Refining ground classification for the distribution of LTE users using supervised learning techniques

Kodjo E. F. Tossou
UTBM, CNRS, Institut FEMTO-ST
F-90000 Belfort, France
LARSI, EPL, Université de Lomé
01 BP 1515 Lomé 01
Lomé, Togo
ORCID: 0000-0002-7280-7562

Sid Lamrous
UTBM, CNRS, Institut FEMTO-ST
F-90000 Belfort, France
ORCID: 0000-0003-1735-2603

Laurent Moalic
Université de Haute-Alsace, IRIMAS UR 7499
F-68100 Mulhouse, France
ORCID: 0000-0003-3749-3227

Tchamye Boroze
Solar Energy Laboratory (LES) 01 BP 1515 Lomé 01
Lomé, Togo
ORCID: 0000-0003-1009-7179

Oumaya Baala
UTBM, CNRS, Institut FEMTO-ST
F-90000 Belfort, France
ORCID: 0000-0002-7247-7874

***Abstract*—Several studies have shown that the layout of an area's infrastructure has a strong impact on the mobility of the mobile network users. Each district in a city has one or more different type of activity areas. Depending on the type of activity areas that a district covers, several profiles emerge. These profiles are closely linked to the impact a district can have on LTE users mobility. In the current work, we propose a first approach to determine the profile of a district in a territory. The territory under study is the city of Lomé, the data used for this analysis come from the geographical data of the OSM database. An alternative approach is proposed that, in the case of missing data, determines the profile of a new district from knowledge built from other districts in the same study area. To validate the proposed approach, evaluations were conducted considering several types of distance (Mahalanobis distance, Euclidian distance, ...). It appears that with the K-NN algorithm, using manhattan distance, we have 61% accuracy in determining the profile of a new district based on the nearest district's profiles.**

***Index Terms*—LTE networks, LTE users, Urban mobility, Supervised learning, Profiling, Urban fabric characterization, Urban fabric classification, k-NN.**

I. Introduction

During decades, there are several studies on distribution of the population of a given territory [1]–[9], whether urban or not. The majority of this population owns a mobile phone and subscribes to mobile network services. These studies provides tools and methods to understand the movement of subscribers and what influences their movement. This makes it possible to know at a given time, what the distribution of the subscribers in an area will be. Many techniques use spatial data [1]–[3], other are based on mobile traffic and census data [2], [4] even if sometimes mobile traffic data can be unavailable [5], it is shown that it gives a real-time estimate of LTE users distribution [4].

Some studies have characterized mobility behavior by determining different mobility patterns. In [6] they built on previous studies that show a relationship between mobile traffic data and urban fabric to look for patterns that emerge in mobile communication activities in relation to different urban fabric. In [7], they used CDR (*Call Detail Record*) to identify weekly patterns of human mobility.

All these studies reveals that the movement or distribution of subscribers in a territory, composed of districts, is strongly correlated with the infrastructure of that area. The different type of activity areas that are covered by a district have an impact on the distribution of subscribers in that district at a given time. This behaviour is not the same for another district that covers other types of activity areas. Therefore, each district has a profile linked to the different types of activity area it covers.

In this research work, we propose a method for determining district's profiles based on the different types of activity areas they cover and use K-NN algorithm to determine the profile of a new district based on the nearest districts. Our main contributions are :

- We design a way to determine the profile of a district;
- We propose a more granular classification of urban fabric;
- We evaluate the determination of the profile of new district knowing the profiles of the nearest districts.

The rest of this paper is organized as follows. Section II provides an overview of the related works on the classification problem. Section III provides the datasets description and the methodology approach whereas Section IV describes the experimentation setup and results and section V concludes the paper and provides future extensions of research.

II. Related Works

Determining the spatio-temporal distribution of a city's inhabitants is an important step in the process of planning and optimizing the services provided. Since most of the population owns is a mobile networks services subscriber, many techniques make it possible to know, at a given time, the distribution of the LTE users in an area. However, it is difficult to know precisely and periodically the distribution of the population based on census data because the distribution of the population changes rapidly according to Rex *et al.* [4]. So it is important to look for alternatives. Thus, several source and type of data have been used in many studies to determine as precisely as possible how the population is distributed and moves. Among these type, we have mobile traffic data [6]–[9], urban fabric [9], Taxicab GPS [8] and spatial data [1]–[3]. According to Yohan *et al.* most studies do not take into account the temporal behavior because this parameter implies a lot of uncertainty [10]. So in their work, they explore the spatio-temporal character of human mobility. After evaluating several models, their results show that location-dependent predictors are more accurate in predicting the temporal behavior of an individual than location-independent predictors. They also showed that the time an individual spends at a location is strongly correlated with his or her arrival time at that location and tendency to return to the next location.

As a result, several studies have focused on understanding human mobility from a spatial displacement perspective, such as Yuan *et al* [3] who worked on geotagged tweet datasets coming from Sweden and spanning 3.6 years on average. It was questioned whether geolocated data from social media can determine the real mobility of a city population. They argue the limits of that type of data by its low and irregular sampling frequency.

Sihan *et al.* proposed a clustering method for classification based on geolocation data from mobile application to predict population movement [2]. Anastasios *et al.* used geographic data provided by the geographic social network platform Foursquare to determine the urban mobility patterns of people in several metropolitan cities. They showed with their analyses that location distributions drive variation in human movement in different locations [1].

Other studies used mobile data. This is the case for instance of Rex *et al.* who developed an explicit link between telecommunication data and population distribution. They used telecommunication and census data to estimate the real-time distribution of the population in time and space.

And Khodabandelou *et al.* who propose an approach to statically and dynamically estimate the population of a region based on telecommunication metadata. They showed that subscriber presence data is a better way to detect the population distribution than other existing metrics [11].

Some studies have characterized mobility behavior by determining different mobility patterns. Furno *et al.* [6] for instance built on previous studies that show a relationship between mobile traffic data and urban fabric to look for patterns that emerge in mobile communication activities in relation to different urban fabric. they proposed a new method of mobile signature classification *The Median Week Signature* to create mobile demand profiles that are more in tune with the ground type. The evaluation of their technique was done on mobile data from mobile operators in 10 cities.

Ethienne *et al.* [7] used CDR (*Call Detail Record*) to identify weekly patterns of human mobility. Evaluation of their model with empirical data from operators shows its relevance. They used an event-based algorithm to cluster individuals and identify 12 weekly patterns. Their methodology is based on the classification of individuals into six distinct presence profiles by focusing on the geographic and temporal parameters inherent to each profile on a territory.

Another factor that is often not taken into account in determining the spatio-temporal behavior of an individual is the building. Most mobility models are based on a 2-dimensional landmark. According to Zimu *et al.* [9] these models are used in network planning and optimization [12] but they are limited. Because the movement of the population is done in a 3D space because of the buildings.

From these studies, it emerges that the distribution of the population of an area depends on the distribution of the infrastructure of that area. Each infrastructure or territory, composed of districts, therefore has a profile that emerges according to the types of activity areas it covers. This profile is then related to the LTE subscribers' distribution.

Thereby, in this paper, we propose a method for determining district's profiles based on the different types of activity areas they cover and use K-NN algorithm to determine the profile of a new district based on the nearest districts. To the best of our knowledge we are the first to use these method to determine the profile of new district.

III. Data set and Methodology

This section presents the description of the datasets and the methodology used. The first subsection presents the study area, the data collected, their structure, the gathering methodology and the representation on a map. The second subsection presents the urban classification methodology.

A. Study area

Our study focuses on the city of Lomé, capital of Togo. It has a total area of 90 km^2. It contains **69** districts, grouped into **5** boroughs and former district. The population estimation is **839,566** in 2010 as specified in the census data obtained from INSEED, the National Institute of Statistics in Togo. The distribution of the number of districts by borough are represented in table I

TABLE I
District distribution

Borough	Number of district
1	11
2	18
3	17
4	4
5	19

B. Building related Data

The classification of a given area considers the activities taking place there. Thus, it is necessary to know the different activities and their position in our study area.

Most areas of the Lomé City are covered with school buildings, public buildings, residences and much more. Thus, knowledge of the position and function of the different buildings becomes important in determining the types of activities carried out in and around the area covered by the building. As we were unable to obtain this data from the competent authorities, we retrieved this data from OpenStreetMap (OSM).

The QGIS software in its version 3.18 and the QuickOSM plugin helped to get this data. QuickOSM uses the Overpass API to download data by specifying the tags as defined in osm features [13] that can then be saved in shapefile format.

C. Determination of urban fabric classes

After obtaining this data, several steps needs to be performed before carry out the experimentation. In order to classify the different areas, it is necessary to classify the different activities. Therefore, several urban fabric classes have been defined. Urban fabric refers to the nature and function of the area present on the ground. Each urban fabric class categorizes certain types of activities. We thus define **11 classes**:

1) *office*: refers to all public and private buildings, as well as all areas used for office work
2) *residence*: refers to all buildings and living areas such as apartments, houses, hotels, ...
3) *trade*: refers to all buildings and areas where products are sold such as supermarkets, markets, etc.
4) *education*: refers to all school buildings such as high schools, universities, etc. It also refers to all training centers regardless of the type of training (dance, cooking, ...)
5) *beach*: refers to the entire area covered by the beach
6) *restaurant*: refers to all the places of restoration such as the canteen, the restaurants, the fast-food, ...
7) *park*: refers to all green spaces open to the public and recreational parks
8) *water*: refers to all water points in our study area
9) *vegetation*: refers to forests and other vegetation that is not public open space
10) *leisure*: refers to bars, nightclubs, and other entertainment venues outside the beach and parks
11) *Other*: refers to all other area that cannot be categorized in previous classes.

D. Profile determination

A district is a delineated area with several buildings and area types. Depending on the proportion of each type of area and type of building present in a district, the latter will have a given profile. The profile of a district is then a function of the proportion of the different types of business areas that are present. Some districts will have all activity areas and others will have only some. Also, other districts will have the same proportion of the different types of activity areas they cover, while others will have more variation in the proportion of activity area types.

In order to derive the different districts profiles, these two parameters are therefore taken into account:

- The proportion of area types
- The types of area covered

Based on these two factors, for two districts to have the same profile, they must cover the same areas of activity and have a proportion of these areas of activity that is close. In other words, two districts are close in terms of the proportion of types of activity areas if the proportions vary by no more than 15% from one district to the other and the ranking of the proportions remains unchanged.

Indeed, if two districts cover office and trade areas where the proportion in the first district is 80% for office and 20% for trade, and in the second district is 80% for trade and 20% for office, then these two districts cannot have the same profile. However, if the second district is 70% office and 30% trade, the proportions for both districts are close, then they will have the same profile.

Once the urban fabric classes are defined, the city of Lomé is divided into meshes of 100m x 100m that are classified according to the defined urban fabric classes. The classification is performed using OSM buildings and nodes data. A list of keywords per tag is defined to categorize OSM data according to the different urban fabric classes. A python program is written to browse OSM data and classify them according to the defined keywords. Then, the meshes are classified according to these classified data.

E. Profile determination using k-NN

Sometimes, data relative to different activity areas of a district may be missing, while for other districts they are available. To manage the missing data, a machine learning approach is proposed to predict the profile of a new district based on the knowledge built from the other districts.

In Lomé, there are few districts with a single profile (*residence, office, etc.*). Most districts have a mix of *offices (SME/SMI), trade (shop, stalls, etc.), etc.* in varying proportions. By considering the profile that emerges from each district and its geographical position, we can draw a line that surrounds them and form an area that circumscribes them. Thus, for a district with no information on its composition, we propose to look at the profiles of the nearest districts to determine its profile and thus determine the activity areas that are potentially located in this district and their proportion. This will help in the planning and deployment of new services. On the other hand, the size of our dataset (*less than 100 000 samples*) and its structure (*labeled data and not text data*) led us to use k-NN as a suitable classification algorithm to the problem. Using k-NN, we propose to predict the profile of a new district after training the model with data from other districts.

IV. Experiments and Results

This section presents the experimentation setup and the obtained results. The subsection IV-A presents the data pre-processing, the subsection IV-B presents the urban fabric classification, and the subsection IV-C describes the profile determination.

A. Data pre-processing

Before determining the district profile, the data to be used must be prepared. The study area depicted in Figure 1 had to be delimited, and the district cut out.

Figure 2 shows the geographical coverage of the various districts of the city of Lomé. Each polygon represents a district. For the area of interest, the data is downloaded from OSM database which allows obtaining two types of data relevant for our study that are polygons and geographic points called nodes. The Python language in its version 3 is used for coding and QGIS 3.18 is used for the visualizations.

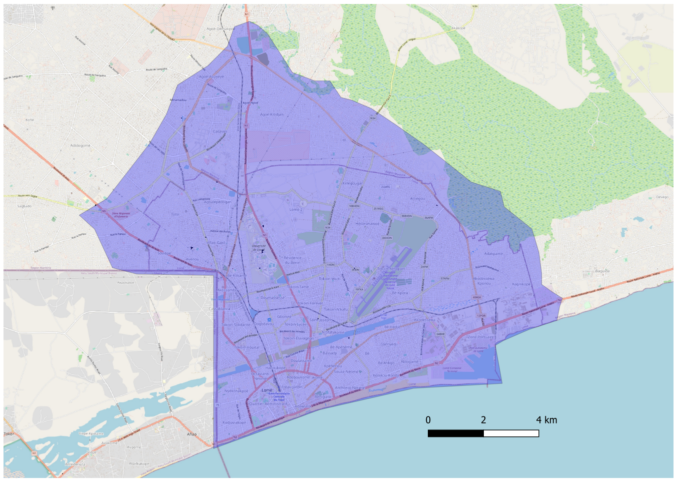

Fig. 1. City of Lomé

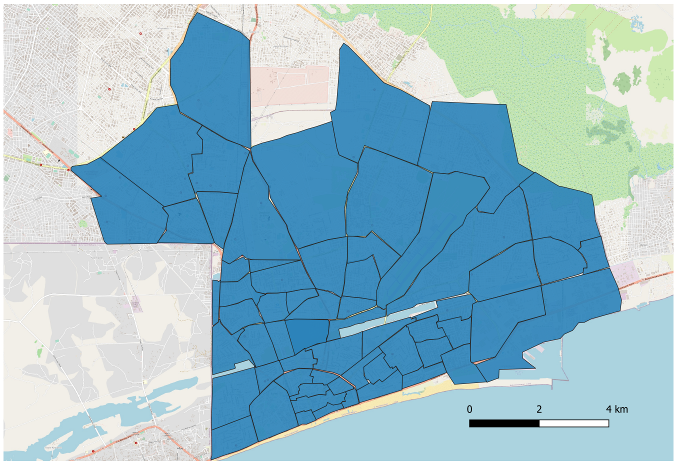

Fig. 2. Districts of Lomé

B. Urban fabric classification

Once data is ready, first, the study area is divided into 13,811 meshes of 100m x 100m that will be classified according to the 11 urban fabric classes defined. Color codes are assigned to each urban fabric class. Figure 3 provides a representation of the city of Lomé once the meshes have been classified. The meshes are grouped by district.

Unfortunately, in our case, not all information is available to identify the nature and function of all the urban fabric of the city as shown in Figure 3. Some meshes could not be classified because the area they cover do not have enough data to enable their classification. For a given district, these meshes are randomly classified according to the proportion of the different urban fabric classes present in this district.

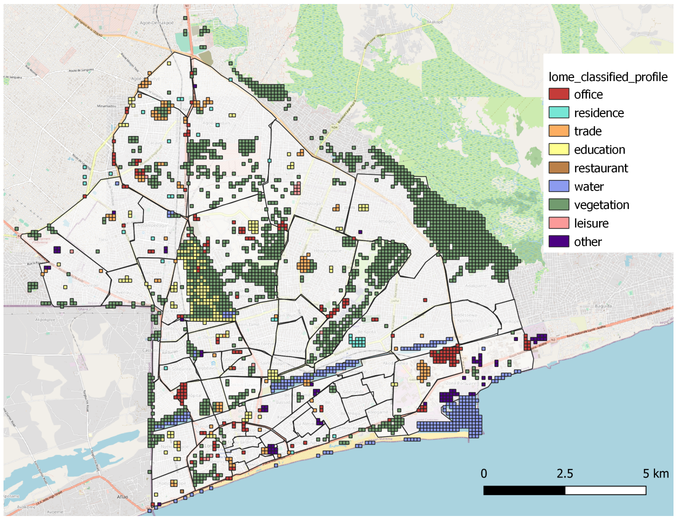

Fig. 3. classified meshes grouped by district

C. Profile determination

The different types of activity areas that make up a district have an impact on district profile. In our study, some districts are uniform, composed of one activity area type, while some others districts are composite and include several activity area types. For convenience, a code is defined for each urban fabric class, as shown in the table II.

TABLE II
Corresponding codes to classes

N°	Class	Class code
1	Office	Of
2	Residence	Rd
3	Trade	T
4	Education	E
5	Beach	B
6	Restaurant	Rt
7	Park	P
8	Water	W
9	Vegetation	V
10	Leisure	L
11	Other	Ot

Once the proportions of each activity type area were determined for a given district, the districts are grouped into two profile categories: the single profile type (uniform district) and the heterogeneous profile type (composite district). In the first category, the profile of a district containing 100% of a given activity type area is simply the urban fabric class name. In the area under study, 13 districts have a uniform profile type. All districts having 100% of the same urban fabric class are assigned a profile equal to that type of class. The table III shows these districts.

TABLE III
UNIFORM DISTRICTS

District	Class Proportion				Profile
	1	2	3	4	
GBONVIE	100.0	0.0	0.0	0.0	Of
KOTOKOU-KONDJI	0.0	100.0	0.0	0.0	Rd
TOKOIN-SOLIDARITE	0.0	100.0	0.0	0.0	Rd
BENIGLATO	0.0	0.0	100.0	0.0	T
ADOBOUKOME	0.0	0.0	100.0	0.0	T
DOGBEAVOU	0.0	0.0	100.0	0.0	T
HANOUKOPE	0.0	0.0	0.0	100.0	E
SAINT-JOSEPH	0.0	0.0	0.0	100.0	E

Not all the data are represented in table III. Districts TOKOIN-OUEST, TOKOIN-WUITI, TOKOIN-ELAVAGNON, ADAKPAME and AKOSSOMBO are classified as vegetation profile in uniform type profile. With the terrain knowledge, we know with certainty that these districts include not only vegetation but other urban fabric class such as education or residence. But due to the lack of data to identify the urban fabric function, only vegetation data were usable.

The second category is districts that contain more than one type of activity area. This is a composite profile type. Table IV summarizes the profile of composite districts. The codes of each class represented in column *Classes and percentage* are indicated in descending order of the percentage of the class. More precisely, a district with 75% office, 18% residential and 7% vegetation would be labeled OfRdV_721. Once the classes and percentage determined for a district, the final urban fabric class is derived considering the class with the highest percentage. With the previous example, the final class value for the district with *OfRdV_721* classes and percentage label becomes *Of*. When a district have the same proportion of the activity areas covered, the more relevant urban fabric class is chosen for the final class value. The profiles of all composite districts are provided in table IV.

TABLE IV
COMPOSITE PROFILE TYPE DISTRICTS

District	Classes and percentage	Final Class
KODJOVIAKOPE	TERt_532	T
NYEKONAKPOE	EVT_631	E
AMOUTIVE	OtOfE_622	Ot
BE	VOf_64	V
BE-APEHEME	VOt_73	V
ANTONIO-NETIME	OfV_55	Of
KOKETIME	TOtOf_631	T
TOKOIN-HOPITAL	VOf_73	V
TOKOIN-GBADAGO	VT_73	V
TOKOIN-LYCEE	VLOf_622	V
DOUMASSESSE	EVOfT_4321	E
ABOVE	VW_73	V
NTIFAFAKOME	VTE_622	V
TOKOIN-TAME	RdTV_811	Rd
BE-KPOTA	RdOfV_622	Rd
AKODESSEWA KPOTA	WT_91	W
LOME-2	VEOf_811	V
BE-KLIKAME	VEWOf_5311	V
HEDZRANAWOE	VTEOf_5311	V
KELEGOUGAN	VTELOf_52111	V
ATTIEGOU	VEOt_811	V
CACAVELI	VTOfRdEL_52111	V
AGOE-ASSIYEYE	TRdVE_4222	T
AGOE-KITIDJAN	VT_91	V
TOTSI	TVE_422	T
AVENOU	VE_55	E
ZONE PORTUAIRE	OtVOf_622	Ot
TOKOIN AEROPORT	VOf_91	V
SOVIEPE	VOtTE_5311	V
AFLAO-GAKLI	VTRd_631	V
AGBALEPEDOGAN	VOfOt_222	Of
OCTAVIANO-NETIME	OfT_73	Of
WETRIVI-KONDI	VOf_82	V
AGBADAHONOU	OfT_55	Of
AKODESSEWA	OfTV_631	Of
ABLOGAME	VOf_91	V

D. *Profile determination using k-NN*

The k-NN algorithm is well known algorithm for its efficiency in the field of classification [14], which we decided to use to predict the profiles of the districts. Furthermore, several distances are used as metrics to evaluate the quality of the profile prediction. Again, we decided to use very common and widely used distances, also called metrics that are: *Mahalanobis, Manhattan, Euclidian, Cosine, Correlation, Canberra, Braycurtis, Infinity, Chebyshev, Minkowski and Cityblock*. Each distance is used to predict the profiles of districts. The data set is divided into 80% for training and 20% for testing. Then, the predicted profiles for each distance are compared with the testing data to determine the accuracy of the prediction. Figure 4 shows the accuracy of each distance which are compared to identify the best distance metric as well as the number of k-neighbors providing the best prediction accuracy.

From these results, we observe that the best distance metric is **Manhattan** and the good number of k-neighbors is **5**. With this metric, we can derive the profile of a district that cannot be identified with data collected from OSM due to missing data. Finally, 14 districts among the 64 districts in the study area cannot be identified.

V. CONCLUSION

For decades, several studies have been conducted to characterize and extract area profiles from predefined urban fabric classes. In this research work, first, we defined a set of urban fabric classes. The territory, composed of districts, is meshed and classified according to the different urban fabric classes. Then, we proposed an approach to determine profiles for specific areas based on the defined urban fabric classes. Data

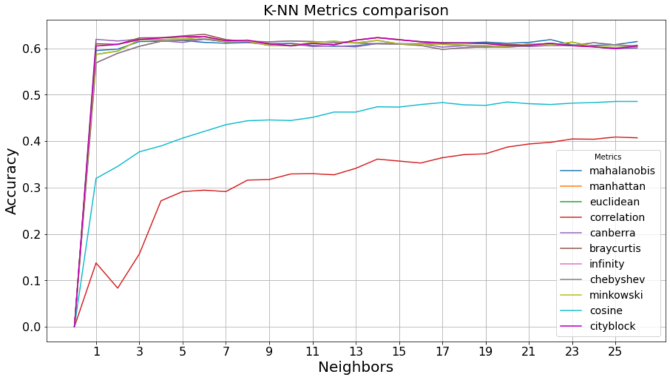

Fig. 4. K-NN Metrics comparison

set comes from OSM database including information related to polygons and geographic points. The major limitation of our study is the lack of OSM data on the nature and function of several urban fabric area in the study area. In order to compensate this, random data were generated in each district respecting the proportion of each activity area. For districts with missing OSM data, the prediction of profiles is performed using K-NN algorithm. To assess the quality of the prediction, we used metrics that are distances. We compared the prediction provided by each distance to identify the best metric and the good number of k-neighbors. The Manhattan distance outperforms the other distance metric with 61% accuracy using 5 neighbors.

For further work, we envisage to gather more real data coming either from enriched OSM database of the study area or another area or from other source of ground-truth data which will help the assessment of our approach. Moreover, we can test other classification algorithms and compare their accuracy to identify the most appropriate one in our context.

References

[1] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027–, 2012.

[2] Sihan Zeng, Huandong Wang, Yong Li, and Depeng Jin. Predictability and prediction of human mobility based on application-collected location data. In *2017 IEEE 14th International Conference on Mobile Ad Hoc and Sensor Systems (MASS)*, pages 28–36, 2017.

[3] Yuan Liao and Sonia Yeh. Predictability in human mobility based on geographical-boundary-free and long-time social media data. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2068–2073. IEEE, 2018.

[4] Megha Ram David Rideout Rex W. Douglass, David A. Meyer and Dongjin Song. High resolution population estimates from telecommunications data. *EPJ data science*, 4(1):1–13, 2015.

[5] Yingxiang Yang, Carlos Herrera, Nathan Eagle, and Marta C González. Limits of predictability in commuting flows in the absence of data for calibration. *Scientific reports*, 4(1):5662–, 2014.

[6] Angelo Furno, Marco Fiore, Razvan Stanica, Cezary Ziemlicki, and Zbigniew Smoreda. A tale of ten cities: Characterizing signatures of mobile traffic in urban areas. *IEEE transactions on mobile computing*, 16(10):2682–2696, 2017.

[7] Etienne Thuillier, Laurent Moalic, Sid Lamrous, and Alexandre Caminada. Clustering weekly patterns of human mobility through mobile phone data. *IEEE transactions on mobile computing*, 17(4):817–830, 2018.

[8] Chaogui Kang, Stanislav Sobolevsky, Yu Liu, and Carlo 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*, UrbComp '13, pages 1–8. ACM, 2013.

[9] Zimu Zheng, Feng Wang, Dan Wang, and Liang Zhang. An urban mobility model with buildings involved: Bridging theory to practice. *ACM transactions on sensor networks*, 16(1):1–24, 2020.

[10] Yohan Chon, Hyojeong Shin, Elmurod Talipov, and Hojung Cha. Evaluating mobility models for temporal prediction with high-granularity mobility data. In *2012 IEEE International Conference on Pervasive Computing and Communications*, pages 206–212, 2012.

[11] Ghazaleh Khodabandelou, Vincent Gauthier, Mounim El-Yacoubi, and Marco Fiore. Population estimation from mobile network traffic metadata. In *2016 IEEE 17th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–9, 2016.

[12] Yao Wei and Mugen Peng. A mobility load balancing optimization method for hybrid architecture in self-organizing network. In *IET International Conference on Communication Technology and Application (ICCTA 2011)*, pages 828–832, 2011.

[13] M Haklay and P Weber. Openstreetmap: User-generated street maps. *IEEE pervasive computing*, 7(4):12–18, 2008.

[14] Masoodzadeh F. Roshandel E. Hosseinzadeh, J. Fault detection and classification in smart grids using augmented k-nn algorithm. *SN Applied Sciences*, 1(8):1627, 2019.