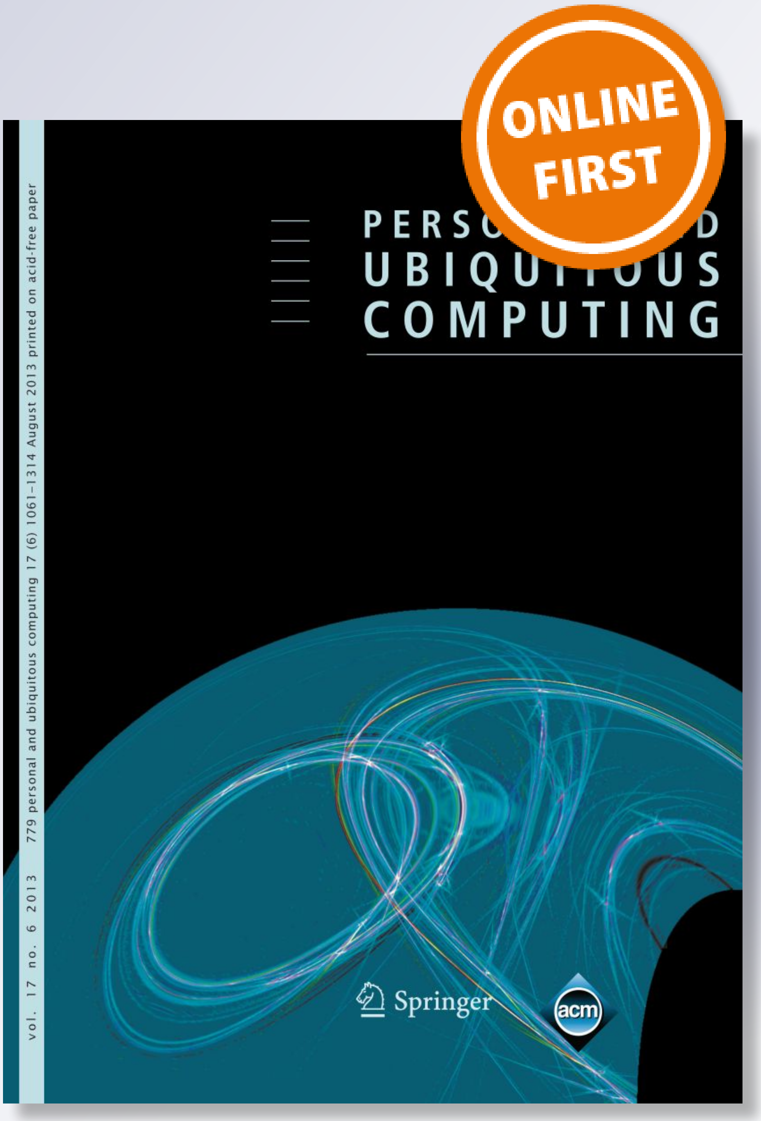
Utilizing digital traces of mobile phones for understanding social dynamics in urban areas

Oded Zinman & Boaz Lerner

Personal and Ubiquitous Computing

ISSN 1617-4909

Pers Ubiquit Comput
DOI 10.1007/s00779-019-01318-w


ONLINE FIRST
PERSONAL AND
UBIQUITOUS
COMPUTING
vol. 17 no. 6 2013 779 personal and ubiquitous computing 17 (6) 1061–1314 August 2013 printed on acid-free paper
Springer
acm

Springer

Personal and Ubiquitous Computing
https://doi.org/10.1007/s00779-019-01318-w

ORIGINAL ARTICLE

Utilizing digital traces of mobile phones for understanding social dynamics in urban areas

Oded Zinman[1] · Boaz Lerner[1]

Received: 1 December 2018 / Accepted: 5 September 2019

Abstract
Understanding land use in urban areas, from the perspective of social function, is beneficial for a variety of fields, including urban and highway planning, advertising, and business. However, big cities with complex social dynamics and rapid development complicate the task of understanding these social functions. In this paper, we analyze and interpret human social function in urban areas as reflected in cellular communication usage patterns. We base our analysis on digital traces left by mobile phone users, and from this raw data, we derive a varied collection of features that illuminate the social behavior of each land use. We divide space and time into basic spatiotemporal units and classify them according to their land use. We categorize land uses with a leveled hierarchy of semantic categories that include different levels of detail resolution. We apply the above methodology to a dataset consisting of 62 days of cellular data recorded in nine cities in the Tel Aviv district. The methodology proved beneficial with an accuracy rate ranging from 84 to 91%, dependent on land use label resolution. In addition, analyzing the results sheds light on some of the limitations of relying solely on cellular communication as a data resource. We discuss some of these problems and offer applicable solutions.

Keywords Land use · Computational social science · Mobile phone data · Urban computing · Classification · Smart cities

1 Introduction

The emergence of data generated by mobile phones has enabled a wide range of academic, social, commercial, and governmental applications [29]. The distribution of mobile use in the modern world is immense, especially in highly populated urban areas. According to a worldwide Pew report conducted in 2016, 88% of the respondents indicated that they owned a cellphone [24]. As mobile phones are usually kept close to the user and contain many useful sensors, they enable geospatial detection; thus, the digital traces these phones leave are effective for mobility-pattern discovery and next-place prediction [17], capturing behavior in everyday life [8, 22], and for identifying and predicting social lifestyles [4, 5]. Data generated by mobile phones is highly efficient for analyzing human and social behaviors from a small-scale individual perspective to large-scale collective behavior with an unprecedented degree of reach and accuracy [9].

Rapid urbanization and the evolution of modern cities have created new challenges for social research in cities and urban planning. Interactions between topography, transportation infrastructure, individual mobility patterns, real estate markets, and social preferences cause problems such as traffic congestion, air pollution, and urban sprawl [1]. Years ago, the vast number of factors influencing the city ecosystem made it almost impossible to tackle these challenges, but recently new insights toward smarter cities have become possible through sensing and computing technologies [39]. Location estimation systems and other sensing abilities inherent in mobile devices are major contributors to better understanding the spatiotemporal properties of collective urban mobility patterns.

Mobile phones are efficient for function identification of urban zones, as they provide new insights into social interactions and activity. In the modern city, different parts of the city function for different social purposes, i.e., residential neighborhoods, commercial areas, industrial areas, etc. Understanding the social functions of urban land use

✉ Boaz Lerner
boaz@bgu.ac.il

Oded Zinman
odedzinman@gmail.com

[1] Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Beer Sheva, Israel

contributes to urban planning and design of better urban strategies, e.g., planning of highways [15]. However, high-resolution and up-to-date mapping of these areas according to their social function is rare, especially in developing countries [15, 28].

In this work, we aim to (1) illuminate different aspects of the social functions as reflected by cellular communication and (2) discuss and offer solutions for some of the obstacles that may stand in the way of a sufficient land use identification based only on cellular communication. We examine and interpret the cellular communication patterns of different land uses, basing our analysis on call detail records (CDRs), which are mobile phone signals routinely collected and stored by telecom operators. To better understand the dynamics of the different social functions, we present a vast collection of features that portray different perspectives of the cellular communication characteristics of the social functions. Thus, we recognize daily and weekly patterns typical to each social function and emphasize the difference between the land uses. We use a supervised land use identification methodology, similar to that used in previous works [23, 30]. We divide urban time and space into small spatiotemporal units and use the random forest classifier to discover areas of similar social function, tagging them with a semantically meaningful label. Resolution of the highest land use categories includes residential, industrial, office, commercial, entertainment, highways, streets, and no activity. We use an hourly land use labeling set that enables more flexibility, recognizing the variations between land uses that cannot be taken into consideration when using a daily label. We implement this methodology on a dataset acquired from a leading Israeli telecommunications company collected during 62 days of communication in different cities in the center district of Israel. By analyzing the results of this use case, we discuss the opportunities of this method—its good accuracy rate with a relatively inexpensive dataset—alongside some of the limitations of such an approach. We discuss the consequences of the location estimation inaccuracy of a CDR-based dataset, and we analyze and elaborate on the possibility to properly identify some social function categories. We offer to unite land use categories that share similar usage and have common cellular communication characteristics.

The rest of this paper is organized as follows: Section 2 presents the needed background and related work; Section 3 describes the work process and method from the raw data to the extracted features, and random forest implementation; Section 4 presents the empirical evaluation of the classifier performance and detailed analysis of this performance for each land use; section 5 presents analysis of the features to identify patterns and characteristics that are typical of the different land uses; and in Section 6, we summarize the work, present our conclusions, and offer suggestions for further research.

2 Related work

Land use research falls into two main categories: *subjective* land use and *collective* land use. Subjective land use is the research of identifying the function of locations with significant importance to specific individuals and denotes them with a semantic label such as "home" or "work" [16]. Other works use semantic land use labeling as a part of mobility pattern analysis and next move prediction [7, 10, 17, 27]. Collective land use aims to identify the social function of the land for groups of people (residential neighborhood, commercial district). In this research, we focus on identifying the collective land use function.

The digital revolution has brought great opportunity for social sciences research in cities as the emergence of enhanced computing power and mobile phones with built-in sensors and location technologies has created enormous amounts of data for understanding and monitoring urban life [3]. Usage of data sources, such as remote sensing imagery, social media data, and taxi trajectories, and mobile phone patterns of usage are now utilized for more cost-effective and enhanced social land use identification research.

Numerous works leverage CDRs to capture spatiotemporal movement patterns and city dynamics [29, 38]. These records contain communication properties such as start time and call duration, and type of communication (call, SMS, internet), as well as the cell tower from which the communication originated. CDRs also include the location where the communication occurred—calculated by triangulating signal strengths from surrounding cell towers [30, 31, 39]. The greatest virtue of CDR as a location tool for human behavior evaluation is that it is routinely produced by telecom equipment when users make a phone call, send or receive a message, or browse Web pages; hence, it is an inexpensive and efficient location estimation source [33].

Several works have used CDR as their main data resource for land use identification. Toole et al. [30] utilized a CDR dataset and classified land use in the city of Boston into one of five categories: residential, commercial, industrial, parks, and other. They achieved an average accuracy rate of 54%, where their algorithm performed relatively well on residential, commercial, and industrial regions, but poorly on "parks" and "other." Pei et al. [23] also relied on CDR and offered a semi-supervised algorithm for classifying land in Singapore into the same five categories as Toole et al. [30]. They used the fuzzy *c*-means clustering algorithm and assumed possession of "real" land use labels of a small number of area segments. Their results also showed a modest detection rate of 58%, where mainly commercial and other classifications were confused.

Some works based their social land use analysis and identification algorithms on other data resources. For example, Lu and Weng [21] used an integration of population density data and remote-sensing systems to measure land surface temperature and spectral reflectance for classifying urban areas. Numerous works used image processing and classification techniques of remote-sensing images to capture the physical aspects of the land [13, 34, 35]. Others relied on data generated from social media, such as check-in data, GPS trajectories, and points of interest (POIs) [19, 25]. Specifically, POIs that are coordinates of a specific point location that carry a title, such as restaurant, shopping center, and theaters [36], were extensively leveraged mainly because they carry semantic information [12, 26].

However, as all data sources are limited and capture specific aspects of urban dynamics, a recent movement in land use identification research is to rely on several data sources of different types. Remote-sensing images and social media data are combined in some works [15, 20]. Both data resources can be seen as complementary, as remote-sensing image data is utilized to extract the physical location features, while social media data captures the social interactions [12, 18]. The work of Yuan et al. [37] integrated POI datasets and a dataset of 3 months of GPS trajectories generated by 12,000 taxicabs in Beijing to identify zones of different social function using an unsupervised clustering algorithm. The work of Tu et al. [32] integrated a mobile phone signal dataset with social media data to infer social function land use. They estimated individuals' home and work locations, and then aggregated their subjective land use together with social knowledge learned from social media check-in data for identifying collective social land use.

3 Method

In this section, we describe our methodology: data preparation (Section 3.1), feature extraction (section 3.2), and land use classification (Section 3.3). The core of the methodology is common with previously presented works [23, 30]. We offer some innovations to the methodology including types of features that were not used before and a different labeling approach that enables labels to vary throughout the day.

3.1 CDR dataset and data preparation

Our dataset consisted of CDRs recorded by an Israeli telecommunication company during a 62-day period, each day between 4 a.m. and 10 p.m., in a region covering a major part of Israel's center district. The data holds information gathered from numerous users, enabling mobility analysis on a large scale. Moreover, the 62 days of recording enabled us to study specific users during a relatively long period of time and examine their movement and communication habits. Aggregating the communication days and hours granted us the ability to locate the "average" circadian activity pattern.

We chose to analyze areas of unambiguous social function that can be used for analyzing the typical behavior in areas with different social functions and, thus, to assess the feasibility of land use identification. The disadvantage is that the chosen areas are less representative of normal urban behavior because the areas we selected are less mixed. We deliberately chose areas of varied social functions, such as neighborhoods, industrial zones, office areas, highways, and commercial districts and shopping malls. We selected 61 areas, spread in nine cities, all located in Tel Aviv and its surrounding areas including Holon, Ramat-Gan, Petah-Tikva, Rosh-Haayin, Ra'anana, Ramat-Hasharon, Givatayim, and Kfar-Saba. Figure 1 illustrates the areas selected in Kfar Saba. The four polygons represent the areas selected for classification. Areas numbered 1 and 2 are neighborhoods, the narrow rectangle of area 3 covers the main commercial street, and area 4 depicts the industrial area.

The chosen areas were divided into smaller units in a grid-like manner, with each unit of land denoted as a cell. Cells are the basic and highest resolution unit for land use classification and analysis in this work. Cell size was set as 200 × 200 m^2, the same shape and size as in the works of Toole et al. [30] and Pei et al. [23]. The inaccuracy of the location estimation in CDR, which can be up to 300 m distant from the actual signal location (depending on the density of cell towers), limits the possible analysis resolution and, hence, smaller cells could be too noisy for proper analysis. Notice that in areas in which a 200 × 200 m^2 cell would not fit, we use narrower cells. In addition to dividing space, we also divide time, as the function of the land can vary throughout the day. We divide the day into 24 hours; thus, the basic spatiotemporal unit used in this paper is "cell in an hour."

We deliberately chose areas whose social function is "pure" and, hence, relatively easy to be labeled semantically. We used knowledge from locals to label each cell in an hour to one of eight land use categories: Residential; Commercial; Industrial; Highway; Entertainment (recreation, nightlife, pubs, bars); Office; Street; and No activity (no human activity is expected in this cell at this specific time, for example, industrial areas before opening time). We refer to these eight categories as *atomic land uses*. Later, in this paper, we will analyze the possibility of distinguishing these land use categories using cellular communication data, and offer unions of categories that (1) share similar social function and (2) share similar communicational behavior and hence indistinguishable. The land use labels are used to analyze the communication behavior in each land use. Moreover, the labels will be used for training the classifier and evaluating its performance.

3.2 Feature extraction

Toole et al. [30] and Pei et al. [23] used a CDR dataset to extract features that depict the spatiotemporal calling variation

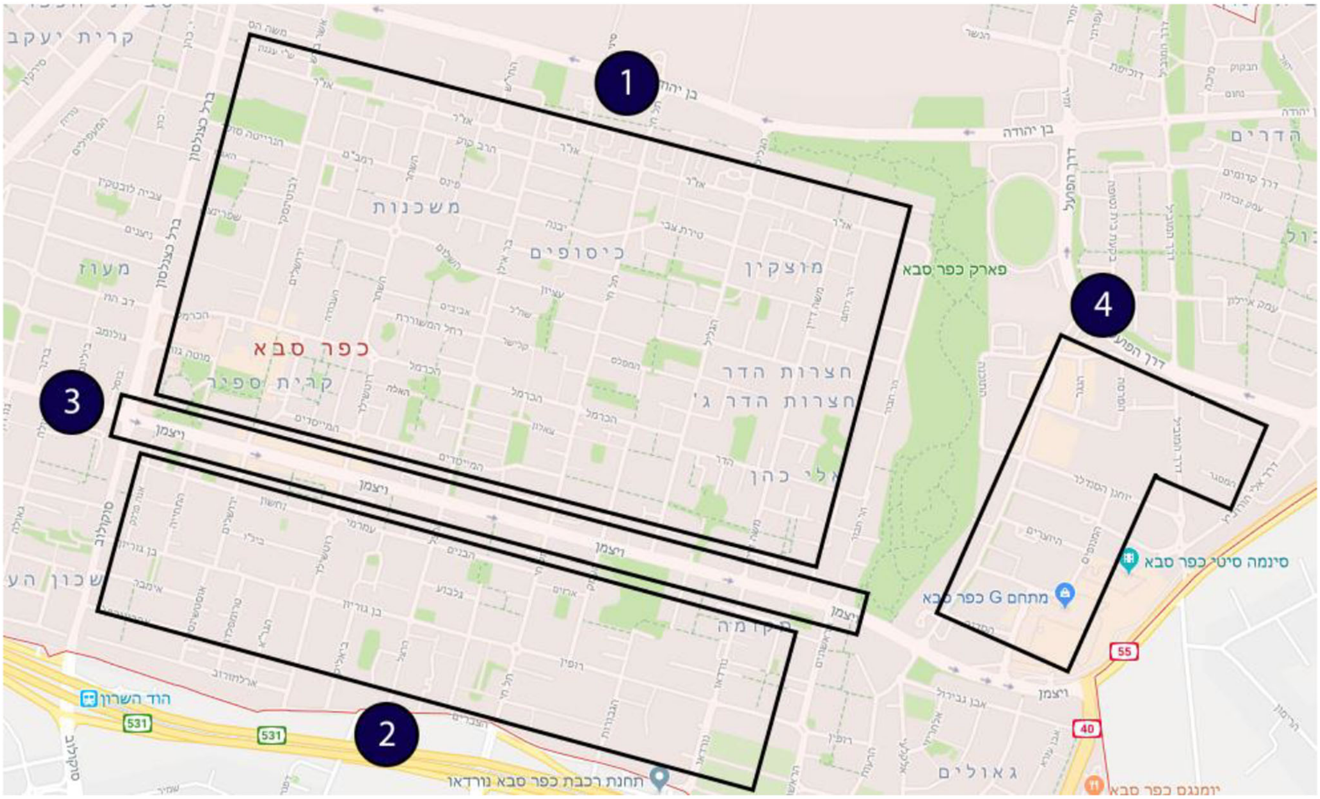
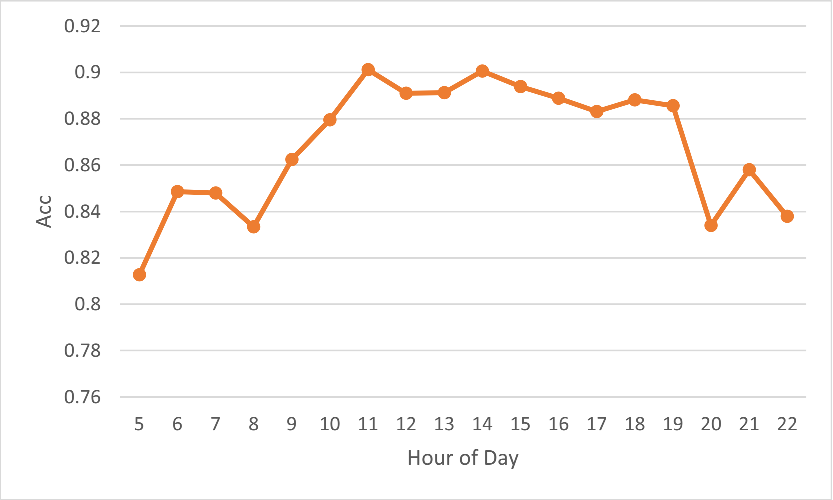


Fig. 1 Areas selected in the city of Kfar-Saba

pattern. One underlying assumption in their work was that the volume of calls in a specific cell (the number of phone calls) in a specific time segment was an indicator of the number of people present at that location at that specific time, and that different land uses have different spatiotemporal calling volume variation patterns. They utilized the circadian rhythm of the activity in the different land uses to differ between them. They used two types of features: *calling volume*, designated to capture the amount of activity in a period of time; and the *relative calling pattern*, extracted from the calling volume at a specific time relative to the calling volume at other hours of the day, and designated to capture the circadian activity pattern cleanly from the influence of the activity volume in the cell.

In this work, we extracted 158 features that illustrate different aspects of cellular communication in the spatiotemporal basic unit. We divided the features into five types, including two that were used in Toole et al. [30] and Pei et al. [23]: *communication volume features* and *daily pattern features*; and three feature types that have not been used in previous

Fig. 2 RF average accuracy as a function of the hour in a day



works: *weekly patterns*, *contacts*, and *communication habits*. The details of these features are:

Communication volume features These measure the degree of communicational activity and are designated to capture the difference between the activity volume typical to a specific social function (e.g., in commercial areas, there is more cellular communication compared with that in residential areas). These features are equivalent to the calling volume features in Toole et al. [30] and Pei et al. [23] but, here, we include different aspects of it: total activity volume, number of users, number of communications per user, etc. Notice that, because the cell area is not equal, the counting features are normalized by the cell size.

Daily pattern features These are calculated by the calling volume in a specific hour relative to the communication volume at different hours of the day in the same zone. These features are designated to identify the circadian pattern of the communication activity typical to an area (e.g., in residential zones, the communication peak hours are in the mornings and evenings, while in industrial areas, the peak is during work hours). These features are the equivalent of the relative calling volume features depicted in Toole et al. [30] and Pei et al. [23].

Weekly pattern features These capture the difference in cellular usage during weekdays compared with that during the weekend. Thus, they differentiate between land uses such as residential, to which its inhabitants return every day, and office zones, where workers usually do not go on weekends.

Contact features These measure the number of different days in which people engage in at least one cellular communication in cell s in hour h. Thus, they differentiate between land uses with frequent visitors and those with occasional ones.

Communication habit features These aim to depict the land from the perspective of its typical cellular communication usage habits, e.g., call duration and usage distribution of different types of cellular communication (phone calls, Internet usage). These features are used to examine if, in lands of different social function, there are prominent differences in communication behavior.

3.3 RF use for land use classification

We examined land use identification using the random forest (RF) algorithm for classification. This is an ensemble learning method for classification or regression. The essence of this method is to build multiple decision trees that are trained on randomly selected subsets of the samples and subspaces of the feature space, and outputting the class that is the mode [6, 14]. RF suits our problem well. Other state-of-the-art classifiers such as neural networks require that the number of samples would be much greater than the number of features. This is not the case in our dataset. We have 158 features and around 400 samples (in each hour)—a situation that RF can handle. Moreover, RF does not require performing feature selection in advance. It shows excellent performance even when most predictive variables are noisy [11] and is designed to resist overfitting. We used 8-fold cross-validation and, in each iteration, 7/8 of the cells were used for training and the other 1/8 of the cells was used as a test set. We partitioned the dataset into eighths because, in a preliminary study, this division was found to fit the problem well.

4 Classification evaluation

In this section, we implement the methodology described in Section 3 on the CDR dataset of 62 days of cellular communication in the center district of Israel. By analyzing these use case results, we demonstrate some of the possibilities and limitations of this methodology. In Section 4.1, we demonstrate the results over all land uses together, and in Section 4.2, we examine the performance over each land use separately and offer label unions.

4.1 Overall classification evaluation

In this section, we evaluate the overall classification results. The classification accuracy is between 91.2% in the lowest labeling resolution binary classification of residential/non-residential and 84% in the atomic labeling set, which is the most detailed. Compared with the works of Toole et al. [30] and Pei et al. [23], who also attempt to identify land use using CDRs, the accuracy rate is exceptionally high; Toole et al. [30] and Pei et al. [23] achieved 54% and 58% accuracy rates, respectively. However, the accuracy rates of the works are incomparable. The main reason for this is that their work performed land use identification for entire cities: Boston in Toole et al. [30] and Singapore in Pei et al. [23]. However, we chose areas from different cities located in Israel, but did not include a whole city. We deliberately focused on areas with a relatively "pure" and clear land use function; hence, these were easier to classify. We elaborate on more reasons in Section 6.

Figure 2 demonstrates the average accuracy rate (Acc) in the different hours of the day using the atomic labels. The classifier performed best between the work hours of 10 a.m. and 7 p.m.; during these hours, accuracy did not fall below 88%.

Table 1 demonstrates the land use classification confusion matrices during morning (Table 1(a)), work hours (Table 1(b)), early evenings (Table 1(c)), and late evenings (Table 1(d)). We separated the day into four parts because

Table 1 Confusion matrices of the classification results. Rows and columns hold true and predicted values, respectively

(a) 4 a.m.–7 a.m.					
	Residential	Street	Highway	No activity	
Residential	*47.15%*	0.61%	0.46%	1.90%	
Street	3.09%	*9.08%*	0.52%	1.24%	
Highway	1.79%	1.31%	*0.54%*	1.81%	
No activity	1.81%	1.35%	0.33%	*27.02%*	
(b) 8 a.m.–5 p.m.					
	Residential	Commercial	Industrial	Office	
Residential	*43.90%*	1.86%	0.61%	0.06%	
Commercial	3.63%	*16.11%*	1.78%	0.03%	
Industrial	0.35%	1.09%	*28.43%*	0.11%	
Office	0.02%	0.31%	1.36%	*0.36%*	
(c) 5 p.m.–7 p.m.					
	Residential	Commercial	Office	No activity	
Residential	*44.15%*	1.91%	0.03%	0.74%	
Commercial	3.30%	*16.82%*	0.03%	1.65%	
Office	0.15%	0.24%	*0.38%*	1.30%	
No activity	0.21%	1.56%	0.21%	*27.33%*	
(d) 8 p.m.–10 p.m.					
	Residential	Street	Highway	Commercial	No activity
Residential	*45.23%*	0.06%	0.24%	0.12%	1.60%
Street	2.38%	*0.36%*	0.18%	1.04%	0.51%
Highway	1.40%	0.51%	*2.23%*	0.39%	0.53%
Commercial	1.13%	0.42%	0.59%	*9.24%*	1.40%
No activity	0.74%	0.00%	0.42%	0.56%	*28.74%*

the set of land use categories changes throughout the day. Some of the social functions, such as Commercial, occur only in specific hours, while other social functions, such as Highway and No activity, occur all day long, but not necessarily in the areas we chose. For example, in our dataset, there is no cell labeled as No activity between 8 a.m. and 5 p.m. Most land uses are well identified: residential, commercial, industrial, and no activity are relatively well identified throughout the day. Office is confused, especially with Industrial (Table 1(b)) and No activity (Table 1(c)). However, notice there are only seven office cells in this work. Highway is confused for other land uses (Table 1(a) and (d)). Street is relatively well identified in the morning (Table 1(a)); however, it is confused in the late evening (Table 1(d)). It is mostly confused for Residential, which is not surprising because both are located inside neighborhoods.

In Fig. 3, we visualized the classification results on a geographical map during the work hours 8 a.m. to 5 p.m., with atomic labeling in three cities: Fig. 3a, Ra'anana; Fig. 3b, Ramat-Gan; Fig. 3c, Tel Aviv. We refer to this map as a *confusion map*. It resembles a confusion matrix, but it displays the results on a geographical map, with each cell (sample) placed where it is located. The legend displays the set of colors representing the four land use classes in these hours. The colored circles beside each batch of cells indicate the "real" land use label of the cell batch which is located to its side. The color of the cells indicates the land use it is classified as. Notice that some of the cells contain more than one color. This is because the results in these maps accumulate 45 classification results (9 h from 8 a.m. to 5 p.m. × 5 iterations of random training–testing partitioning).

Figure 4 focuses on Ra'anana (exactly Fig. 3a, but with marks used for explanation). The cell marked "a" is in blue, yellow, and green. Yellow is the most dominant color; it indicates that this cell was mostly classified as Industrial. Blue and green indicate that, in some of the runs, this cell was classified as Residential and Office, respectively. The cell to its left marked with "b" is all blue, indicating that it was classified as Residential in all the runs unanimously.

Communication behavior is similar across different cities: One encouraging result that can be noticed when examining the maps is that human land uses characterized by cellular communication are similar across different cities. For example, commercial cells from different cities such as Tel Aviv

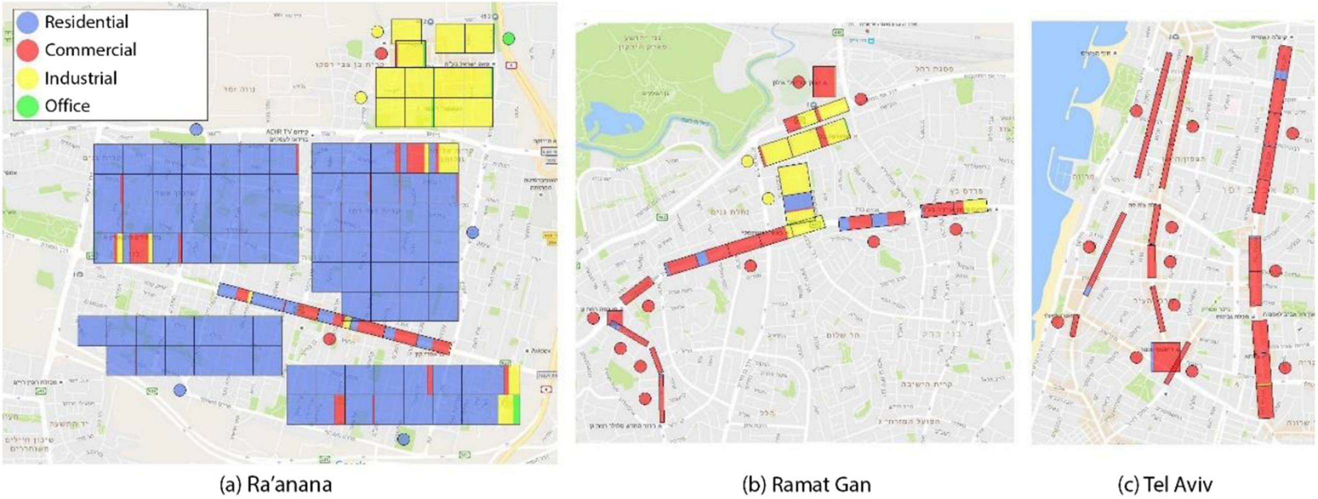
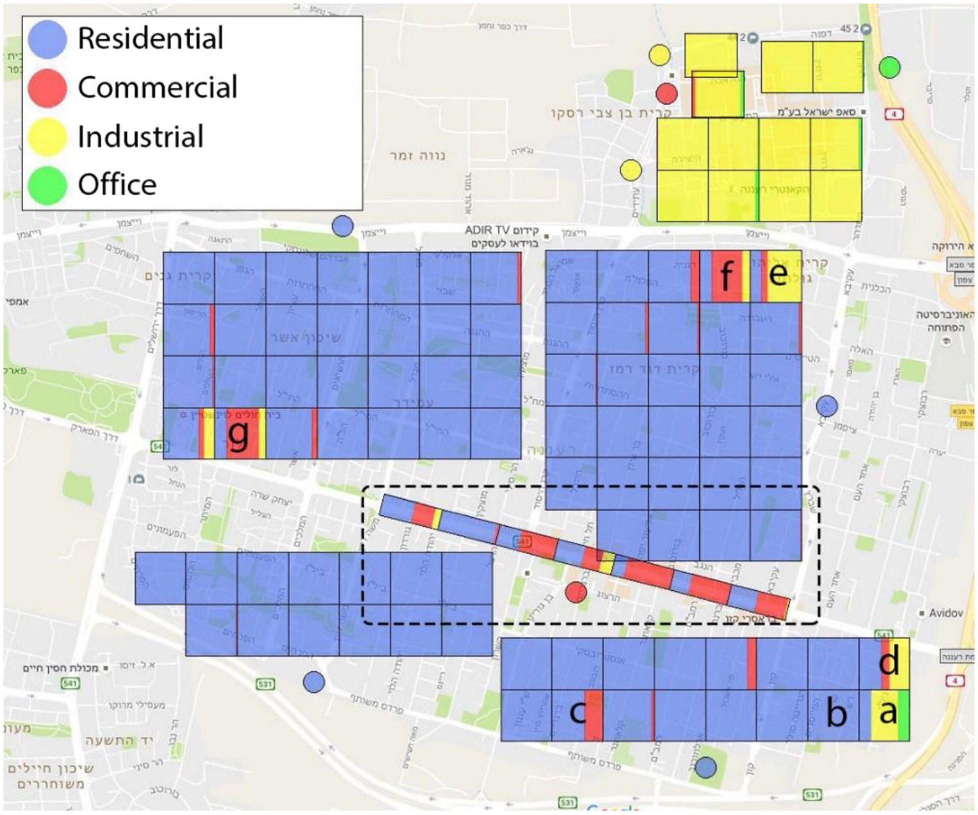


Fig. 3 Confusion maps of the classification results during work hours (8 a.m.–5 p.m.)

and Ramat-Gan are well distinguished from other land uses; the algorithm recognizes that they share a similar land use and classify both as Commercial.

Relatively low variance The results are relatively consistent, as indicated by a significant number of cells with only one color, meaning that all 45 repetitions classified as the same class. For example, see the residential neighborhoods in Ra'anana in Fig. 4, where 69 of the 86 cells are blue, indicating that they were classified as Residential in all the repetitions. Moreover, retrospective analysis reveals that most of the cells that were not classified as Residential in all the runs do not include a "pure" residential land use: the cells marked with "a" and "d" are on the outskirts of the neighborhood and are mostly unpopulated, while the cell marked "c" mostly includes a large school. For another example, see the commercial streets of Tel Aviv in Fig. 3c (from east to west): Ibn Gabirol, Dizengoff, and Ben-Yehuda. Fifteen of the 22 cells that are included from these three streets are red, indicating they were classified as Commercial in all runs. Five of the 22 cells in the commercial

Fig. 4 "Zoom in" on Ra'anana's confusion map of the classification results



streets in Tel Aviv have a negligible blue color (indicating they were seldom classified as Residential), and only two cells are non-negligible blue, indicating that, in some of the runs, they were classified as Residential. Tel Aviv's commercial streets also serve residential purposes; therefore, it is not surprising that sometimes they are classified as Residential.

Relatively low bias Examining the cells that are not correctly classified in all the repetitions, it seems that the algorithm correctly classifies them in most cases. Examining the majority vote of each cell (the most dominant color in the cell), 82 out of 86 Ra'anana's residential cells, 12 out of 16 commercial cells in Ramat-Gan, and all 24 commercial street cells in Tel Aviv are correctly classified. Examining Residential and Commercial classifications at other hours and in other cities indicates that this example of RF identifying them with a high rate is not alone: 93.5% of the residential cells are classified by the RF as Residential, and 74.7% of the commercial streets and malls are identified as Commercial.

Effect of the leaking phenomenon The classifier has difficulty identifying what we refer to as "island" land uses—small or narrow-shaped land uses surrounded by a different land use. For example, see Ra'anana's main commercial street, Ahuza, surrounded by a broken line in Fig. 4. Ahuza St. is located in the heart of a residential neighborhood, and it is frequently classified as Residential. The classifier has difficulty in correctly identifying island land uses because location estimation inaccuracy causes communication transmissions originating from one cell to fall inside the borders of its neighboring cell. We refer to this as the leaking phenomenon. Because the triangulated signal strength location estimation technology used for the location estimation suffers from high inaccuracy, the extent of the problem is not negligible. Small and narrow streets which are surrounded by a "sea" of residential neighborhoods are especially affected by this leaking. The leaking phenomenon also causes cells located at the border between different land uses to frequently be confused. For example, notice the cells in the outer layers of the residential neighborhoods of Ra'anana marked with "e," "f," and "g." Retrospective analysis indicated that these three cells are all part of the neighborhood. However, cells "e" and "f" are confused for the industrial and commercial areas that are to its side, and "g" is confused for the commercial street that is close by. The cells in the outer layer of the areas are not classified as well as the cells located in the center of them.

4.2 Analyzing classification of each land use separately and consideration of alternative labeling sets

In this section, we analyze the performance of the classifier on each land use separately. We focus on land uses that are often confused, and suggest labeling sets that unite some of the problematic land uses. For evaluating the quality of the classification to a specific land use, we use the F1 score. This score incorporates two crucial aspects for analyzing the quality of a classification to a specific land use: the land use precision, which is the percentage of cells classified to the specific land use c that are classified correctly; and the land use recall, which is the percentage of cells of the specific land use that are classified correctly:

$$Precision_C = \frac{TP_C}{TP_C + FP_C} \tag{1}$$

$$Recall_C = \frac{TP_C}{TP_C + FN_C} \tag{2}$$

where TP_C is the number of samples of class c that are classified as class c (classified correctly), FP_C is the number of samples classified as class c but are not c (classified incorrectly), and FN_C is the number of samples of class c that are classified to another class (classified incorrectly).

F1 score is the harmonic average of precision and recall:

$$F1_C = 2\frac{Precision_C \cdot Recall_C}{Precision_C + Recall_C} \tag{3}$$

Table 2 illustrates the precision, recall, and F1 scores for the classification of each land use over all cells in the nine cities. The table also demonstrates scores of label unions. For example, the label that unites Commercial and Entertainment labels is referred to as {Commercial, Entertainment}. In that case, both cells with Commercial and Entertainment atomic labels are labeled the same {Commercial, Entertainment}.

As reflected by Table 2, residential is well identified and distinguished (F1 score is 0.91). It is the most common land use in urban areas; therefore, correct identification of it is important. In our work, 47% of the cells are Residential.

Industrial and Commercial are also relatively well identified (0.91 and 0.66, respectively), and both are also prominent in the areas chosen for this work. Such cells during work hours are 30% and 21% of the cells, respectively. The Commercial identification rate is high even though it suffers from the leaking phenomenon more than the other land uses. That is because most Commercial cells are located on streets, close to a residential neighborhood, and sometimes surrounded by a neighborhood; they are more vulnerable to location estimation inaccuracy.

Notice that Table 2 does not display the change of the classification accuracy throughout the day. That is because the identification rate of the land uses is relatively stable throughout the day. One exception is Commercial. Although it is well identified overall, between 8 p.m. and 10 p.m., all Commercial cells were classified as Entertainment. In the areas we chose, Entertainment occurs only between 8 p.m. and 10 p.m., and apparently, Commercial and Entertainment were indistinguishable by the classifier.

We would like to offer a set of land use categories that is more suited for land use classification by uniting some of the atomic land use categories. For uniting land uses, we would like to consider two aspects: (1) the land use categories share similar social function; and (2) the land use categories share similar communicational behavior, and hence, they are indistinguishable by the classifier. For example, Commercial and Entertainment satisfy the two requirements for union as they are obviously indistinguishable by the classifier and both share similar social function.

No activity is also well identified (F1 is 0.89). Intuitively, we could assume that it would be well distinguished because the classifier can recognize the small amount of communication in these areas.

The algorithm does not identify Highway and Street well, especially that Highway in the morning and Street in the evening suffer from low F1 scores. At least to a certain extent, this is because of the leaking phenomenon as both Highway and Street are narrow and lie beside other land use areas.

We examined label unions for Street: {Highway, Street} and {Residential, Street}. The motivation for uniting Street is that the classifier poorly identifies it at certain hours, and uniting with one of the two options is reasonable because, by definition, Street cells are between highway and residential zones.

We compared the union options using the sub-land use recall, demonstrating the ability to correctly identify the cell l in union with land uses m, $Recall_{l,\{l,m\}}$.

Table 2 Precision, recall, and F1 of each land use and some of their unions

Land uses	Precision	Recall	F1
Residential	0.90	0.93	0.91
Commercial	0.70	0.62	0.66
Industrial	0.89	0.95	0.91
Office	0.69	0.16	0.25
Entertainment	0.80	0.68	0.74
Highway	0.44	0.27	0.34
Street	0.54	0.36	0.43
No activity	0.87	0.91	0.89
{Residential, Street}	0.91	0.93	0.92
{Highway, Street}	0.67	0.54	0.60
{Highway, Street, No activity}	0.83	0.84	0.83
{Commercial, Entertainment}	0.86	0.65	0.74
{Commercial, Industrial}	0.89	0.91	0.90
{Commercial, Office}	0.80	0.71	0.75
{Industrial, Office}	0.92	0.95	0.93

The percentages of cells labeled l that are classified correctly to the union label of l and m, label $\{l,m\}$, is,

$$Recall_{l,\{l,m\}} = \frac{TP_{l,\{l,m\}}}{TP_{l,\{l,m\}} + FN_{l,\{l,m\}}}$$

where $TP_{l,\ \{l,m\}}$ is the number of cells labeled l that are classified as $\{l,m\}$ (classified correctly), and $FN_{l,\ \{l,m\}}$ is the number of cells labeled l that are *not* classified as $\{l,m\}$ (classified incorrectly).

The recall values of $\mathrm{Recall}_{\mathrm{Street},\ \{\mathrm{Street}\}}$, $\mathrm{Recall}_{\mathrm{Street},\ \{\mathrm{Street,Road}\}}$, and $\mathrm{Recall}_{\mathrm{Street},\ \{\mathrm{Street,Residence}\}}$ are 0.348, 0.517, and 0.778, respectively. For example, $\mathrm{Recall}_{\mathrm{Street},\ \{\mathrm{Street}\}}$=0.348 means that 34.8% of the cells with an atomic label of Street are correctly classified, and $\mathrm{Recall}_{\mathrm{Street},\ \{\mathrm{Street,Residence}\}}$ = 0.778 means that 77.8% of the cells whose atomic label is Street are correctly classified to the union label of Street and Residential. Street recall is significantly higher when united with Residential. We already saw an indication of the resemblance between Street and Residential in Table 1; Street cells are frequently classified as Residential, especially in the evening (Table 1(d)).

Another option for a label union is Highway and Street with No activity. It is reasonable to unite these land uses because people in these zones use them only for mobility purposes. The results of this union, as indicated in Table 2, are good, much better than those of Highway and Street separately. However, because 30.3% of the cells are No activity compared with only 14.1%, which are labeled Highway or Street, F1 may be majorly influenced by the classifier's ability to identify No activity. The recall values $\mathrm{Recall}_{\{\mathrm{Highway,\ Street}\},\ \{\mathrm{Highway,\ Street}\}}$ and $\mathrm{Recall}_{\{\mathrm{Highway,\ Street}\},\ \{\mathrm{Highway,\ Street,\ No\ activity}\}}$ are 0.534 and 0.718, respectively. Uniting Highway and Street with No activity seems to be beneficial.

The classifier does not identify Office well (F1 is 0.25). We assume the main reason is the lack of Office-labeled samples—only seven cells in the data are labeled as Office, i.e., in an average fold of the eight-fold cross-validation, there are only six Office-labeled samples in the training set, and one sample in the test set. We examined two options for uniting Office: {Commercial, Office} and {Industrial, Office}. The sub-land use recall values $\mathrm{Recall}_{\mathrm{Office},\ \{\mathrm{Office}\}}$, $\mathrm{Recall}_{\mathrm{Office},\ \{\mathrm{Office,\ Industrial}\}}$, and $\mathrm{Recall}_{\mathrm{Office},\ \{\mathrm{Office,\ Commercial}\}}$ are 0.177, 0.895, and 0.434, respectively. Office is best identified when in the {Industrial, Office} union—around 90% of the Office-labeled cells are identified as {Industrial, Office}, significantly better than uniting with Commercial or not uniting it at all.

The land use unions that we found to be effective and offer advantage to non-union labels are as follows: {Commercial, Entertainment}, {Office, Industrial}, and {Residential, Street} or {Highway, Street, No activity}. Thus, we would suggest using one of two sets of land use categories:

Label set 1:

- {Residential, Street}
- {Commercial, Entertainment}
- {Office, Industrial}
- Highway
- No activity

Label set 2:

- Residential
- {Commercial, Entertainment}
- {Office, Industrial}
- {Highway, Street, No activity}

Figure 5 compares the average accuracy of the atomic label sets, label set 1 and label set 2, throughout the day.

Both land use sets that include label unions improved accuracy. The most significant improvement was in the mornings and evenings, in which the atomic label set is not as good as during work hours. However, notice the accuracy versus land use resolution trade-off. Label set 2 that has the highest accuracy also has the lowest resolution (four land uses); and the atomic set that has the lowest accuracy also has the highest resolution (eight land uses).

5 Feature analysis

The 158 features used in this work capture different dimensions of the cellular communication pattern and thus can be utilized to better understand the relations between cellular communication and human activity. Table 3 illustrates a representative sample of seven of the 158 features, selected to demonstrate the different feature types introduced earlier (Section 3.2): CountComsWeek and CountComsEnd of the Communication volume type, PropHourCallsWeek of the daily pattern type, PropWeekAtEnd and CountComsWW of the weekly pattern type, and AvgContactCallsWeek and AvgContactCallsEnd of the contact type. Later in this section, we will elaborate and explain each of these seven features.

Table 3 displays the average value of these features in each land use. It can be used to point out the typical behavior in each land use. The feature values are normalized over all cells and in each hour separately, so they will be distributed as standard normal. As an example, see the feature CountComsWeek—this feature measures the number of cellular signals conducted in the cell during weekdays (weekend is not included) per square meter. The land use with the highest volume cellular communication volume is Office with the exceptionally high value of 2.33, whereas it is not surprising to see that the least activity is conducted in No activity–labeled cells. Thus, this feature can be effective to distinguish between Office and No activity. Notice the colors emphasize the normalized values of the features, where warm orange colors indicate high values and cold blue colors indicate low values.

In addition, Table 3 illustrates in curved brackets the *variable importance* (VI) measure suggested by Breiman [6] that measures the importance of the feature in the classification. It is computed throughout the RF training process and is based on a permutation test. The idea is that if the variable is not important (the null hypothesis), then rearranging the values of that variable will not degrade prediction accuracy. The VI of a feature is computed as the average accuracy decrease on the out-of-bag samples when the values of the respective feature are permuted randomly [2]. A high VI measure indicates that a feature is important for good classification.

The *communication volume features* are designated to differentiate between land uses with different levels of cellular activity. They are the equivalent of the calling volume features in the works of Toole et al. [30] and Pei et al. [23].

We demonstrate these features' behavior by examining CountComsWeek and CountComsEnd. CountComsWeek, which is the 49th best feature with a VI measure of 0.88, and estimate the average number of cellular communications per hour in a square meter on weekdays (Sunday to Thursday in Israel). The value of this feature for Office is exceptionally high (2.33). However, it cannot be generalized to all Office land uses because, in our dataset, there are only seven Office cells. Nevertheless, it is rather unsurprising that there is high cellular communication activity in Office areas. Although shadowed by the Office exceptional values, Commercial and Street also have a relatively high communications volume (0.45 and 0.49, respectively). Loyal to its definition, there is indeed a low number of communications in No activity (− 0.48).

CountComsEnd is the equivalent weekend feature to CountComsWeek. It estimates the average number of cellular communications per hour in a square meter on weekends (Friday and Saturday in Israel). It is the feature with the highest VI of all the features of this type (a VI measure of 1.1, leading to the 25th best feature). Commercial and Street have high activity, as on weekdays (0.85 and 0.70 respectively). Again, there is a low number of communications in No activity (− 0.63). In Industrial, the number of communications is low (− 0.70) because most businesses are closed then. Also, the number of communications in Office is much more moderate on weekends because most offices are also closed.

The *daily pattern features* examine the communication behavior of a cell in an hour compared to the communication behavior in the same area during the rest of the hours. As such, they differentiate between land uses with different day routines.

These features are the equivalent of the relative calling volume features depicted in Toole et al. [30] and Pei et al.

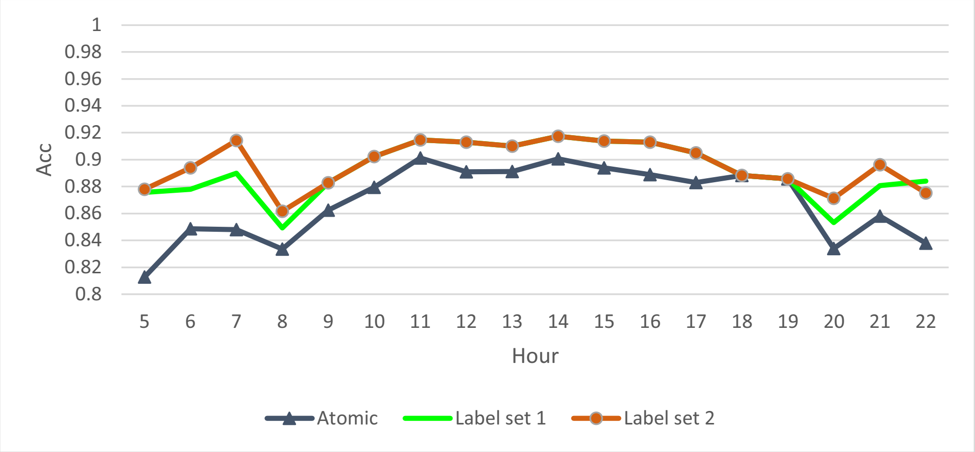


Fig. 5 Label union set accuracy comparison (accuracies are identical for label sets 1 and 2 between 9 a.m. and 7 p.m.)

[23]. They extracted these features to capture the spatiotemporal variations patterns of human mobility without the noise inherent in the communication volume features (e.g., some commercial streets are very crowded in big cities, and in smaller cities, they can be significantly less active).

The daily pattern features have an average VI of 1.02, which is second only to the contact features. We will demonstrate the daily pattern features' behavior by analyzing the feature PropHourCallsWeek, which has the highest VI of all daily pattern features (a VI measure of 1.51, which is the 14th best feature). PropHourCallsWeek compares the number of communications in cell s in hour h to the number of communications in cell s on hours different from h. A positive value indicates that the number of phone calls in cell s in hour h is higher than the mean number of phone calls in cell s in hours different from h; likewise, a negative value indicates that the number of phone calls in cell s in hour h is lower than the mean number of phone calls in cell s in hours different from h.

The feature differentiates between land uses where the core of activity is at different hours. For example, it captures the difference between Industrial, whose core hours are during work hours, and Commercial, whose core hours are after work hours. Mainly, it differentiates between occupation-related land uses, in which the activity concentrates mainly during work hours, and leisure-related land uses, in which most of the activity does not occur during work hours.

Because this feature captures the daily pattern of the land use, the values change throughout the day. Therefore, we use Fig. 6 to demonstrate the values of the normalized PropHourCallsWeek changing throughout the hours of the day (between 5 a.m. and 10 p.m. because the dataset does not include nighttime). It may seem like most values are positive; however, it is just because the land use categories are not of the same size, e.g., Residential that has negative normalized values during the work hours is a big category that includes 47% of the cells. Industrial and Office are most active during work hours (9 a.m. to 5 p.m.). Most cells labeled as Industrial or Office are labeled as No activity before and after work hours; thus, No activity is somehow complementary to Industrial and Office at those hours and as we expected the activity in these places is lower than during work hours. Therefore, it is low during leisure hours. Residential, Commercial, Highway, and Street are active during leisure hours when people are not at work. In the morning, people are at home (Residential is high). From 6 a.m., people start driving to work (Highway and Street are high). During work hours, fewer people are present at home or in commercial areas (Residential and Commercial are low), and in the evening, people return home or go to commercial areas (Residential and Commercial are high).

The *weekly pattern features* capture the difference in communication activity on weekdays compared with weekends. They include PropWeekAtEnd, which has the highest VI (1.97). It measures the percentage of users who had at least one communication usage in cell c in hour h on weekdays (Sunday to Thursday) who also had at least one communication usage in cell c in hour h on the weekends (Friday and Saturday). This feature may help identify cells where visitors go in different parts of the week, i.e., differentiating between cells visited only on weekdays and those visited on weekends as well.

Naturally, this feature has a high value in residential areas (0.75); people spend time in their homes all through the week; therefore, Residential has the highest value. People are in their office and in industrial areas only on weekdays; therefore, Office and Industrial have the lowest values (− 1.09 for both). Commercial has a moderate value (− 0.07), probably because most of the commercial centers and streets in our data are open on weekends, and workers are present on weekends as well.

We will examine another weekly pattern feature: CountComsWW (VI measure 0.94, 43rd best feature). It measures the weekday to weekend ratio of the number of communications. We expect it to differentiate land uses that have a different volume of cellular usage on weekdays compared

Table 3 A sample of seven features and their average normalized value in each land use (Resid, Residential; Highw, Highway; Comm, Commercial; Ind, Industrial; Off, Office; No Act, No activity)

	Land uses						
Feature {VI}	Resid	Street	Highw	Comm	Ind	Off	No Act
CountComsWeek {0.88}	− 0.02	0.49	− 0.05	0.45	− 0.15	2.33	− 0.48
CountComsEnd {1.1}	0.13	0.70	− 0.04	0.85	− 0.70	− 0.18	− 0.63
PropHourCallsWeek {1.51}	− 0.01	0.15	0.10	− 0.04	0.88	0.73	− 1.09
PropWeekAtEnd {1.97}	0.75	0.23	− 0.37	− 0.07	− 1.09	− 1.09	− 0.98
CountComsWW {0.94}	− 0.38	− 0.44	− 0.08	− 0.41	0.83	1.70	0.69
AvgContacCallsWeek {1.18}	0.48	− 0.29	− 0.60	− 0.71	− 0.02	1.47	− 0.73
AvgContacCallsEnd {1.81}	0.75	− 0.15	− 0.45	− 0.31	− 0.94	− 0.87	− 0.81

with weekends. As one could expect, in Office and Industrial, there is significantly more activity in the weekdays comparing to the weekend. Their non-normalized values of 7 and 4, respectively (the values in Table 3 demonstrate the normalized values), indicate that, on weekdays, they have seven and four times more communication activity than on weekends. Residential, Commercial, Street, and Highway have lower values. Their non-normalized values are around 1, indicating that the number of communications on weekdays is similar to the number on weekends.

The *contact features* measure the number of distinct dates in which people perform a cellular communication (i.e., contact) in cell s in hour h. Thus, they differentiate between land uses with frequent visitors and those with occasional ones. This is the feature type with the highest average VI (1.04). Moreover, 13 of the 14 features with the highest VI are contact features. We will demonstrate these by analyzing the features AvgContactCallsWeek and AvgContactCallsEnd. AvgContactCallsWeek (VI measure 1.18, 20th best feature) measures the average number of weekday contacts per user in cell s in hour h over the 62-day recording period. Residential has a high value (0.48) because the same residents frequently return home usually around the same hours; likewise, Industrial and Office have higher values than most of the other land uses (− 0.02 and 1.47, respectively) because workers go to offices, or business, in Industrial areas, every weekday. Commercial has a low value (− 0.71) because on commercial streets, there are workers who frequently visit their workplace, but they are negligible in the vast stream of occasional visitors shopping or passing by. Highways also host a variety of different people; hence, Highway has a low value (− 0.6).

The equivalent weekend AvgContactCallsEnd feature (VI measure 1.81, 2nd best feature) measures the average number of weekend contacts per user in cell s in hour h. This feature is valuable in distinguishing Residential from other land uses. On weekends, this is the only land use with frequent visitors. Residential has a high value (0.75) because people tend to be at home in weekends, and all the other land uses have a significantly lower frequent visitor value in the weekend. Office and Industrial, which had a high contact rate in the parallel week feature, are very low on weekends because workers are not present then (− 0.87 and − 0.94, respectively).

The *communication habit features* added little contribution to the classification. The average VI of this feature type is the lowest of all five types (0.28), and the most valuable feature of this type is AvgDurationIntrWeek that measures the average duration of internet usage. It is only the 55th best feature with a VI of 0.76.

The two feature types that were used in Toole et al. [30] and Pei et al. [23], the communication volume features and the daily pattern features, are also useful for classification in this work. We introduced three more feature types that were crucial for the classification, the weekly pattern features that include the feature with the highest VI, and the contact features that have the highest average VI.

The typical communication behavior in the different land uses can be captured by the features as demonstrated in Table 3. However, the features contain noise and even exceptional values that complicate land use classification.

Figure 7 illustrates a Tukey boxplot that demonstrates the normalized values of the feature AvgContactCallsEnd in each of the land uses active during work hours. The red graph is the median value, and the bottom and top of the box are the first and third quartiles. The boxplot includes whiskers, the dotted lines extending vertically from the boxes. The lower whisker is the lowest datum still within 1.5 IQR (interquartile range), and the higher whisker is the highest datum still within 1.5 IQR. Any data not included in the whiskers range is plotted as an outlier with a cross.

As demonstrated in Fig. 3, this feature captures the typical behavior in the different land uses; however, using only this feature, it is impossible to distinguish between the land uses with a high accuracy. This feature is valuable to distinguish between Residential and the rest of the land uses; however, still 25% of the Residential cells are inside the IQR of

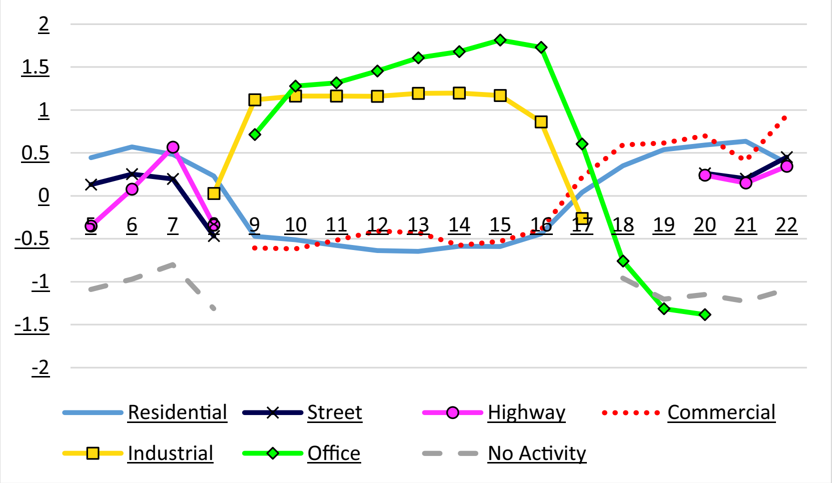


Fig. 6 PropHourCallsWeek for each land use per hour of the day

Commercial. Office and Industrial are almost indistinguishable, and Commercial can be confused with any of the other land uses.

6 Conclusions

In this paper, we offered the reader a chance to take a deep look into the different sides of land use identification based only on cellular communication—its advantages along with the shortcomings. We implemented a common methodology for solving this problem, with some innovations, to illuminate different angles of the problem. We applied the methodology on a dataset of 62 days of cellular communication in the center district of Israel. We utilized this use case to point out patterns and behavior as reflected by features aggregating cellular communication. We demonstrated that communication behaves reasonably similar across cities, and thus, we believe, it can be used for classifying land use in a city with data of other cities to achieve reasonable results.

We also offered some innovations to the methodology. We introduced types of features not used in previous works, and

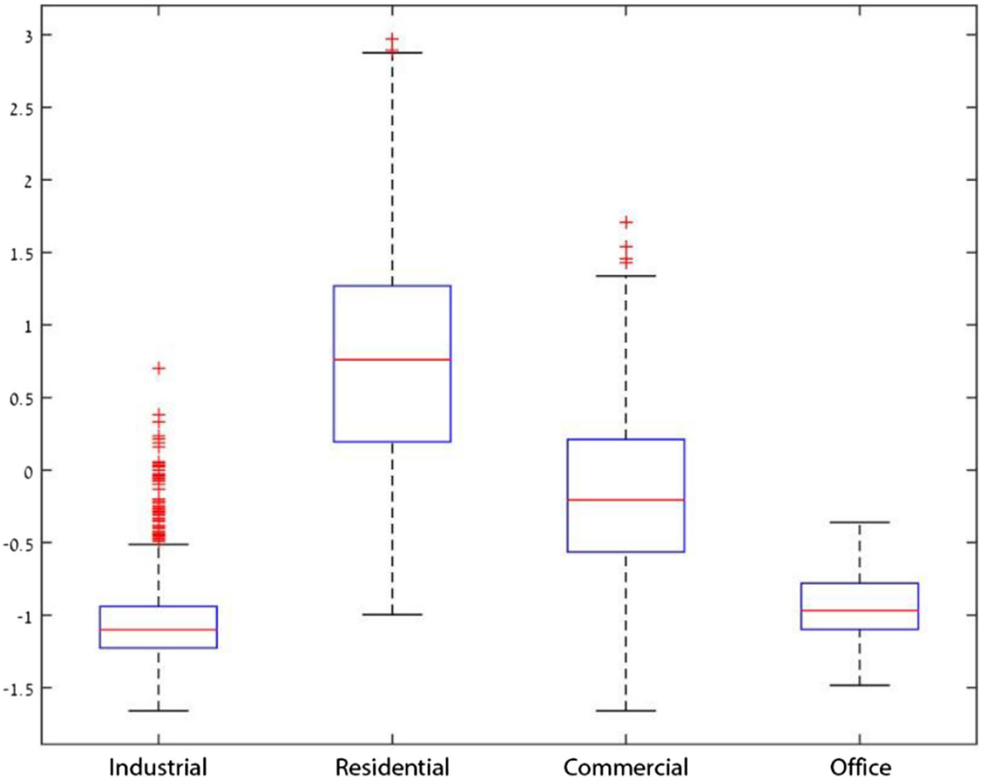


Fig. 7 Boxplot of the normalized values of AvgContactCallsEnd in each land use (during work hours)

two of these proved to be dominant in the classification process: the weekly pattern features that differentiate between land uses with major differences in activity on weekdays versus weekends, and the contact features that differentiate between land uses with frequent visitors (e.g., residential neighborhood) and land uses of occasional visitors (e.g., commercial streets). We used an hourly labeling method instead of the constant through time land use label that has been used in previous works. An hourly land use labeling set enables the flexibility to recognize variations and temporal social functions.

The accuracy rate of identifying the land uses in this work showed a relatively high accuracy rate ranging from 91.2%, when using the lowest binary residential/non-residential classification, to 84% when using a high-resolution labeling set. The classification performs best during work hours, as its accuracy does not drop below 88% even in the high-resolution atomic label set. Although the accuracy rate in this work is higher than that in other works that also focused only on cellular communication as a data resource, such as the works of Toole et al. [30] and Pei et al. [23], it cannot be deduced that the difference in the methodology led to the higher accuracy. The works are incomparable because the earlier works performed land use identification for entire cities: Boston in Toole et al. [30], and Singapore in Pei et al. [23]. However, we chose areas from different cities located in Israel, but did not include a whole city. We deliberately chose areas with a relatively "pure" and clear land use function; hence, it was easier to classify them. Moreover, Toole et al. [30] and Pei et al. [23] used the label "open spaces" to label parks and areas without buildings, whereas we did not include these in our work. Another difference is our hourly labeling method, whereas Toole et al. [30] and Pei et al. [23] used the constant through time land use labeling.

Although the overall accuracy is satisfactory, checking each land use separately indicates that some of the social functions such as Office, Entertainment, and Street were poorly identified. Therefore, we tested options for uniting land uses and recommended uniting some land uses that were indistinguishable and shared similar social function and communication behavior. We suggested combining Commercial with Entertainment and Office with Industrial because these two pairings were barely distinguishable by the classifier. We discussed the consequences of the leaking phenomenon caused by the inaccuracy of location estimation. The communication transmissions that originate from one cell and fall inside the borders of its neighboring cell damage the ability to correctly classify land uses and land located nearby lands of other use, particularly narrow land uses such as commercial streets.

In future work, it would be interesting to further investigate and offer solutions for the leaking phenomenon caused by the inaccuracy of location estimation. It would also be valuable to evaluate this method on a whole city dataset to understand if the additions to the methodology significantly improved its efficiency.

Funding information This research was supported by a grant from the Israel Ministry of Science and Technology.

References

1. Alberti M, Marzluff JM, Shulenberger E, Bradley G, Ryan C, Zumbrunnen C (2003) Integrating humans into ecology: opportunities and challenges for studying urban ecosystems. AIBS Bull 53(12):1169–1179
2. Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: a corrected feature importance measure. Bioinformatics 26(10):1340–1347.3
3. Arribas-Bel D, Tranos E (2018) Characterizing the spatial structure(s) of cities "on the fly": the space-time calendar. Geogr Anal 50(2):162–181
4. Ben Zion E, Lerner B (2017) Learning human behaviors and lifestyle by capturing temporal relations in mobility patterns. European Symposium on Artificial Networks, Computational Intelligence and Machine Learning (ESANN2017), Bruges
5. Ben Zion E, Lerner B (2018) Identifying and predicting social lifestyles in people's trajectories by neural networks. EPJ Data Sci 7(45):1–27
6. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
7. Calabrese F, Diao M, Di Lorenzo G, Ferreira J Jr, Ratti C (2013) Understanding individual mobility patterns from urban sensing data: a mobile phone trace example. Transp Res C 26:301–313
8. Calabrese F, Ferrari L, Blondel VD (2015) Urban sensing using mobile phone network data: a survey of research. ACM Comput Surv 47(2):25
9. Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási AL (2008) Uncovering individual and collective human dynamics from mobile phone records. J Phys A Math Theor 41(22):224015
10. Diao M, Zhu Y, Ferreira J Jr, Ratti C (2016) Inferring individual daily activities from mobile phone traces: a Boston example. Environ Plann B Plann Des 43(5):920–940
11. Díaz-Uriarte R, De Andres SA (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7(1):3
12. Gao S, Janowicz K, Couclelis H (2017) Extracting urban functional regions from points of interest and human activities on location-based social networks. Trans GIS 21(3):446–467
13. Heiden U, Heldens W, Roessner S, Segl K, Esch T, Mueller A (2012) Urban structure type characterization using hyperspectral remote sensing and height information. Landsc Urban Plan 105(4):361–375
14. Ho TK (1995) Random decision forest. In: Proceedings of the 3rd International Conference on Document Analysis and recognition. IEEE, Montreal, pp 278–282
15. Hu T, Yang J, Li X, Gong P (2016) Mapping urban land use by using landsat images and open social data. Remote Sens 8(2):151
16. Isaacman S, Becker R, C'aceres R, Kobourov S (2011) Identifying important places in people's lives from cellular network data. In: International Conference on Pervasive Computing. Springer, pp 133–151
17. Khoroshevsky F, Lerner B (2017) Human mobility-pattern discovery and next-place prediction from GPS data. In: IAPR workshop on multimodal pattern recognition of social signals in human-computer-interaction. Springer, Berlin, pp 24–35

18. Liu Y, Liu X, Gao S, Gong L, Kang C, Zhi Y, Chi L, Shi L (2015) Social sensing: a new approach to understanding our socioeconomic environments. Ann Assoc Am Geogr 105(3):512–530
19. Liu X, Kang C, Gong L, Liu Y (2016) Incorporating spatial interaction patterns in classifying and understanding urban land use. Int J Geogr Inf Sci 30(2):334–350
20. Liu X, He J, Yao Y, Zhang J, Liang H, Wang H, Hong Y (2017) Classifying urban land use by integrating remote sensing and social media data. Int J Geogr Inf Sci 31(8):1675–1696
21. Lu D, Weng Q (2006) Use of impervious surface in urban land-use classification. Remote Sens Environ 102(1):146–160
22. Patel S, Kientz J, Hayes G, Bhat S, Abowd G (2006) Farther than you may think: an empirical investigation of the proximity of users to their mobile phones. In: International conference on ubiquitous computing. Springer, Orange County, pp 123–140
23. Pei T, Sobolevsky S, Ratti C, Shaw SL, Li T, Zhou C (2014) A new insight into land use classification based on aggregated mobile phone data. Int J Geogr Inf Sci 28(9):1988–2007
24. Poushter J (2016) Smartphone ownership and internet usage continues to climb in emerging economies. Pew Research Center 22:1–44
25. Shen Y, Karimi K (2016) Urban function connectivity: characterisation of functional urban streets with social media check-in data. Cities 55:9–21
26. Sheng C, Zheng Y, Hsu W, Lee ML, Xie X (2010) Answering top-k similar region queries. In: International conference on database systems for advanced applications. Springer, Berlin, Heidelberg, pp 186–201
27. Siła-Nowicka K, Vandrol J, Oshan T, Long JA, Demšar U, Fotheringham AS (2016) Analysis of human mobility patterns from GPS trajectories and contextual information. Int J Geogr Inf Sci 30(5):881–906
28. Theobald DM (2014) Development and applications of a comprehensive land use classification and map for the US. PLoS One 9(4): e94628
29. Toch E, Lerner B, Ben-Zion E, Ben-Gal I (2018) Analyzing large-scale human mobility data: a survey of machine learning methods and applications. Knowl Inf Syst:1–23
30. Toole JL, Ulm M, González MC, Bauer D (2012) Inferring land use from mobile phone activity. In: Proceedings of the ACM SIGKDD international workshop on urban computing. ACM, Beijing, pp 1–8
31. Trasarti R, Olteanu-Raimond AM, Nanni M, Couronné T, Furletti B, Giannotti F, Smoreda Z, Ziemlicki C (2015) Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. Telecommun Policy 39(3):347–362
32. Tu W, Cao J, Yue Y, Shaw SL, Zhou M, Wang Z, Chang X, Xu Y, Li Q (2017) Coupling mobile phone and social media data: a new approach to understanding urban functions and diurnal patterns. Int J Geogr Inf Sci 31(12):2331–2358
33. Wang H, Calabrese F, Di Lorenzo G, Ratti C (2010) Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: Intelligent transportation systems (ITSC), 2010 13th international IEEE conference. IEEE, Funchal, pp 318–323
34. Wen D, Huang X, Zhang L, Benediktsson JA (2016) A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation. Geosci Remote Sens 54(1):609–625
35. Wu C, Zhang L, Zhang L (2016) A scene change detection framework for multi-temporal very high resolution remote sensing images. Signal Process 124:184–197
36. Ye M, Yin P, Lee WC, Lee DL (2011) Exploiting geographical influence for collaborative point-of-interest recommendation. In: Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval. ACM, Beijing, pp 325–334
37. Yuan J, Zheng Y, Xie X (2012) Discovering regions of different functions in a city using human mobility and POIs. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, Beijing, pp 186–194
38. Zhao Z, Shaw SL, Xu Y, Lu F, Chen J, Yin L (2016) Understanding the bias of call detail records in human mobility research. Int J Geogr Inf Sci 30(9):1738–1762
39. Zheng Y, Capra L, Wolfson O, Yang H (2014) Urban computing: concepts, methodologies, and applications. ACM Trans Intell Syst Technol (TIST) 5(3):38

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.