to do clustering. This method can be separated into feature extraction, distance calculation and clustering [11]. The convolutional autoencoder has extracted the feature vector from the high dimensional matrix. Pearson correlation coefficient is used as the formula for the similarity calculation. For the two box areas, their distance is defined as follows:

\[
dist(A_i, A_j) = 1 - \text{pearson}(A_i, A_j)
\] (2)
5. Results and analysis

We analysed the clustering results through combining the previously obtained field data. It is worth noting that since the ground truth data contains only about 30% of the blocks, all of our analysis results are limited.

In the process of cluster analysis, we divide all the squares blocks into 10 categories. Since cluster analysis belongs to one type of unsupervised learning, the results of clustering are completely learned by the machine. Other aspects of data need to be verified to explain the results. In this paper, we plot the average change in mobile traffic for each category over the course of a week. It can be seen in figure 8 that communication behaviours occur more evenly in some areas, which may belong to areas where people are comparatively stable. In other areas, communication behaviours change significantly within a week and may belong to the work area.

Figure 6. The clustering results of 10,000 squares.

Figure 7. The number of squares included in each category.

Figure 8. Average behaviour patterns of different categories.

When the communication behaviours of people in the two regions are completely similar, their distance is 0; when the behaviour patterns of people in the two regions have nothing to do with each other, their distance is 1; when the behaviour patterns of people in the two regions are completely absent, their distance is 2. By defining such distances, we can categorize areas that are closer together into the same category and find urban areas with similar functions. We make \( k = 10 \), which can divide the urban area into 10 categories to get our clustering results. As figure 6 shows, the areas with the same colour in the figure represent similar urban functional areas. And figure 7 shows the number of squares in each category. Overall, the geographically similar areas have similar functions, which is consistent with our consistent perception.

Figure 9. The number of universities in different categories.

Figure 10. The number of commercial facilities in different categories.
Combined with the ground truth data, we calculated the number of different functional regions in the clustering results to evaluate our clustering results. In general, squares of the same functional area are classified as much as possible. Due to the complexity of the structure of the city, the distribution of various functional areas will be intertwined [12]. We have to tolerate a certain classification error rate. Here we have selected several typical urban functional areas and observed their distribution in the clustering results. We find that most of the universities belong to cluster 0, and some universities are classified as cluster 1, which is shown in Figure 9. Most of the commercial area is divided into cluster 1, and some are divided into cluster 0, which is shown in Figure 10.

6. Conclusion
In this paper, we analyse the urban structure based on mobile phone data. With the continuous development of big data today, we could improve the urban development methods and urban planning studies through big data analysis and research methods [13]. We also build a snapshot of the mobile traffic behaviour of people in the city, which can represent people's behaviour patterns. We refer to the image clustering method to extract the features from the flow snapshot map based on the convolution auto encoder. Then use the clustering algorithm to explore the distribution of urban functional areas. This unsupervised learning method can divide urban areas into several types according to people's communication behaviours. After that, we discuss and analyse the clustering results based on the ground truth data, which turns out to be positive results. In the future work, the feature extraction and cluster analysis could be integrated in a unified framework to achieve end-to-end cluster analysis. In addition, we can collect a variety of ground truth data to verify and analyse our clustering results.

Acknowledgments
This work was funded in part by the National Natural Science Foundation of China under Grant 61331009 and Intelligent Connected Vehicle Pilot Demonstration Project under Grant 2019B090912002. This is also a part of project that has received funding from the University of Hong Kong & Tongji University under the fund of Ministry of Education, China – Research and Communication of Cultural and Creative industries and Regional Development under the background of smart city.

References
[1] Naboulsi D, Fiore M, Ribot S, et al. Large-scale mobile traffic analysis: a survey[J]. IEEE Communications Surveys & Tutorials, 2015, 18(1): 124-161.
[2] Furno A, Stanica R, Fiore M. A comparative evaluation of urban fabric detection techniques based on mobile traffic data[C]/Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015. Paris. pp. 689-696.
[3] Cici B, Gjoka M, Markopoulou A, et al. On the decomposition of cell phone activity patterns and their connection with urban ecology[C]/Proceedings of the 16th ACM International Symposium on Mobile Ad Hoc Networking and Computing. ACM, 2015: 317-326.
[4] Barlacchi G, De Nadai M, Larcher R, et al. A multi-source dataset of urban life in the city of Milan and the Province of Trentino[J]. Scientific data, 2015, 2: 150055.
[5] Zhang C, Ouyang X, Patras P. ZipNet-GAN: Inferring fine-grained mobile traffic patterns via a generative adversarial neural network[C]/Proceedings of the 13th International Conference on emerging Networking Experiments and Technologies. ACM, 2017: 363-375.
[6] Bajardi P, Delfino M, Panisson A, et al. Unveiling patterns of international communities in a global city using mobile phone data[J]. EPJ Data Science, 2015, 4(1): 3.
[7] Bianchi F M, Rizzi A, Sadeghian A, et al. Identifying user habits through data mining on call data records[J]. Engineering Applications of Artificial Intelligence, 2016, 54: 49-61.
[8] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis[C]/International conference on machine learning. 2016: 478-487.
[9] Ghahramani M, Zhou M C, Hon C T. Spatio-temporal analysis of mobile phone data for interaction recognition[C]//2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC). IEEE, 2018: 1-6.

[10] Madiraju N S, Sadat S M, Fisher D, et al. Deep Temporal Clustering: Fully unsupervised learning of time-domain features[J]. arXiv preprint arXiv:1802.01059, 2018.

[11] Marom Y, Feldman D. k-Means Clustering of Lines for Big Data[J]. arXiv preprint arXiv:1903.06904, 2019.

[12] Wu C, Ye X, Ren F, et al. Check-in behaviour and spatio-temporal vibrancy: An exploratory analysis in Shenzhen, China[J]. Cities, 2018, 77: 104-116.

[13] Tan J, Dong L, Gao J, et al. The Methods of Extracting Spatiotemporal Characteristics of Travel Based on Mobile Phone data[C]//2018 IEEE 7th Data Driven Control and Learning Systems Conference (DDCLS). IEEE, 2018: 1174-1179.