

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

5,800

Open access books available

142,000

International authors and editors

180M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Mapping of Social Functions in a Smart City When Considering Sparse Knowledge

Oded Zinman and Boaz Lerner

Abstract

In recent years, technological advances, specifically new sensing and communication technologies, have brought new opportunities for a less expensive, dynamic, and more accurate mapping of social land use in cities. However, most research has featured complex methodologies that integrate several data resources or require much prior knowledge about the examined city. We offer a methodology that requires little prior knowledge and mainly relies on call detail records, which is an inexpensive available data resource of mobile phone signals. We introduce the Semi-supervised Self-labeled K-nearest neighbor (SSK) algorithm that combines distance-weighted k-nearest neighbors (DKNN) with a self-labeled iterative technique designed for training classifiers with only a small number of labeled samples. In each iteration, the samples (small land units) that we are most confident of their classification by DKNN are added to the training set of the next iteration. We perform neighbor smoothing to the land-use classification by considering feature-space neighbors as in the regular KNN but also geographical space neighbors, and thereby leverage the tendency of approximate land areas to share similar social land use. Based only on a few labeled examples, the SSK algorithm achieves a high accuracy rate, between 74% without neighbor smoothing, and 80% with it.

Keywords: call detail records, classification, computational social science, k-nearest neighbors, land use, machine learning, mobile phone data, smart cities, urban computing

1. Introduction

A city is a complex ecosystem and, as such, it is not the sum of its components; each component contributes but does not form the behavior of the whole [1]. The modern city is characterized by a sophisticated structure and zones of diverse urban social function, that is, residential neighborhoods, commercial areas, and industrial areas [2]. Functional city parts enable better orientation and support people's different needs [3, 4]. Rapid urban development has led to larger cities with more complex social dynamics, and this creates a great challenge for the accurate mapping of urban land use [5], for example, to promote social equity [6].

A smart city is a platform to facilitate technological and social innovation that enhances productivity, sustainability, and livability [7]. It opens the door for research designated for dynamic and automated identification of social function land use—understanding and classifying city lands of different social functions. Mapping of urban land use can be utilized for urban planning and designing of better urbanization strategies [8–10], urban air quality management [11], promotion of sustainable ecocities [12, 13], and green utilization efficiency of urban land [14]. Knowledge of the function of city parts and their management can help govern a city [15] and contribute to a better understanding of mobility patterns and interconnections between city parts, which is crucial for efficient planning decisions within cities, for example, planning of highways. Moreover, it can serve businesses looking for the right spot for their business, advertisers choosing a location for enhanced advertisement, and social recommendations [3].

The digital revolution has brought a great opportunity for social sciences research in cities; the emergence of enhanced computing power and mobile phones with built-in sensors and location technologies has created an enormous amount of data for understanding and monitoring urban life [16]. Data sources, such as remote sensing imagery, social media data, taxi trajectories, and mobile phone patterns of usage, have been utilized for cheaper and enhanced social land-use identification research.

Most research in recent years has offered complex methodologies that require the integration of several data resources of different types or substantial prior knowledge about the examined city. The motivation for conducting this research is to offer a method that requires only sparse knowledge of the examined land and relies on an inexpensive data resource. Previous works have yet to achieve high accuracy in such conditions; therefore, research and creative solutions are needed to solve this problem. Although incorporating several data resources can definitely improve the identification rate, in this work, we aim to achieve solid land-use mapping with a simple and efficient methodology that requires one data resource. Our main assumption is that sparse prior knowledge about the examined city's functional zones can be obtained by a local or domain expert at a low cost. We mainly rely on call detail records (CDR), an inexpensive and available data source routinely collected by telecom operators, and assume that areas of different social functions cause different typical cellular communication behavior [17]. For example, one can expect the communication pattern in a residential neighborhood to have different characteristics than that used for industry; perhaps at night and in the early morning, there will be more communication in a residential neighborhood. We utilize this behavior to identify different area categories with different functions.

This paper presents a semi-supervised algorithm, denoted as SSK (Semi-supervised Self-labeled K-nearest neighbor), which requires only sparse prior knowledge of the examined urban area, meaning it assumes we possess only a small number of land-use labeled areas. SSK combines both the distance-weighted k-nearest neighbor (DKNN) with a self-labeled iterative technique aimed to enlarge the training set in an iterative manner. We also perform a neighbor smoothing approach that offers a unique interpretation of neighbors in the context of the KNN process. In addition to considering feature-space neighbors as in the regular KNN, we also consider the geographical space neighbors, and thus we utilize the geographical homogeneity of social functions in urban areas.

The contributions of this work are as follows:

1. We offer a simple methodology that relies solely on one data resource (CDR). Previous works dedicated to this problem used more than one data resource and complex methodologies that integrate them.
2. We offer a method designated to perform in a condition of sparse prior knowledge about the social functions of the lands in the examined city. Most works assumed substantial prior knowledge about the examined lands, while others, such as Pei et al. [18], offered a semi-supervised method that requires relatively little knowledge about the examined city; however, the accuracy rate achieved in their work is yet not satisfactory.
3. SSK offers methodological innovations as it combines self-labeling techniques aimed at the condition of sparse knowledge and a fresh perspective on KNN—a KNN that considers not only the feature-space neighbors as in regular KNN but also the geographical space neighbors.
4. The presented methodology although relying only on few labeled samples and only one data resource, achieves a high accuracy rate, between 74% without neighbor smoothing, and 80% with it.

The rest of this paper is organized as follows: Section 2 presents recent developments and research on land-use mapping, Section 3 describes the methodology and SSK land-use classification algorithm, Section 4 evaluates the efficiency of SSK and compares its performance with other algorithms that require more prior knowledge about the examined area, Section 5 presents the neighbor smoothing integrated into SSK, Section 6 evaluates the usage of neighbor smoothing in SSK and discusses its merits and drawbacks, and Section 7 summarizes the work, presents conclusions, and offers directions for further research.

2. Related works

Several techniques have been developed for identifying social land-use functions. Traditionally, land-use identification was inferred by human trajectory patterns as reflected by individual travel surveys recorded by respondents [19–21]. However, self-reported diaries suffer from major disadvantages, including a relatively small number of respondents, difficulty in obtaining a representative sample of the city population, and an experimental period that is usually limited to a few days because of high costs. Moreover, the diaries are self-reported; therefore, they are not considered to be fully reliable.

Sensing technologies, ubiquitous connectivity, and computing power have brought a variety of opportunities for smart cities, and specifically to land-use mapping [22]. Data sources, such as remote-sensing imagery, social media data, taxi trajectories, and mobile phone signals, have been utilized for cheaper and enhanced social land-use mapping research.

Some works have used spectral and textural characteristics. For example, Lu and Weng [23] integrated population density data and remote-sensing systems measuring land surface temperature and spectral reflectance to classify urban lands. Image processing and classification techniques of remote-sensing images were used in numerous research studies to capture physical aspects, such as land surface reflectivity and texture of urban space [24–26] or to accomplish urban land-use mapping [9]. However, inferring land use by analyzing remote-sensing images tells only part of the story because they cannot recognize functional interactions between city segments and social behavior [27–29].

Social media can be seen as complementary to remote-sensing image methodology, as it is valuable for identifying movement patterns and social dynamics [27, 29, 30]. A varied collection of social media data, such as social media check-ins, GPS trajectories, and points of interest (POI), has been used for monitoring urban residents' land-use dynamics [31]. Liu et al. [32] offered an unsupervised method that extracts patterns of temporal activity variations and spatial interactions between places based on taxi trajectories and discovers the common characteristics of lands of similar social function. Long and Thill [33] combined one-week period bus smart card data and household travel survey to analyze jobs–housing relationships in Beijing. Commuting trips from three typical residential communities to six main business zones were mapped and compared to analyze commuting patterns in Beijing, and then validated with those extracted from the survey. Also, Zhou et al. [34] used smart card data. They investigated how a rider allocates time in the vicinity of metro stations spatially and temporally to classify space–time activity patterns that may explain inter-personal and intra-personal behavioral variability. Shen and Karimi [35] used check-in-based data and analyzed the interaction between places in the city to infer their urban structure and related socioeconomic patterns. POIs associated with coordinates and a label such as “restaurant,” “shopping center,” and “theater” have been extensively leveraged for land-use identification [36]. Their biggest virtue is that they carry semantic information. Some methodologies offer to leverage POI datasets to discover regions of similar social function by classifying together lands of similar POI types' distribution and patterns [27, 37]. However, social media data's main demerit is its sparsity in space and time [29]. Social information hidden in GPS records allowed Khoroshevsky and Lerner [38] to discover mobility patterns and predict users' geographic and semantic locations alike, with no privacy violation by using only the user's own data and no semantic data voluntarily shared by him or by others. By properly selecting an evaluation metric of trajectory clustering and accounting for cluster density, they traded between prediction accuracy and information, providing more clusters that are smaller and denser, showing more meaningful locations, but less predictable, and vice versa. Using semantic mobility patterns determined from POIs in people's daily trajectories, Ben Zion and Lerner [39] could identify and predict person's lifestyle both for a novel trajectory and a novel user.

As all data sources are limited and capture specific aspects of urban dynamics, a recent movement in the research of land-use identification is to rely on several data sources of different types. Both the works of Liu et al. [31] and Hu et al. [8] combined remote-sensing images and social media data. The work of Yuan et al. [3] integrated POI datasets and datasets of 3 months of GPS trajectories generated by 12,000 taxis in Beijing to identify lands of different social functions using an unsupervised clustering algorithm. The work of Tu et al. [29] integrates a mobile phone signals dataset with social media data to infer the social function of land use. They estimated individuals' “home” and “work,” and then aggregated the individuals

together with social knowledge learned from social media check-in data into a collective social land-use map.

Numerous works leverage call detail records (CDR) for capturing spatiotemporal movement patterns and city dynamics [17, 30, 40]. CDR holds data of mobile phone signals collected and stored by telecom operators mainly for billing reasons [41]. They contain communication properties, such as start time and call duration, type of communication (call, SMS, internet), as well as the cell tower from which the communication originated. CDR also includes the location at which the communication occurred, calculated by triangulating the signal strengths from surrounding cell towers [4, 41, 42]. Its greatest virtue, as a location tool for human behavior evaluation, is that it is routinely produced by the telecom equipment when users make a phone call, send or receive a message, or browse web pages; hence, it is a low-cost and efficient location estimation source [43]. The respondents in an experiment are unaware of it, and are, thus, not interrupted by it, but still, their personal information is not violated, as the actual user identification is ciphered. CDR contains an enormous amount of data and covers the major part of civilized areas in the world, depicting a variety of users. However, CDRs have two prominent limitations as a source for tracking human activity: First, they are sparse in time because they are generated only when a user engages in cellular communication. Second, they are coarse in space because they record location only at the granularity of a cell tower [30, 44]; CDR-rendered coordinates have a varied inaccuracy of 50–350 m, depending on the density and arrangement of the towers. Another shortcoming is their lack of semantic information [30, 45].

Although incorporating several data resources is beneficial for achieving a high accuracy rate [9], in this work, we focused on achieving solid land-use identification with a simple and efficient methodology that requires only one data resource and little prior knowledge that can be obtained by domain experts. We wish to extract the most out of the information embodied in CDRs, and it can also be integrated with additional resources in future works. Several other works have already used CDRs as their main data resource for land-use identification. Toole et al. [40] utilized them for a supervised land-use classification method with a dataset consisting of CDRs for a period of three weeks in the greater Boston area. They classified urban space into five categories—residential, commercial, industrial, parks, and other, and relayed possession of ground truth land use as obtained by a zoning map. For the classification, they used Breiman's [46] random forest classification algorithm and post-processed the classification results with a neighbor smoothing algorithm. However, even with smoothing performed, in classifying the five land-use classes, the accuracy was relatively low, 54%. Pei et al. [18] also relied on CDRs and offered a semi-supervised algorithm for classifying the land of Singapore into the same five classes as Toole et al. [42]. They relied on the classification of a small number of labeled places, choosing 200 places to be labeled based on a few criteria aimed to ensure reliable labeling, and labeled them based on Singapore locals and Google Earth. They used the fuzzy c-means algorithm [47] and assumed possession of the “real” land-use labels of a small number of area segments. Their results also showed a relatively low detection rate of 58%. Zinman and Lerner [17] divided the space and time into spatiotemporal units, derived a varied collection of features to illuminate the social behavior of the units, and classified, with accuracies ranging from 84% to 91%, units in 62 days of cellular data recorded in nine cities in the Tel Aviv district according to their land use using a leveled hierarchy of semantic categories that include different levels of detail resolution.

3. SSK methodology

Our dataset consists of CDRs recorded by an Israeli telecommunications 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, including the city of Tel Aviv and its neighboring cities. The data include a diverse collection of human activity—a variety of settlements (cities and villages), open areas, highways, and industrial areas.

Our workflow (**Figure 1**) can be divided into five steps: (3.1) Area selection, (3.2) Division of smaller units of land with grid-like partitioning, (3.3) Land-use labeling, (3.4) Feature extraction, and (3.5) Usage of the SSK algorithm for land-use classification.

3.1 Area selection

We selected 61 areas with varied and known social functions, such as neighborhoods, industrial areas, office areas, highways, commercial streets, and shopping malls spread over nine cities, all located in the Tel Aviv metropolitan and its surrounding area (**Figure 2**): Tel Aviv, Holon, Ramat Gan, Petah Tikva, Rosh Haayin, Ra'anana, Ramat Hasharon, Givatayim, and Kfar Saba.

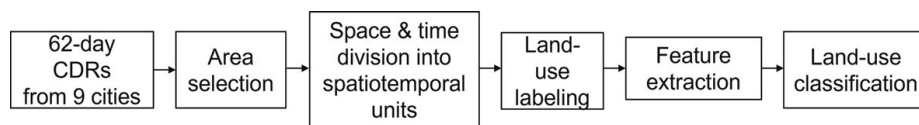


Figure 1.
Workflow of land-use classification using the SSK algorithm.

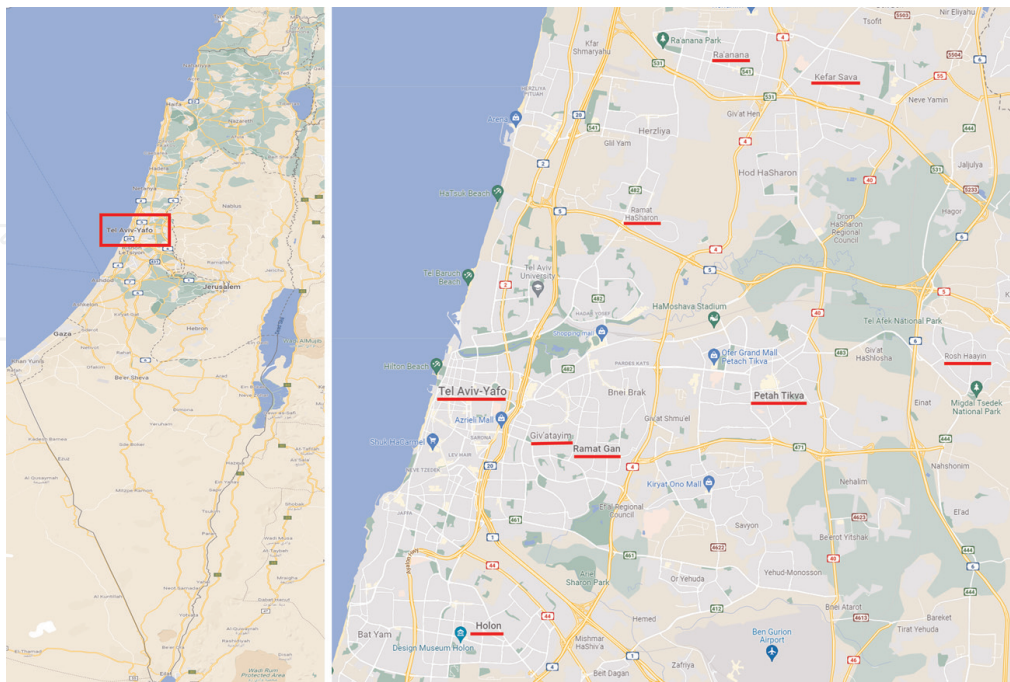


Figure 2.
(Left) A map of Israel with the area covered in the study marked by a red rectangular. (Right) A zoom-in map of this area including the Tel-Aviv metropolitan and its surrounding area with nine cities participating in the study (underlined in red): Tel Aviv, Holon, Ramat Gan, Petah Tikva, Rosh Haayin, Ra'anana, Ramat Hasharon, Givatayim, and Kfar Saba. Approximately 1 million people are living in this area.

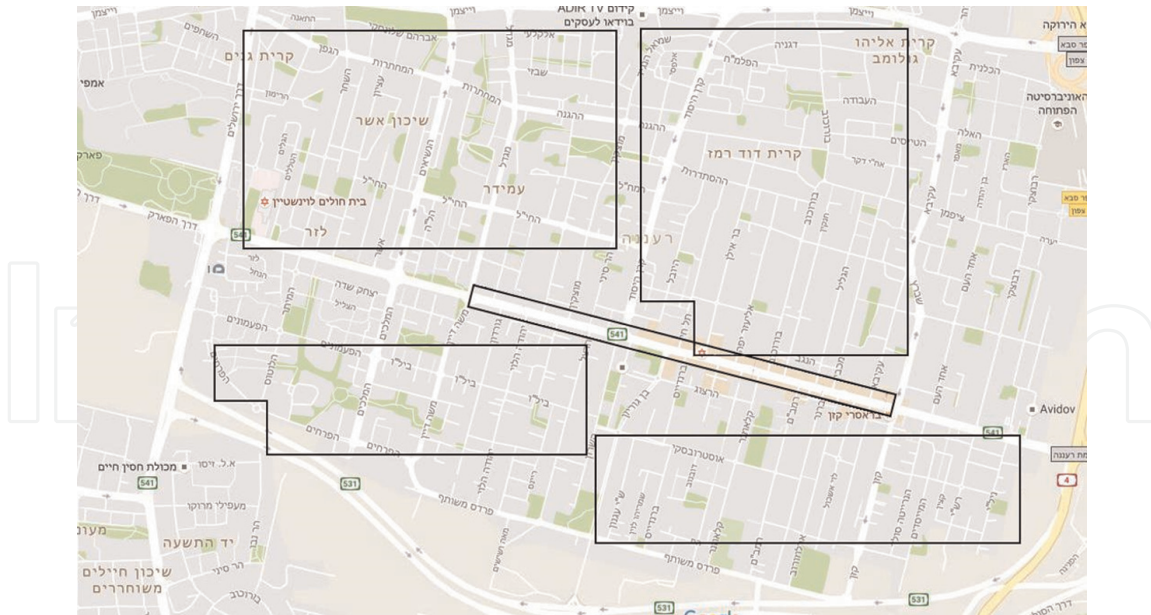


Figure 3.
Areas selected in the city of Ra'anana.

For example, see the five selected areas in the city of Ra'anana shown in **Figure 3**. Each area is represented as a polygon on the map. Four of the areas are wide; these cover residential neighborhoods. There is one narrow rectangle representing Ahuza Street, the main commercial street. It is narrow to include only the street without the surrounding area.

There is a need to discuss the choice to analyze segments of several cities that were deliberately chosen. Previous works, such as the works of Yuan et al. [3],

Toole et al. [42], and Sun et al. [9], performed land-use mapping of whole cities, Beijing, Boston, and Shenzhen, respectively. However, in intentionally chosen areas, the social functions are less mixed. Classifying land of “pure” social function is easier; hence, we expect implementing this method on a whole city to yield a lower accuracy than achieved on this dataset. However, the deliberate choice of areas also has some notable advantages. Analyzing social land uses in their “pure” condition enables us to recognize the core behavior and patterns of the social functions. The areas chosen from different cities enable the examination of the inter-cities’ resemblance of social function, as reflected by the use of cellular communication. Deliberately choosing areas causes the labeling process to be less expensive and time-consuming. More importantly, the granted labels are more accurate, and it enables more reliable tests and conclusions. Thus, this dataset enables a careful analysis, which is valuable for the assessment of the feasibility of the method.

3.2 Division of time and space into basic spatiotemporal units

We divided space and time into spatiotemporal units. The chosen areas were divided into smaller geographical units in a grid-like manner; we refer to each unit as a cell. Dividing the land into smaller parts reduces the variety of the social functions that take place in each; therefore, there is more homogenous land use, which is more suitable for land-use categorizing. However, using small fixed-size land parts may lose accuracy when the use of space is dynamic due to a mix of buildings of different uses in close proximity, or even different uses in the same building on different floors. We

further note that others [48] found hexagonal cells advantageous over square cells, although the former are less intuitive for the urban environment, or used census blocks, where each partitioning system has its advantages and disadvantages [49]. We preliminarily found the square grid suitable for our needs and selected the default size of the cell as $40,000 \text{ m}^2$, shaped as a $200 \times 200 \text{ m}^2$. This is the same cell size and shape specified by Toole et al. [42] and Pei et al. [18]. However, because 30 of the 61 areas contained an edge smaller than 100 m, in these areas, we used narrower rectangles.

Land use is dynamic and varies during the day. For example, activity habits in a residential neighborhood at 7 p.m. (say, eating dinner and watching TV) are greatly different than the activity habits in the same neighborhood at 3 a.m. (say, sleeping). Therefore, in addition to dividing space, we also divided the day hourly, that is, 00:00 a.m. to 01:00 a.m. is one time unit, 01:00 a.m. to 02:00 a.m. is another time unit, and so on.

3.3 Land-use labeling

We labeled each cell per hour with a semantic social function of land use. As mentioned above, we chose to focus on areas that were relatively easy to label and, hence, we could label them with the help of a few locals. The labeled areas were then used as ground truth for training the land use classifier and evaluating its accuracy.

The semantic land-use labels include Residential, Commercial, Industrial, Highway (arterial roads), Office, Street, and No activity (no human activity is expected in this cell at this specific time, e.g., in industrial areas before work hours begin).

3.4 Feature extraction

In this work, we used 158 features that include varied aspects of the circadian nature of the activity in the cell [17]. We divided the features into five types: (1) Communication volume features measure the degree of communication activity. These features are designated to capture the difference between the activity volume typical to a specific social function (e.g., in commercial zones, there is more cellular communication compared to residential areas). (2) Daily pattern features 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 that area (e.g., in a residential area, the communication peak hours are in the mornings and evenings, while in industrial areas, the peak is during working hours). (3) Weekly pattern features capture the difference in cellular usage on weekdays compared to the weekend. Thus, it differentiates between land uses, such as residential, where their inhabitants return daily, and those like office zones, where workers do not go on weekends. (4) Contact features measure the number of different days on which people engage in at least one cellular communication in cell s in hour h , thus, differentiating between land uses with frequent visitors and those with occasional ones. (5) Communication habits features are a collection of features that aim to illustrate the land from the perspective of typical cellular communication usage habits, for example, call duration and usage distribution of different types of cellular communications (phone calls and internet usage). These 158 features were found very successful in land-use classification [17]. They predicted residential, industrial, and no activity land uses with F1 (see Eq. (5) below) values higher than 0.9 and provided average accuracy over seven land uses between 81% and 90% at any time of the day.

3.5 Semi-supervised self-labeled k-nearest neighbor

We developed a variation of the k-nearest neighbor algorithm combined with a self-labeled iterative technique that enlarges a labeled dataset when only a few labeled samples exist. We call this method the Semi-supervised Self-labeled K-nearest neighbor (SSK).

Gathering land-use labels of a few segments of an urban area is relatively attainable. This information can be gathered by inquiring locals. However, getting additional land-use labels is often out of reach or too expensive. In a condition of only a small number of labeled samples, the effectiveness of conservative supervised classification algorithms deteriorates. Therefore, we used the self-labeled technique designated to generate more labeled samples as an input for the classifier to tackle the lack of labeled data [50]. The self-labeled technique follows an iterative procedure—in each iteration, unlabeled data is labeled and added to the training set for the next iterations. In the first iteration, a classifier is trained based only on the labeled samples and classifies the unlabeled samples. In every iteration, the samples that the algorithm is most confident of classifying correctly are added to the labeled sample pool.

In our implementation, we used the Distance weighted variation of K-Nearest Neighbor (DKNN) as the classifier. We assumed possessing the “real” land use label of 5% of the samples. In every iteration, 5% of the samples, which the DKNN classifier is most confident of, are added to the training set. The samples used in the classification are the basic spatiotemporal unit described in Section 3.2, which we refer to as cell. We use x_i to refer to the cell i 's sample.

We used DKNN, as introduced by Dudani [51]. In the classic version of KNN, assigning a class to each query sample (unlabeled sample) is determined by its k nearest neighbors in the training set, and each of the k neighbors has the same impact. In the distance-weighted version, again the k -nearest neighbors contribute to the classification of the query sample, but here, the closer the sample is to the query sample, the more impact on the classification it has. Each of the k neighbors of the query sample x_q 's gets a weight $w_q^{(i)}$ that depends on how close it is to the query sample:

$$w_q^{(i)} = \frac{1}{d(x_q, x_i)^2} \quad \forall i \in 1, \dots, k, \quad (1)$$

where $d(x_q, x_i)$ is the feature-space Euclidean distance between the query sample x_q and its labeled neighbor x_i , and other distance-weighted versions may be considered as well, for example, the harmonic mean distance [52]. k determines the number of neighbors considered in the calculation. Since training the DKNN does not exist (all computation is done during prediction), the classifier training time and space complexities are $O(1)$, and the prediction time complexity is $O(knd)$ for n d -dimensional samples (and the prediction space complexity is also $O(1)$). Setting the number of neighbors k and a discussion about the considerations leading to its choice will follow below.

For example, let us assume that $k = 2$ and that x_q 's two closest neighbors (labeled samples closest in the feature space to x_q) are x_a and x_b , and that their feature-space distance from x_q are 2 and 3, respectively. Then, according to Eq. (1), the weight of X_a is $\frac{1}{4}$ and that of X_b is $\frac{1}{9}$, as x_a is closer to x_q .

The SSK algorithm demonstrated in this section comprises the self-labeled technique and the DKNN classification algorithm. However, we have made some adjustments to make a version of DKNN that is more suitable for our problem. In regular classification, the labels used for training are assumed to be correct. However, this assumption cannot be taken when using the self-labeled technique because only the labels in the first iteration are ground truth labels, and the labels in the next iterations are samples that were not labeled but have been classified through the process. To address this issue, we would like neighbors whose label we are more confident is correct to have more impact on the classification.

Let O be the set of all cells (samples) and L be the original set of predefined ground truth labeled cells. The set of cells that are currently labeled in a certain iteration is G , and its complement set of cells that are not yet labeled Q ($Q = O \setminus G$). In the first iteration of the algorithm, $G = L$. When describing the process, we will refer to the cell that its class is being considered as the query cell.

We would like to introduce the term land-use array, which is an object that we use to discuss the method. The number of entries in a land-use array is equal to the number of land uses. We denote the land-use array of x_i as A_i . Each array entry in A_i represents a land use, for example, entry 1 would be Residential, entry 2 Commercial, etc. The value of entry j represents the certainty that cell x_i is attributed to class j . Consider $A_i = (v_1, v_2, \dots, v_c)$. v_i is a value that represents the confidence we have that the land use of cell x_i is i . The sum of all entries in A_i is always 1. c is the number of land-use categories.

In the first iteration of the algorithm, the classification of the unlabeled cells is determined using the predefined labeled cells L , of which we assume 100% confidence. Before the first iteration, we initialize the land-use arrays of all the cells in L . Let us denote the land-use classes of the cells in L as C , meaning that the label of $x_i \in L$ is C_i . The initialization of the land-use array of cell $x_i \in L$ follows—entry number C_i (the class of x_i) in A_i is set to 1, and all the other entries are set to 0. For example, if cell x_i is labeled as Commercial, and we assume that Commercial is represented in the second entry, then its land-use array $A_i = (0, 1, 0, \dots, 0)$.

Setting the land-use arrays of the yet unlabeled cells is computed by the land-use arrays that were already calculated. Thus, the computation of the land-use array A_q for a query cell x_q is given by

$$A_q = \frac{\sum_{i=1}^k w_q^{(i)} A_i}{\sum_{i=1}^k w_q^{(i)}} \quad q \in Q, \quad (2)$$

where k is the number of neighbors configured for x_q , and $w_q^{(i)}$ is set by Eq. (1).

In the first iteration, the calculation of the land-use arrays is based only on the land-use arrays of the cells in L . At the end of the first iteration, the land-use arrays of the cells that were selected to be added to training set G of the next iteration will be set according to (2), and they will be used for the calculation of land-use arrays in the next iterations, and the process repeats itself in the next iterations.

For example, we will examine an hour with four land-use classes. For simplicity, let us assume that $k = 2$, meaning that for computing the land-use array A_q , we will consider only the two neighbors closest in the feature space. The two nearest neighbors of the query cell x_q are x_i and x_j . x_i is labeled as class 2 and x_j is labeled as class 4; therefore, their land-use arrays are $A_i = (0, 1, 0, 0)$ and $A_j = (0, 0, 0, 1)$. Their

weights are $w_i = 6$ and $w_j = 2$. Notice, the weights indicate that x_i is closer to x_q than x_j . Calculating A_q :

$$A_q = \frac{w_q^{(i)} A_i + w_q^{(j)} A_j}{w_q^{(i)} + w_q^{(j)}} = \frac{6(0, 1, 0, 0) + 2(0, 0, 0, 1)}{6 + 2} = \left(0, \frac{3}{4}, 0, \frac{1}{4}\right). \quad (3)$$

A_q is calculated by the weighted average of the land-use arrays of its feature-space neighbors. For example, the value of the fourth entry in A_q ($\frac{1}{4}$), which represents the fourth land use, is the result of a weighted average of the fourth entry in A_i (equals 0) and A_j (equals 1), and it is calculated by $\frac{6 \cdot 0 + 2 \cdot 1}{6 + 2} = \frac{1}{4}$. The weighted average value $\frac{1}{4}$ is closer to A_i (equals 0) than to A_j (equals 1) because x_q is closer to x_i . Notice that (2) guarantees the land-use array entries always sum up to 1. In the example, the highest entry value is $\frac{3}{4}$, and its corresponding land-use class is 2; therefore, it is most reasonable to assign q to class 2. If x_q will be added to G at the end of the iteration, then A_q will be used to calculate land-use arrays in the next iterations.

However, we will classify x_q to class 2 only if it has high enough classification confidence, meaning only if we have relatively high confidence that its attribution is correct, we classify it and add it to the training set of the next iteration. The classification confidence of x_q is estimated by the entry with the maximal value in the land-use array:

$$\text{confidence}_q = \max(A_q). \quad (4)$$

In the example, the classification confidence level of x_q is $\frac{3}{4}$ of it being attributed to class 2. In the example, x_q is a candidate for being classified as class 2, and it will be classified as class 2 if the confidence level $\frac{3}{4}$ is high enough.

In each iteration, we add 5% of all the cells to the training set for the next iteration. To consider a proper balance between the labels in the training set over the iterations, we do not blindly add to the training set the top 5% of the samples with the highest classification confidence. The number of cells added to the training set is proportional to the number of candidates for each land use in this iteration. For example, consider a simple case with only two land-use classes. Let us assume that the number of cells $|O| = 1000$, and therefore the number of cells added to the training set G in each iteration is 50 (5% of 1000). If in a specific iteration, 60% of the cells (600 cells) are candidates for class 1 (i.e., in 60% of the cells, the highest entry in the land-use array is 1), and the other 40% (400 cells) are candidates for class 2, then accordingly, 60% (30) of the cells added to the training set will be from class 1 and 40% (20) of the cells from class 2. The cells with the highest confidence are added to each class separately. In this example, the 30 cells with the highest values in entry 1 (represent class 1) will be labeled accordingly and added to the training set of the next iteration.

We would like to demonstrate in **Figure 4** the process of land-use classification using SSK with an example. We demonstrate classifying a query cell x_q to land use in the first iteration (**Figure 4(top)**), and then we demonstrate classifying another query cell x_s in the second iteration (**Figure 4(bottom)**). The bars in **Figure 4** represent the values of each entry in the land-use arrays. In the example, for simplicity, the neighborhood parameter $k = 2$, that is, the classification is based on the two samples that are closest to the query cell in the feature space. In this example, there are

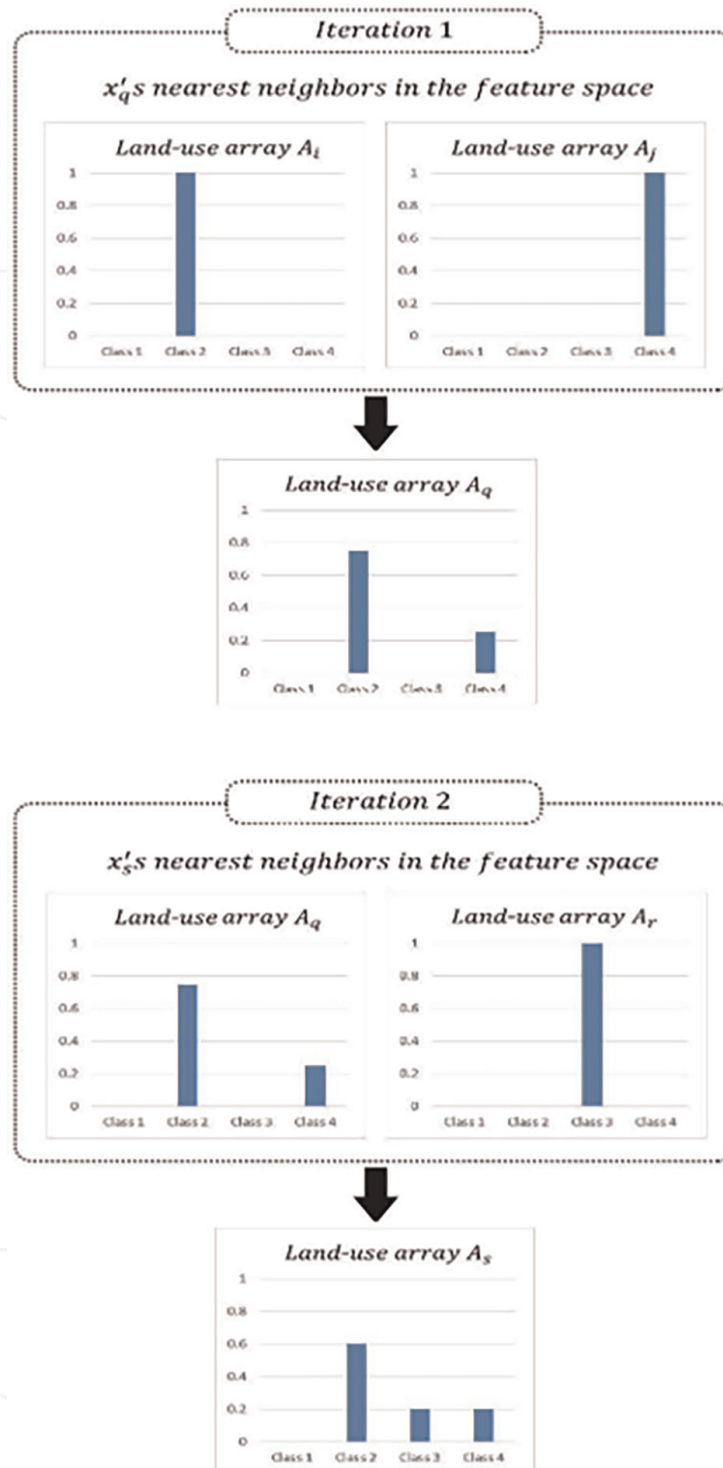


Figure 4. Computing land-use arrays for (top) first and (bottom) second iterations of an example.

four land use classes. The computation of the land-use array of A_q in the first iteration (**Figure 4(top)**) was already demonstrated in the previous examples. We saw that after considering x_q 's two nearest neighbors x_i and x_j , and based on their land-use arrays A_i and A_j , then $A_q = (0, \frac{3}{4}, 0, \frac{1}{4})$ and $confidence_q = \frac{3}{4}$. Let us further assume that this confidence level of x_q was high enough, and thus x_q was labeled by class 2 and added to the training set for the second iteration.

In the second iteration (**Figure 4(bottom)**), there is another query cell x_s . In the example, x_s 's two nearest neighbors are x_q (the cell that was added to the training set

in iteration 1) and another cell x_r , and their land-use arrays are $A_q = (0, \frac{3}{4}, 0, \frac{1}{4})$ (as already computed) and $A_r = (0, 0, 1, 0)$ and weights are $w_q = 4$ and $w_r = 1$, respectively. The land-use array of query cell A_s (Eq. (2)) is:

$$A_s = \frac{w_s^{(q)} A_q + w_s^{(r)} A_r}{w_s^{(q)} + w_s^{(r)}} = \frac{4(0, \frac{3}{4}, 0, \frac{1}{4}) + 1(0, 0, 1, 0)}{4 + 1} \left(0, \frac{3}{5}, \frac{1}{5}, \frac{1}{5}\right). \quad (5)$$

Figure 4(bottom) demonstrates that the land-use array A_s of cell x_s is mainly affected by cell x_q (belonging to class 2), which was labeled and introduced into the training set only in the previous iteration.

There is a need to specify the neighborhood parameter k that specifies the number of cells considered in the classification of each query cell. k controls the volume of the neighborhood and, consequently, the smoothness of the density estimates; thus, it plays an important role in the performance of the nearest neighbor classifier [53]. Increasing k decreases variance and increases bias; conversely, decreasing k increases variance and decreases bias [54]. Since the number of labeled cells gradually increases during the process of the self-labeled technique, we offer a dynamic k that changes through the iterations; its value depends on the size of $|G|$ —the number of cells currently in the training set G . Through the iterations, k grows with the set of cells (samples) available for training. We used a rule-of-thumb offered by Duda et al. [55], setting the k value by:

$$k \approx \sqrt{|G|}. \quad (6)$$

For example, if the number of labeled cells $|G|$ in the first iteration is 50, then in the first iteration, $k = \sqrt{50} = 7.07 \approx 7$, and therefore the closest seven neighbors of each query cell will be considered in the classification. By the next iteration, 50 cells are added to G , then $|G| = 100$ and $k = \sqrt{100} = 10$, thus 10 neighbors will be considered next.

4. Empirical evaluation of SSK

In this section, we evaluate the performance of SSK classification. We compare it to the results of a classifier that possesses significantly more prior knowledge, demonstrate its performance with a few examples from different cities in Israel, analyze the process of the self-labeled technique, and discuss its overall accuracy and the accuracy in each land use separately.

We used the ground truth land-use labels for two purposes—for training the SSK classifier and for evaluating its performance. Five percent of the cells were randomly chosen at the beginning of the process, and the labels of these cells were treated as ground truth and were used for training the classifier. The performance of the classifier was estimated by the labels of the other 95% of the cells. We performed the classification in each hour separately, and in each hour, repeated the process five times, each with another randomly chosen 5% of the cells. Thus, using these permutations, we diminished the variance caused by the random aspect.

The accuracy rate of SSK averaged over all permutations and hours using labels for only 5% of the cells is 74.4%. Compared to the works of Toole et al. [42] and Pei et al. [18] who also attempted to identify land use based on CDR, our accuracy rate is

exceptionally high; Toole et al. [42] and Pei et al. [18] achieved 54% and 58% accuracy rates, respectively. However, it is not possible to make conclusions based on comparing the accuracy rates of these works. The main reason is that these studies performed land-use mapping of a whole city, Boston in the work of Toole et al. [42], and Singapore in the work of Pei et al. [18], whereas we deliberately chose areas with a relatively “pure” and clear land-use function from different cities in Israel. Identification of the land use in lands of “pure” social function is an easier process.

Tables 1 and **2** illustrate the classification results in greater detail and the quality of the classification of each land-use category separately. **Table 1** demonstrates the confusion matrices of the results—predicted (columns) vs. true values (rows)—in different day parts: (a) between 4 a.m. and 7 a.m., (b) between 8 a.m. and 5 p.m., (c) between 5 p.m. and 7 p.m., and (d) between 8 p.m. and 10 p.m. Notice the set of social

	Residential	Street	Highway	No activity	
(a) 4 a.m.–7 a.m					
Residential	46.27	1.43	2.27	0.50	
Street	8.40	4.00	1.03	0.60	
Highway	3.13	0.67	1.03	0.63	
No activity	10.57	3.40	3.10	12.90	
	Residential	Commercial	Industrial	Office	
(b) 8 a.m.–5 p.m.					
Residential	44.71	1.33	0.29	0.12	
Commercial	9.99	10.30	1.13	0.12	
Industrial	4.99	2.27	22.42	0.28	
Office	0.66	0.14	0.62	0.62	
	Residential	Commercial	Office	No activity	
(c) 5 p.m.–7 p.m					
Residential	38.50	8.15	0.10	0.10	
Commercial	6.55	14.85	0.10	0.35	
Office	0.65	0.55	0.50	0.45	
No activity	4.55	6.15	1.55	17.00	
	Residential	Street	Highway	Commercial	No activity
(d) 8 p.m.–10 p.m					
Residential	41.80	2.40	2.15	0.05	0.85
Street	2.95	0.70	0.45	0.15	0.25
Highway	2.80	0.20	1.00	0.10	1.00
Commercial	2.05	0.30	1.35	7.20	1.90
No activity	5.65	0.55	2.85	0.80	20.70

Rows—true values; columns—predicted values. All values in %.

Table 1. Confusion matrices of the classification results in four day parts: (a) 4 a.m.–7 a.m., (b) 8 a.m.–5 p.m., (c) 5 p.m.–7 p.m., and (d) 8 p.m.–10 p.m.

Land uses	Precision	Recall	F1
Residential	0.73	0.92	0.82
Commercial	0.70	0.52	0.59
Industrial	0.91	0.74	0.82
Office	0.46	0.28	0.35
Highway	0.25	0.19	0.21
Street	0.30	0.20	0.24
No activity	0.82	0.52	0.64

Table 2.
 Precision, recall, and F1 of each land use.

land uses changes throughout the day. Some of the social functions, such as Commercial, occur only in specific hours (**Table 1b–d**), 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. While **Table 1** provides detailed accuracies for the different land uses in different time parts of the day, **Table 2** averages performance over the land uses and time parts and illustrates the precision, recall, and F1 score for the classification of each land use over all cells in the nine cities. Precision is the percentage of cells correctly classified to specific land use c , recall is the percentage of cells of the specific land use that are classified correctly, and the F1 score considers both recall and precision by calculating their harmonic average

$$F1 = 2 \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

Thus, we use the F1 score as the best indicator for the quality of classification of certain land use.

Residential and Industrial are well identified (both have an F1 score of 0.82). Residential 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. All the land-use categories except Residential have higher precision than recall. It indicates that the classifier tends to classify as Residential, and all the other land uses are under-classified. Residential has a high Recall (0.92) and lower precision, while Industrial has high Precision (0.91) and lower recall. Commercial is relatively well-identified (F1 is 0.59). The commerce identification rate is damaged by the inaccuracy of location estimation more than other land uses. As mentioned in Section 2, CDR-rendered coordinate location estimation is inaccurate and can reach 350 m. Commercial streets, because of their long and narrow shape, are vulnerable to location estimation mistakes. Because they are often surrounded by a “sea” of residential neighborhoods, transmissions originating from the neighborhoods are mixed with transmissions originating from the commerce street. The result is a mixed cellular communication behavior that makes correct identification harder. Indeed, Commercial is often confused with Residential, as is shown in **Table 1b** and **c**. Later in the paper, we demonstrate an example of a Commercial street in the city of Ra’anana that is confused with its neighboring residential buildings. The same problem occurs in other narrow-shaped land uses, such as streets and highways; both have a low identification rate.

Street is also frequently confused with Residential (see **Table 1a** and **d**), rather not surprisingly because they are located in the heart of neighborhoods. No activity is relatively well-identified (F1 is 0.64).

We compared SSK performance that assumes possession of the social function of only 5% of the cells to a supervised random forest (RF) [46] classifier that assumes significantly more labeled cells. The RF classifier was trained on the same dataset and the same areas, except that it was trained with 8-fold cross-validation, thus in each fold, RF classified 1/8 of the cells based on the other 7/8 cells. Meaning, that compared to SSK, which assumed possession of 5% of the cells, RF assumed possession of 87.5% (7/8) of the cell. As expected, RF did achieve a higher accuracy rate of 84%; however, the accuracy rate of SSK (74.4%) is considerably high, considering the lack of labeled samples.

In **Figure 5**, we visualize the results on a map we refer to as a geographical 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. **Figure 5** compares the geographical confusion maps of RF (**Figure 5a–c**) and SSK (**Figure 5d–f**) classification on the work hours between 8 a.m. and 5 p.m. in three cities: Ra’anana (RF **Figure 5a** and SSK **Figure 5d**), Ramat-Gan (RF **Figure 5b** and SSK **Figure 5e**), and Tel Aviv (RF **Figure 5c** and SSK **Figure 5f**). The legend displays the 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 that lies to its side. The color of each of the cells indicates the land use it is classified to. Notice, some of the cells have more than one color. This is because the results in these maps accumulate 45 classification results, 9 hours from 8 a.m. to 5 p.m. X 5 random training–testing permutations.

Figure 6(left) focuses on part of Ramat-Gan’s RF classification results (**Figure 5b**). See the cell marked “1”; it has three colors: blue, yellow, and a thin line of red. Fifty-three percent of the cell is blue, indicating it was classified as Residential in 53% of the runs (24 of the 45 runs). Also, almost half of the cell is yellow, indicating that it was frequently classified as Industrial, and it includes a thin red line that indicates it was also classified as Commercial (in 2 of the 45 runs). In contrast, the cell marked “2” is completely yellow, indicating that it was classified as Industrial in all runs.

Comparing the visualized results, one can see that SSK, which relies on a small number of labeled cells, suffers from higher classification variance than RF. In SSK, more cells are not unanimously classified to the same cell in all 45 runs, as indicated by more cells containing more than one color. For example, in **Figure 5c**, most of the cells of the commercial streets Ibn Gabirol and Dizengoff in Tel-Aviv classified by RF are uniformly red. This indicates that they were classified as Commercial in all runs.

However, the same streets classified by SSK (**Figure 5f**) are mostly red, indicating that in most runs, they are correctly classified as Commercial, but blue is also prominent, indicating that in a non-negligible number of the runs, they were classified as Residential (note, however, that in both streets, the ground floor of the buildings is stores and restaurants, that is, should be labeled Commercial, but the remaining, usually three, floors are residential, and thus should be labeled as Residential). SSK heavily relies on a random selection of the 5% cells used in the initial training set, in contrast to RF that relies on a large and consistent training set. Raanana’s commerce street, Ahuza St. (**Figure 5a** and **d**), is confused with Residential. This is mostly because of the location estimation inaccuracy described earlier in this section, as the street is surrounded by neighborhoods and, hence, receives cellular transmissions of the neighboring Residential land use and is thereby confused with Residential.



Figure 5. Geographical confusion map comparison of RF (a)–(c) and SSK (d)–(f) for three cities shown in **Figure 2** (bottom): (a) and (d) Ra'anana, (b) and (e) Ramat Gan, and (c) and (f) Tel Aviv.

Moreover, this geographical confusion may be caused by residential buildings on the street itself that mix the social use of the land (as in the two streets in Tel Aviv).

SSK classification is more biased. As an example, we will examine the results of the commercial streets marked with a red circle beside them in Ramat-Gan (**Figure 5b** and **e**). Both algorithms classified the commercial streets inconsistently, sporadically classifying them as Commercial (correct) or as Residential (incorrect), but RF correctly classified the cells in most runs as Commercial (most cells are mostly red), whereas SSK classified some of the Commercial cells more as Residential (cells that are mostly blue).

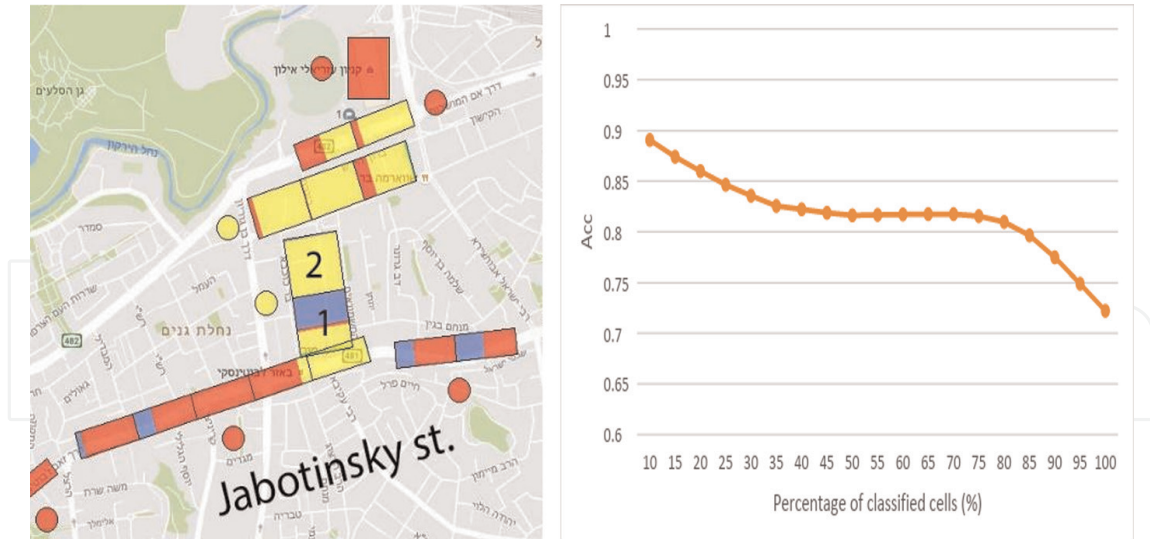


Figure 6. (left) “Zoom in” on part of Ramat Gan’s geographical confusion map of the RF classification results (**Figure 5b**). (right) Accuracy rate (Acc) vs. the percentage of classified cells added in the self-labeled process.

The accuracy of SSK is different across the different streets. Dizengoff St. (**Figure 5f**) for example, is correctly classified as Commercial in most runs. Another Commercial street in Tel Aviv, Ibn Gabirol St. (**Figure 5f**), is correctly classified at a lower rate than Dizengoff, while Jabotinsky St. in Ramat-Gan (**Figure 5e**) is mostly classified as Residential instead of Commercial. Analyzing the three streets indicates that they have different characteristics. Dizengoff and Ibn Gabirol have higher commercial densities than Jabotinsky, with many more shops, cafes, and bars. The automobile traffic on those streets is also different. All three have noticeable car traffic, but Ibn Gabirol is a wider road than Dizengoff, and Jabotinsky is much wider than Ibn Gabirol and serves as the main artery that connects several cities to Tel-Aviv. It may be that Jabotinsky is confused with Residential because there are more residents living there. On Jabotinsky, there are four-story residential buildings (and some 10–20-story ones as well), mainly inhabited by families. In comparison, on Dizengoff and Ibn Gabirol Streets, there are three-story buildings inhabited mostly by young single people. For all these reasons, it is not surprising that these streets are classified differently, as their social function differ.

In **Figure 6(right)**, we illustrate the accuracy rate through the self-labeled iterations. The figure demonstrates the accuracy rate (Acc) in accordance with the percentage of cells that were labeled. After the first iteration, 10% of the cells are classified (5% labeled by ground truth knowledge +5% classified in the first iteration), and the accuracy rate is high (89%). However, notice that, in this stage of the process, 90% of the cells are yet to be classified. Through the process, as more cells are classified, the accuracy rate gradually declines—from 89% after the first iteration to 72% at the end of the process when all cells are classified. There are two reasons for this. First, in each iteration, incorrect labels (due to erroneous labeling of previous iterations) are added to the training set, causing the quality of the training set to decline. Second, as the iterations go on, the samples added to the training set are those that the algorithm was the least confident of in previous iterations. Notice we could have stopped the iterations before all the cells were classified. The accuracy rate drops more rapidly in the classification of the last 20% of the cells. If we would have stopped the process when 80% of the cells were classified, then the accuracy rate would have

been 81%. However, in that case, 20% of the cells would have been left unclassified, so this is left as a trade-off for the user.

5. Neighbor smoothing integrated into SSK

The lack of labeled data in our SSK semi-supervised methodology diminishes the classifier's ability. To achieve a more accurate classification, we used a smoothing process, which utilizes geographic neighbor similarity. Cells located close by in the geographic space have a greater chance of sharing the same land use because lands of unified social function are arbitrarily divided into cells and, thereby, neighboring cells tend to share similar social functions. To prevent confusion, we would like to emphasize that there are two different types of neighbors in the context of SSK—feature-space neighbors and geographic neighbors. Until this point in the paper, we have discussed feature-space neighbors. Two cells are considered feature-space neighbors if the Euclidian distance between their feature representations is relatively small. In the SSK without smoothing, only feature-space neighbors were considered. Geographic-space neighbors are cells closely located on the geographical map, and therefore, we use them for geographical smoothing.

Smoothing makes the results more homogenous in the geographical space. It causes the algorithm to be more accurate overall, but less sensitive to island land uses, relatively small lands that include a social function that is different from its surrounding areas. Because geographical space smoothing diminishes the chance of identifying these lands, we evaluated different degrees of smoothing, thus, controlling the trade-off between accuracy and sensitivity to island land uses.

The smoothing is integrated into the SSK process; in each iteration, before assigning a class, the geographic neighbors are also considered. The land-use array A_q , computed by the feature-space neighbors of x_q , is weighted with the geographical neighbors' land-use arrays (computed by their feature-space neighbors) to create an integrated array that is used for classification and confidence estimation. The rest remains the same—in every iteration, 5% of the samples are added to the training set G , with a proportion of the number of samples assigned to each class, and the process ends when all samples are labeled (or before, depending on the user/application).

To weigh between the query cell land-use array and its geographic neighbors' land-use arrays, we first need to define a neighbor. x_i is considered as x_q 's geographic neighbor if the geographical distance between them is smaller than a distance denoted as $radius_q$. The distance between two cells is defined as the distance between their geographical centers. The neighbors' radius of query cell x_q is given by:

$$radius_q = 3\sqrt{(width_q/2)^2 + (height_q/2)^2}, \quad (8)$$

where $width_q$ and $height_q$ are x_q 's width and height (meters).

The square root expression in Eq. (6) is the length of half of the cell's diagonal. That way, the radius is fitted to the size and shape of the cell. Half the diagonal is multiplied by 3 because, in a preliminary study, it was found to fit the problem. **Figure 7(left)** demonstrates the query cell's neighbor radius. Cell x_q is the default squared cell— $200 \times 200 \text{ m}^2$; therefore, $radius_q = 3\sqrt{(200/2)^2 + (200/2)^2} = 424.3\text{m}$. In the example



Figure 7. (left) The neighbors' radius for the query cell q . (right) An example in which query cell x_q has two equally closed neighbors x_a and x_b .

in **Figure 7**, six cells' centers fall inside the circle formed by the neighbors' radius and, thus, those six cells, numbered 1 to 6, are considered as x_q 's neighbors.

In **Figure 7(left)**, the cells within the neighbors' radius of x_q lay on different geographical distances from the center of x_q . For example, the centers of cells x_3, x_1 , and x_6 are 200, 283, and 400 meters away, respectively. We want to weigh the contribution of a neighbor according to its distance from the query cell because the closer the neighbor is, the greater the chance that it shares the same land use as the query cell. The weights are given by:

$$W_q^{(i)} = \frac{1}{D(x_q, x_i)^2} \quad \forall i \in nbrs_q, \quad (9)$$

where $nbrs_q$ is the set of x_q 's neighbors, and $D(x_q, x_i)$ is the geographical-based distance between query cell x_q and its neighbor x_i .

In the example demonstrated in **Figure 7(left)**, the weights of cells x_3, x_1 , and x_6 are $W_q^{(3)} = 1/200^2$, $W_q^{(1)} = 1/283^2$, and $W_q^{(6)} = 1/400^2$. Notice that between these three cells, cell x_3 is the closest to cell x_q , thus its weight is the highest accordingly.

Notice, we denote distances differently in the feature space and the geographical space. Lower case d is a distance in the feature space (Eq. (1)), and upper case D is a distance in the geographical space (Eq. (7)).

We then compute an array NA_q that combines land-use array for x_q 's neighbors by weighting every neighbor's distance from x_q :

$$NA_q = \frac{\sum_{i \in nbrs_q} W_q^{(i)} A_i}{\sum_{i \in nbrs_q} W_q^{(i)}}. \quad (10)$$

For demonstrating the mathematical equations used for integrating neighbor smoothing in SSK, we will use the example illustrated in **Figure 7(right)**. x_q has only two neighbors, x_a and x_b . Since x_a and x_b are located at the same distance from x_q , their weights are equal, $W_q^{(a)} = W_q^{(b)} = 1/268^2$.

Let us assume the land-use arrays are $A_a = (0, 0, 1, 0)$ and $A_b = (0, 0.8, 0, 0.2)$. Then NA_q is calculated by the weighted average of A_a and A_b : $NA_q = \frac{W_q^{(a)} A_a + W_q^{(b)} A_b}{W_q^{(a)} + W_q^{(b)}} = \frac{(1/268^2) A_a + (1/268^2) A_b}{(1/268^2) + (1/268^2)} = \frac{A_a + A_b}{2} = \frac{(0, 0, 1, 0) + (0, 0.8, 0, 0.2)}{2} = \frac{(0, 0.8, 1, 0.2)}{2} = (0, 0.4, 0.5, 0.1)$. As

can be seen, the value in entry 3 (0.5) is the highest in the array, indicating that x_q 's neighbors tend to be attributed to class 3, that is because x_a and its corresponding land-use array A_a are 100% attributed to class 3. However, x_q 's neighbors also tend to be attributed to class 2, which is because x_b is most likely attributed to class 2.

A_q and NA_q , the query cell land-use array and its neighbor's land-use array, are integrated to IA_q by calculating their weighted average:

$$IA_q = P \cdot NA_q + (1 - P) \cdot A_q, \quad (11)$$

where P is the weight of NA_q and, therefore, it is given to all of x_q 's neighbors together. We denote P as the neighbor weight. For example, consider again the example in **Figure 7(right)** and assume $P = 0.3$ and $A_q = (0.1, 0.8, 0.1, 0)$. Then,

$$IA_q = 0.3 \cdot (0, 0.4, 0.5, 0.1) + 0.7 \cdot (0.1, 0.8, 0.1, 0) = (0.07, 0.68, 0.22, 0.03). \quad (12)$$

Examining A_q extracted by x_q 's feature-space neighbors, it seems like x_q has the highest chance to be attributed to class 2, but examining NA_q , extracted by x_q 's geographic-space neighbors, it seems most likely that it belongs to class 3. However, after incorporating both spaces, x_q is most likely attributed to class 2

The neighbor weight P depends on the number of geographic neighbors x_q has. The more neighbors it has, the more reliable their weighted array is, and we want it to have a more significant role in determining x_q 's class. The formula for computing P

$$P(|nbrs_q|, \sigma) = \begin{cases} \sigma + \sigma \frac{(|nbrs_q| - 1)}{11} & |nbrs_q| > 0 \\ 0 & |nbrs_q| = 0 \end{cases}, \quad (13)$$

where $|nbrs_q|$ is the number of neighbors that x_q has, and σ is the smoothing parameter that determines the degree of influence that the neighbors have in the classification of the query cell. Setting a low σ , for example, will cause the neighbors of the query cells to be less significant in the classification.

In the example above, $P = 0.3$, because the number of neighbors $|nbrs_q| = 2$ (as can be seen in **Figure 7(right)**), and $\sigma = 0.275$. Therefore, $P(|nbrs_q|, \sigma) = 0.275 + 0.275 \frac{(2-1)}{11} = 0.3$.

Eq. (9) is designed in a way that when x_q has only one neighbor, its neighbor weight is $P(|nbrs_q| = 1, \sigma) = \sigma$, whereas if x_q has 12 neighbors (the maximal number of neighbors because more neighbors cannot fit inside the neighbor's radius considering the shape and size of the cells), then $P(|nbrs_q| = 12, \sigma) = 2\sigma$. The value of P grows linearly between the case of only one neighbor and the case of 12 neighbors. If the query cell does not have any neighbors, then $P(|nbrs_q| = 0, \sigma) = 0$, and $IA_q = 0 \cdot NA_q + (1 - 0) \cdot A_q = A_q$. Because there are no neighbors to consider, NA_q will have no influence on setting IA_q , and $IA_q = A_q$.

The classification confidence is calculated as in Eq. (3), but here it is calculated over IA_q instead of A_q

$$confidence_q = \max(IA_q), \quad (14)$$

where in the example, $confidence_q = \max(0.07, 0.68, 0.22, 0.03) = 0.68$.

Again, in each iteration, the number of samples added to G from each class is proportional to the number of cells assigned to that class in this iteration. If $confidence_q$ is high enough, then x_q is classified as the class with the highest value in IA_q . The algorithm ends when all samples are added to G (or before based on the user/application).

The procedure of the SSK algorithm with neighbor smoothing:

1. Set σ (the smoothing parameter; can be set using a validation set)
2. $G \leftarrow L$ (set the training set G to be the predefined labeled samples L)
3. $Q \leftarrow O \setminus G$ (Q and O are the sets of unlabeled samples and all samples, respectively)
4. For each $x_q \in Q$ (for each yet unlabeled sample)

- a. $A_q = \frac{\sum_{i=1}^k w_q^{(i)} A_i}{\sum_{i=1}^k w_q^{(i)}}$ (land-use array) (Eq. (2))

- b. $radius_q = 3 * \sqrt{\left(\frac{width_q}{2}\right)^2 + \left(\frac{height_q}{2}\right)^2}$ (neighbor radius) (Eq. (5))

- c. $nbrs_q \leftarrow \emptyset$

- d. For each $x_i \in G$

If $D(x_q, x_i) < radius_q$ then $nbrs_q \leftarrow (nbrs_q \cup x_i)$ (add to $nbrs_q$ the x_i neighbor)

- e. $w_q^{(i)} = \frac{1}{D(x_q, x_i)^2} \forall i \in nbrs_q$ (Eq. (7))

- f. $NA_q = \frac{\sum_{i \in nbrs_q} w_q^{(i)} A_i}{\sum_{i \in nbrs_q} w_q^{(i)}}$ (neighbors' land-use array) (Eq. (8))

- g. If $|nbrs_q| > 0$ then $P(|nbrs_q|, \sigma) = \sigma + \sigma \frac{(|nbrs_q| - 1)}{11}$

Else $P(|nbrs_q|, \sigma) = 0$ (Eq. (10))

- h. $IA_q = P \cdot NA_q + (1 - P) \cdot A_q$ (integrated land-use array) (Eq. (9))

- i. $confidence_q = \max(IA_q)$ (Eq. (11))

5. For each land-use class c

- a. $Z \leftarrow$ sub areas with the highest confidence assigned to c

- b. $G \leftarrow G \cup Z$; $Q \leftarrow Q \setminus Z$ (the cells assigned to class c with the highest confidence are added to G and subtracted from Q)

6. If $|Q| > 0$, then go to step 4, else output G

5.1 Example

Figure 8 illustrates an example of the classification of a query cell x_s after considering both spaces: x_s 's neighbors in the feature space, under the title "Feature space" (**Figure 8(left)**), and x_s 's neighbors in the geographical space, under the title "Geographical space" (**Figure 8(right)**).

In this example, the class assignment is based on the two samples that are closest in the feature space, and there are four land-use classes. x_s 's two nearest neighbors in the feature space are x_r and x_q , and their land-use arrays are $A_r = (0, 0, 1, 0)$ and $A_q = (0, \frac{3}{4}, 0, \frac{1}{4})$ with computed weights $w_s^{(r)} = 1$ and $w_s^{(q)} = 4$, respectively. Notice that $w_s^{(r)}$ and $w_s^{(q)}$ are set, respectively, according to the x_r and x_q feature space distances from the query cell x_s . In **Figure 8(left)**, under the title "Feature space," the two bar graphs represent the land-use arrays of x_r and x_q , which are A_r and A_q , respectively. For example, because A_r has 100% confidence of being attributed to class 3, the value

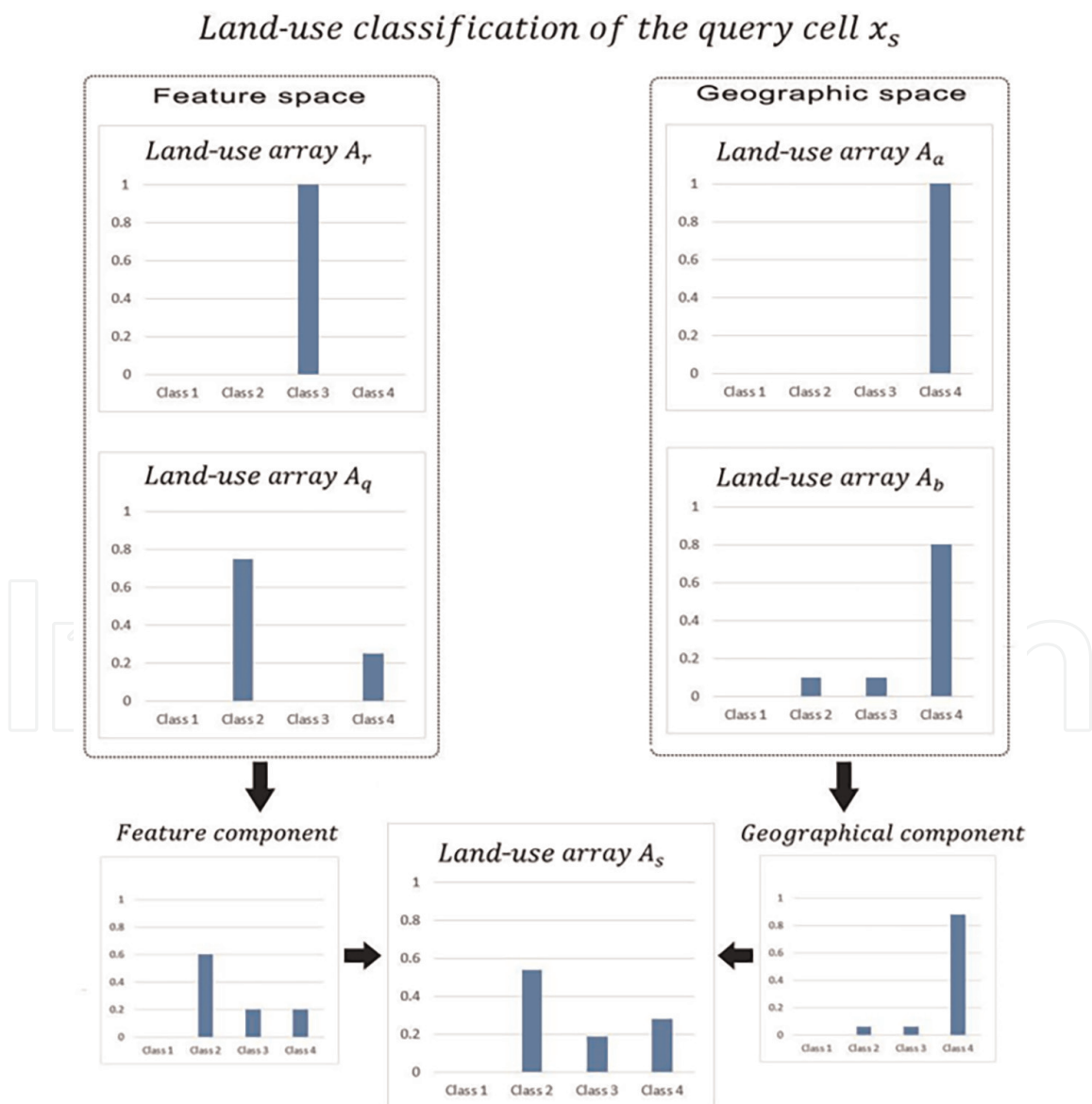


Figure 8. Land-use classification of a query cell based on (left) only the feature space, (right) only the geographical space, and (bottom) both using neighbor smoothing.

of the bar of class 3 is 1, and the values of the other bars are 0. x_s 's land-use array is computed as a weighted average of A_r and A_q (Eq. (2)): $A_s = \frac{w_s^{(r)}A_r + w_s^{(q)}A_q}{w_s^{(r)} + w_s^{(q)}} = (0, 0.6, 0.2, 0.2)$, as is demonstrated in **Figure 8(left)**, and it is the result of the weighted average of A_r and A_q . Without neighbor smoothing, assigning a class to x_s would have been decided at this point, and x_s would have been assigned to the class which is the highest in A_s , that is, class 2.

But here, we integrate the neighbors' land use in the classification decision. Let us assume x_s has two geographical neighbors x_a and x_b , and their land-use arrays are $A_a = (0, 0, 0, 1)$ and $A_b = (0, 0.1, 0.1, 0.8)$, and their weights are $W_s^{(a)} = 2$ and $W_s^{(b)} = 3$, respectively. Notice that $W_s^{(a)}$ and $W_s^{(b)}$ are set according to the Euclidean geographic distance of x_a and x_b from the query cell x_s . In **Figure 8(right)**, under the title "Geographic space," the two bar graphs represent the land-use arrays of x_a and x_b . x_s 's neighbors' land-use array is computed by a weighted average of A_a and A_b (Eq. (7)): $NA_s = \frac{W_s^{(a)}A_a + W_s^{(b)}A_b}{W_s^{(a)} + W_s^{(b)}} = (0, 0.06, 0.06, 0.88)$. NA_s is demonstrated in **Figure 8(right)** under the title "Geographical component." The maximal value of 0.88, based on the influential geographic neighbors of x_s 's, challenge the cell's previous assignment of class 2 to that of class 4.

The final decision about assigning a class to x_s is after combining the feature component A_s and the geographic component NA_s . Let us set the smoothing parameter σ at 0.1, and thus the weight of the neighbors' component is (Eq. (9)): $(|nbrs_q| = 2, \sigma = 0.1) = 0.11$. x_s 's integrated land-use array (Eq. (8)) is $IA_s = 0.11 \cdot NA_s + (1 - 0.11) \cdot A_s = (0, 0.54, 0.18, 0.28)$, as is demonstrated in **Figure 8(bottom)** under the title "Land-use array x_s ." If we consider IA_s 's 0.54 confidence high enough, then x_s would be classified as class 2 and added to the training set G for the next iteration.

6. Empirical evaluation of neighbor smoothing integrated into SSK

In this section, we evaluate the effect of the neighbor smoothing integrated into SSK. **Figure 9** compares the SSK accuracy with different neighbor smoothing values σ , varying from 0 (no smoothing performed) to 0.25. As σ is higher, the accuracy rate is higher, varying from 74% when no smoothing is performed to 80% when σ is 0.25.

Recall that the accuracy rate of RF is 84%. Although not reaching RF's accuracy rate, the smoothing enables SSK accuracy to be significantly close to that of RF even though the latter is a supervised paradigm that uses a much bigger training set (87.5% of the cells are labeled and used as ground truth for training the RF in each of the eight cross-validation folds, comparing to only 5% of the cells that are used by the SSK). However, the effectivity of the smoothing process is overestimated because the neighbor similarity property that the neighbor smoothing relies on is exaggerated in our dataset. In the process of selecting the areas, we chose ones that are homogenous in land use, and their "real" land-use label is relatively easy for locals to determine. This means that most areas include only one land use in a specific hour. Homogenous areas have some advantages—they are practical for labeling, and they can serve to assess the process feasibility, but they are less representative of normal urban behavior. Thus, the areas we selected are overly homogenous. Therefore, the chance of neighboring cells sharing the same land use is higher than in normal urban behavior.

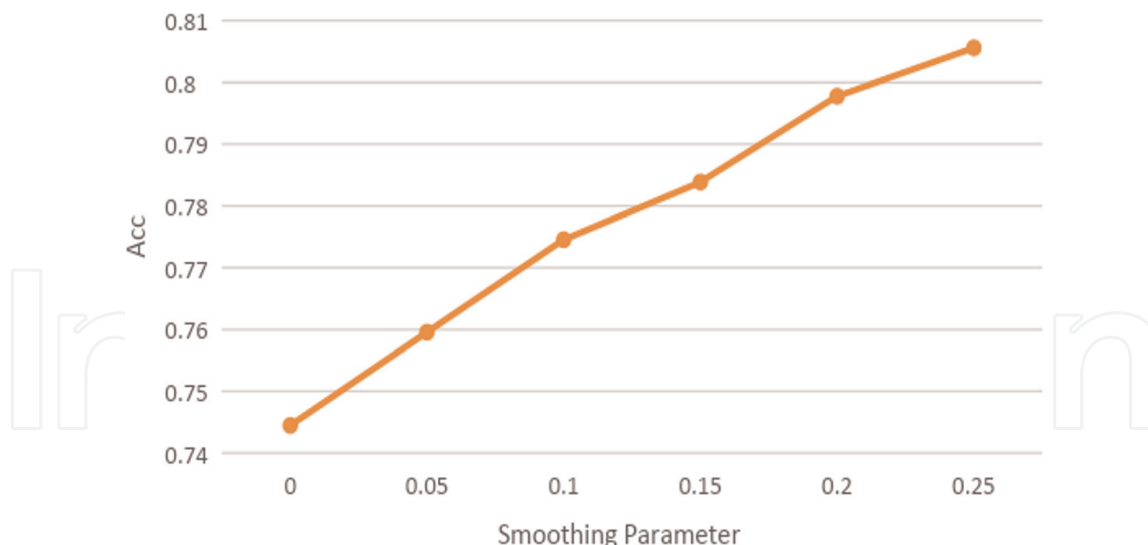


Figure 9.
Effect of smoothing parameter σ on the accuracy rate (Acc).

Island land uses located in the heart of other land uses, to which the neighbor smoothed SSK is less sensitive, occur less frequently in our data. We do expect this process to also perform well in a less homogenous dataset, however, in a more limited manner. We expect the algorithm to perform better when setting a higher smoothing parameter value, up to a point where the results become too homogenous, causing too many errors in identifying island land uses.

Figure 10 compares the geographical confusion maps of SSK classification without (**Figure 10a** and **b**) and with (**Figure 10c** and **d**) neighbor smoothing with $\sigma = 0.25$ on the work hours 8 a.m. to 5 p.m. in Ra'anana (**Figure 10a** and **c**) and Kiryat Arye, an industrial area of Petch Tikva (**Figure 10b** and **d**). Recall that the colors in each cell demonstrate accumulation of the classification results of the different hours and various random cells chosen to be used for the initial set of labeled cells.

The smoothing causes the classification assignment to be more consistent and less influenced by the randomness effect caused by randomly chosen cells with predefined land use. Considering more factors in the cell class assignment, that is, considering the cell's neighbors, diminishes the effect of randomness and lowers the classification variance. For example, see the classification of the industrial cells in Kiryat Arye. This is an area of homogenous social function, and the smoothing makes classification there more consistent. The cells are more uniformly colored in the same color (yellow) indicating that they were classified to the same class in more of the iterations. The smoothing also lowers SSK's bias. Because of the smoothing, all cells in Kiryat Arye are correctly classified as Industrial in most of the algorithm iterations. Without smoothing, 35 out of the 42 cells are well classified in most of the runs, while with smoothing, all 42 cells are well classified in most of them. For example, the bottom-right cell in Kiryat-Arye without smoothing (**Figure 10b**) is incorrectly classified in most runs (note the small yellow area indicating "Industrial" compared to the other colors), whereas with smoothing (**Figure 10d**), this cell is mostly correctly classified as "Industrial."

On the downside, neighbor smoothing diminishes the ability to identify "island" land uses. For example, see the commercial island street in Ra'anana located in the heart of several neighborhoods. Notice that even before smoothing (**Figure 10a**), SSK mostly classified it as Residential, as it is affected by nearby residential cells (as described above). Because the triangulating signal strength location estimation

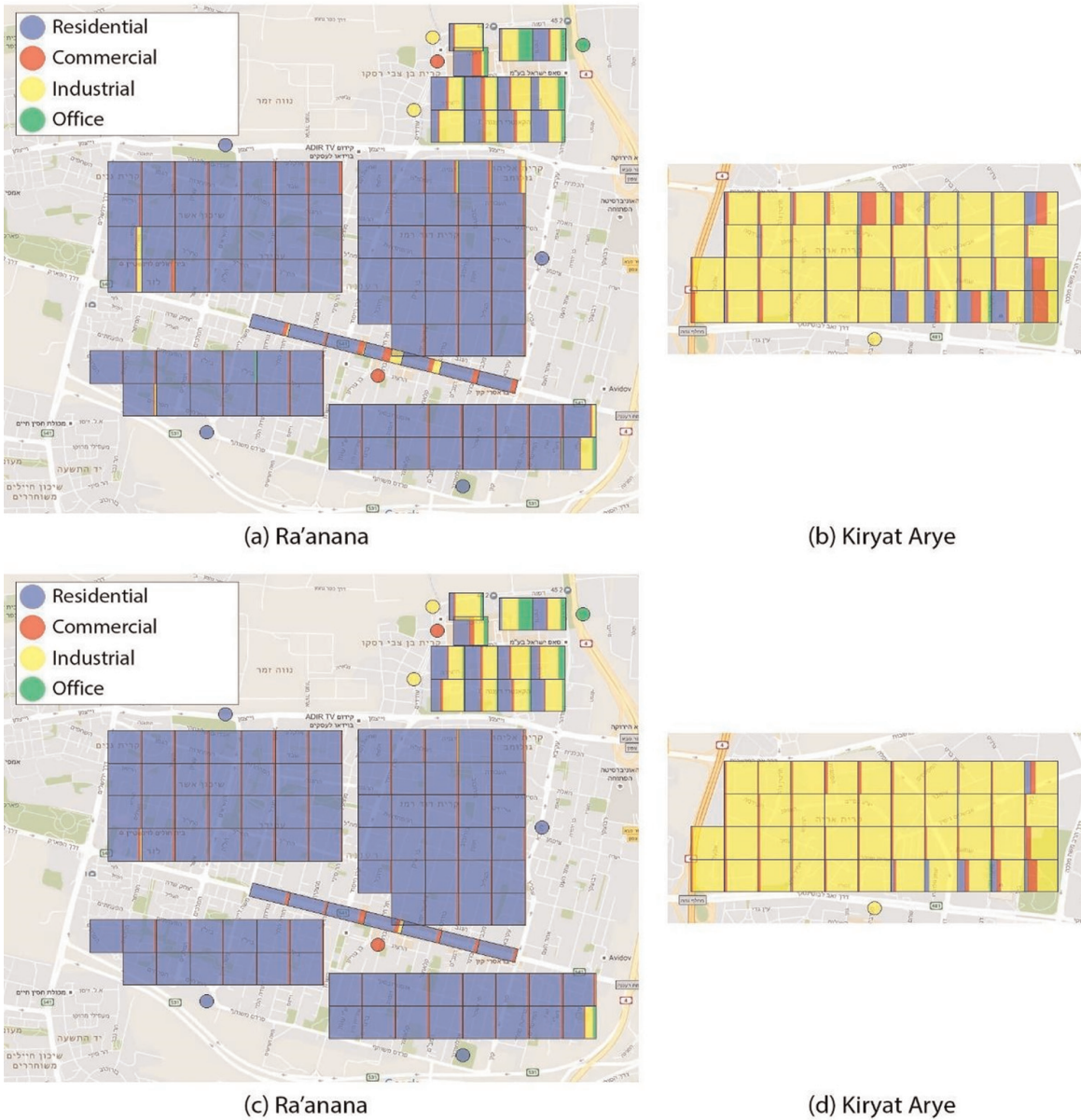


Figure 10. Geographical confusion maps of SSK without (a, b) and with (c, d) smoothing ($\sigma = 0.25$).

technology used for the location estimation in this work suffers from inaccuracy, the extent of the problem is not negligible. Especially, small and narrow (“island”) streets that are surrounded by a “sea” of residential neighborhoods are affected by this inaccuracy. Smoothing complicates the task of identifying island land use, as it makes the results more homogenous, and thus, the classifier is more decisive and mistakenly classifies more to Residential (in the case of Ra’anana; **Figure 10c**).

Smoothing influence depends on the geographical structure of the land use. We will distinguish between geographically wide-stretching land uses, such as Residential, and island land uses, which are usually located in the heart of a wide-stretching land use, such as commercial streets or shopping malls, or located at the borders between them, such as highways.

Neighbor smoothing causes the wide-stretching land uses to expand over island land uses and, consequently, more lands are classified as wide-stretching. Therefore, wide-stretching land uses recall increases—more cells are classified as wide-stretching with more cells identified correctly, but precision declines because some of the “new”

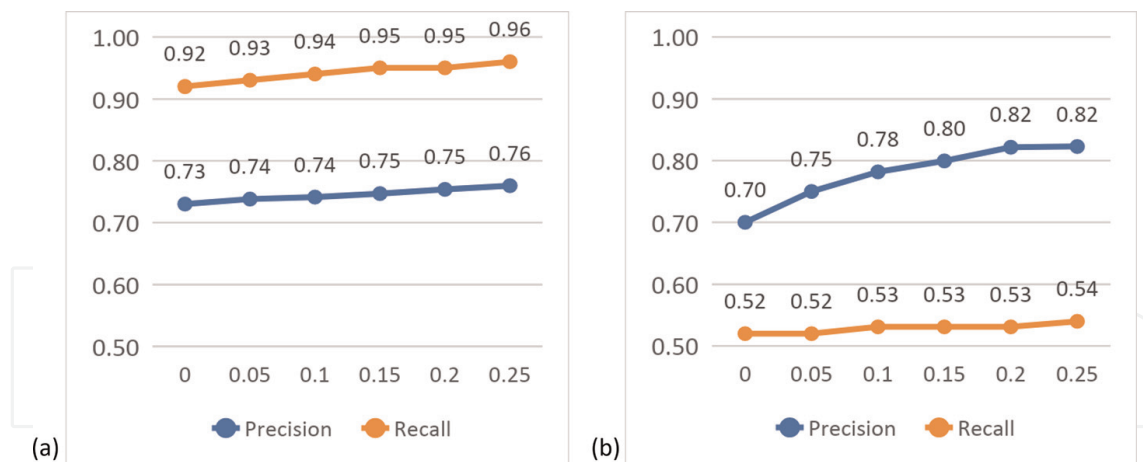


Figure 11. Smoothing effect (σ) on the precision and recall performance measures in classifying (a) wide-stretching residential land uses and (b) narrow commercial island land uses.

wide-stretching cells belong to the neighboring island land use; thus, the percentage of correctly classified cells declines. The recall of island land uses decreases because fewer islands are identified, whereas precision increases because

the cells classified as islands are those that are the most unambiguously correctly classified.

However, because our dataset is homogenous, both precision and recall improve in all land uses. **Figure 11** demonstrates the effect of the smoothing parameter on recall and precision of wide-stretching Residential (**Figure 11a**) and Commercial island land (**Figure 11b**) uses.

In the wide-stretching Residential example, recall ascends from 0.92 to 0.96; thus, 50% of the unidentified Residential cells are identified due to the smoothing. Whereas in the Commercial island land use, recall ascent is less prominent, from 0.52 to 0.54; thus, a 4% rise of the unidentified Commercial cells is identified due to the smoothing. As we would expect, the recall improvement in the wide-stretching land uses is considerably more significant. In the wide-stretching Residential cell, precision ascends from 0.73 to 0.76; thus, the percentage of cells incorrectly assigned as Residential is slightly reduced from 27–24%. Whereas in the Commercial island land use, precision rises significantly from 0.70 to 0.82; thus, the percentage of cells incorrectly assigned as Commercial is reduced from 30–18%. As we would expect, the precision improvement in the island land uses is considerably more significant.

7. Discussion and conclusions

Previous works dedicated to social land-use mapping mostly used more than one data resource and complex methodologies that integrate them. Other works assumed substantial prior knowledge about the examined lands but when used relatively little knowledge about the examined city achieved not satisfactory accuracy rates [18]. The main contribution of this paper is that it offers a method for social land-use mapping when only sparse prior knowledge about the examined city exists, and by relying on the CDR, an inexpensive and available data resource is routinely gathered by telecom operators.

We introduced SSK, a semi-supervised algorithm that requires a relatively small number of labeled samples and, therefore, fits the condition of sparse prior

knowledge. The heart of SSK is the combination of the KNN classifier and the self-labeled technique that enables the enlargement of the training set in an iterative manner. SSK achieves an accuracy rate of 74.4%, a significantly higher rate than that achieved in the works of Toole et al. [42] and Pei et al. [18] of 54% and 58%, respectively. These works also relied mainly on CDR as their main data resource. However, it is not possible to infer that SSK performs better than their methodologies because our validation was on a very different dataset. Whereas they performed land-use mapping of a whole city, Boston in the work of Toole et al. [42], and Singapore in the work of Pei et al. [18], we chose areas of relatively homogenous social function from different cities in Israel. The task of classification in deliberately chosen areas of more “pure” social function is easier. We also compared the SSK’s performance to that of a random forest (RF) classifier trained using many more labeled places, with 87.5% of the surface labeled (7/8 of the data set is used for training) compared to 5% in SSK. As expected, RF lowered the bias and variance of the classification and achieved a higher accuracy rate than SSK, but relative to the prior knowledge used in SSK, the performance gaps are mild. In a condition of only a small number of labeled samples, the effectiveness of conservative supervised classification algorithms, such as RF, deteriorates. Therefore, if getting additional land-use labels is out of reach or too expensive, it is better to use SSK.

SSK heavily relies on few labeled cells. If the land use in these cells is relatively mixed, then it has the potential to heavily damage the classification. Therefore, if cells of relatively “pure” social function cannot be obtained, then it is better to consider using an unsupervised method. The good thing is that, in most cases, the ground truth labeled cells are easier to be categorized to one land use (that is the reason they are chosen to be labeled); thus, they are relatively not mixed. Through the iterative steps, coverage of classified lands grows, but accuracy declines. We offer the option to stop the process before all land use is classified. For example, stopping the process at 80% of classified areas raises the accuracy rate to 81%, instead of 74.4%, if all areas are classified.

We also introduced a version of SSK that includes neighbor smoothing. We rely on the neighbor social land-use similarity property and offer a unique interpretation of KNN—a KNN that considers both the feature-space neighbors as in the regular KNN and the geographic space neighbors. We discussed the merits of incorporating smoothing, along with its drawbacks. Smoothing improves the overall accuracy; however, it degrades the chances to discover narrow land of a social function that is different than its surroundings. Therefore, the algorithm enables a parameter that sets the level of smoothing performed and, thus, controls the trade-off between overall accuracy and sensitivity to an exceptional social function. High levels of neighbor smoothing should be most effective in cities that are more “planned”; these cities tend to be more divided into functional parts of homogenous social function. Validating neighbors’ smoothing shows that it indeed improves SSK’s accuracy rate to 80% with the most smoothed results. In our dataset, it also improves the discovery rate of island land uses. This is mainly due to the homogeneity of the social function of the areas we chose to include in this work.

SSK is assembled of several components, each aiming to tackle some of the difficulties in the problem of mapping social functions (e.g., lack of labeled samples). In addition, SSK leverages opportunities inherent in the problem:

1. Self-labeled technique – While it might be costly to attain sufficient labeled samples needed for a classic classifier, it is relatively easy to attain labels of few

locations in a city. Residents can participate in the process of self-labeling of their city and thereby contribute to the efforts to make their own city smarter.

2. Neighbor smoothing – Usage of only CDR as a data resource requires creative solutions for improving the accuracy of the identification. One property that can be utilized is the resemblance in terms of the social function of neighboring parts of the city. Neighbor smoothing incorporates the geographic neighbors in the classification, and, in our case, it proved to improve the average accuracy rate from 74% to 80%. By integrating a smoothing parameter, we limited the effect of neighbor smoothing to prevent overly homogenous classification that is not sensitive to an exceptional social function.
3. Usage of KNN classifier—KNN fits perfectly for integrating the two spaces—feature space and geographic space and thus incorporates neighbor smoothing.
4. Usage of the distance weighted version of KNN-DKNN, which gives in the classification higher weight to closer neighbors, is mainly implemented for integrating the geographic space. Obviously, adjacent lands tend to share a similar social function, while lands that are relatively close but not adjacent have a lower probability to share the same social function. Therefore, we chose to use DKNN, which would cause the classification to rely more on the closest lands. The same logic is applied to the feature space, mainly for uniformity purposes between the two spaces.

In future work, we would like to validate the offered methodology on a whole city. Because some of the social functions are not well identified, creative solutions will be needed to identify them more consistently. In addition, further research may lead to an enhanced smoothing logic that is more sensitive to island land uses. A limitation of our approach may be that cellular communication cannot always capture the differences between some land uses (e.g., when the communication is limited in less populated areas), and then more data resources will be needed. Therefore, it may also be interesting to examine combining this methodology with other data resources, such as POI and remote-sensing imagery.

Acknowledgements

This work was supported by the Israel Ministry of Science and Technology.

IntechOpen

IntechOpen


Author details

Oded Zinman and Boaz Lerner*

Industrial Engineering and Management, Ben-Gurion University of the Negev, Israel

*Address all correspondence to: boaz@bgu.ac.il

IntechOpen

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Alberti M, Marzluff JM, Shulenberg E, Bradley G, Ryan C, Zumbrunnen C. Integrating humans into ecology: Opportunities and challenges for studying urban ecosystems. *AIBS Bulletin*. 2003;**53**(12): 1169-1179
- [2] Zhang X, Du S. A linear Dirichlet mixture model for decomposing scenes: Application to analyzing urban functional zonings. *Remote Sensing of Environment*. 2015;**169**:37-49
- [3] Yuan J, Zheng Y, Xie X (2012) Discovering regions of different functions in a city using human mobility and POIs. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Beijing, pp. 186-194.
- [4] Zheng Y, Capra L, Wolfson O, Yang H. Urban computing: Concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*. 2014; **5**(3):38
- [5] Li C, Wang J, Wang L, Hu L, Gong P. Comparison of classification algorithms and training sample sizes in urban land classification with Landsat thematic mapper imagery. *Remote Sensing*. 2014; **6**(2):964-983
- [6] Okafor CC, Aigbavboa C, Thwala WD. A bibliometric evaluation and critical review of the smart city concept—making a case for social equity. *Journal of Science and Technology Policy Management*. 2022. Available from: <https://doi-org.ezproxy.bgu.ac.il/10.1108/JSTPM-06-2020-0098>
- [7] Kim HM, Sabri S, Kent A. Smart cities as a platform for technological and social innovation in productivity, sustainability, and livability: A conceptual framework. *Smart Cities for Technological and Social Innovation*. 2021. pp. 9-28
- [8] Hu T, Yang J, Li X, Gong P. Mapping urban land use by using Landsat images and open social data. *Remote Sensing*. 2016;**8**(2):151
- [9] Sun B, Zhang Y, Zhou Q, Zhang X. Effectiveness of semi-supervised learning and multi-source data in detailed urban landuse mapping with a few labeled samples. *Remote Sensing*. 2022;**14**(3):648
- [10] Pan S, Zhou W, Piramuthu S, Giannikas V, Chen C. Smart city for sustainable urban freight logistics. *International Journal of Production Research*. 2021;**59**(7):2079-2089
- [11] Kaginalkar A, Kumar S, Gargava P, Niyogi D. Review of urban computing in air quality management as smart city service: An integrated IoT, AI, and cloud technology perspective. *Urban Climate*. 2021;**39**:100972
- [12] Bibri SE. Eco-districts and data-driven smart eco-cities: Emerging approaches to strategic planning by design and spatial scaling and evaluation by technology. *Land Use Policy*. 2022; **113**:105830
- [13] Bibri SE. Data-driven smart sustainable cities of the future: Urban computing and intelligence for strategic, short-term, and joined-up planning. *Computational Urban Science*. 2021; **1**(1):1-29
- [14] Wang A, Lin W, Liu B, Wang H, Xu H. Does smart city construction improve the green utilization

efficiency of urban land? *Land*. 2021; **10**(6):657

[15] Laurini R. A primer of knowledge management for smart city governance. *Land Use Policy*. 2021:111

[16] Arribas-Bel D, Tranos E. Characterizing the spatial structure(s) of cities “on the fly”: The space-time calendar. *Geographical Analysis*. 2018; **50**(2):162-181

[17] Zinman O, Lerner B. Utilizing digital traces of mobile phones for understanding social dynamics in urban areas. *Personal and Ubiquitous Computing*. 2020;**24**:535-549

[18] Pei T, Sobolevsky S, Ratti C, Shaw SL, Li T, Zhou C. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*. 2014;**28**(9): 1988-2007

[19] Goodchild MF, Janelle DG. The city around the clock: Space-time patterns of urban ecological structure. *Environment and Planning A*. 1984;**16**(6):807-820

[20] Jiang S, Ferreira J, González MC. Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*. 2012;**25**(3): 478-510

[21] Yue Y, Lan T, Yeh AG, Li QQ. Zooming into individuals to understand the collective: A review of trajectory-based travel behaviour studies. *Travel Behaviour and Society*. 2014;**1**(2):69-78

[22] Batty M. Big data, smart cities and city planning. *Dialogues in Human Geography*. 2013;**3**(3):274-279

[23] Lu D, Weng Q. Use of impervious surface in urban land-use classification. *Remote Sensing of Environment*. 2006; **102**(1):146-160

[24] Heiden U, Heldens W, Roessner S, Segl K, Esch T, Mueller A. Urban structure type characterization using hyperspectral remote sensing and height information. *Landscape and Urban Planning*. 2012;**105**(4):361-375

[25] Wen D, Huang X, Zhang L, Benediktsson JA. A novel automatic change detection method for urban high-resolution remotely sensed imagery based on multiindex scene representation. *Geoscience and Remote Sensing*. 2016;**54**(1):609-625

[26] Wu C, Zhang L, Zhang L. A scene change detection framework for multi-temporal very high resolution remote sensing images. *Signal Processing*. 2016; **124**:184-197

[27] Gao S, Janowicz K, Couclelis H. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*. 2017; **21**(3):446-467

[28] Liu Y, Liu X, Gao S, Gong L, Kang C, Zhi Y, et al. Social sensing: A new approach to understanding our socioeconomic environments. *Annals of the Association of American Geographers*. 2015;**105**(3):512-530

[29] Tu W, Cao J, Yue Y, Shaw SL, Zhou M, Wang Z, et al. Coupling mobile phone and social media data: A new approach to understanding urban functions and diurnal patterns. *International Journal of Geographical Information Science*. 2017;**31**(12): 2331-2358

- [30] Toch E, Lerner B, Ben-Zion E, Ben-Gal I. Analyzing large-scale human mobility data: A survey of machine learning methods and applications. *Knowledge and Information System*. 2019;**58**:501-523
- [31] Liu X, He J, Yao Y, Zhang J, Liang H, Wang H, et al. Classifying urban land use by integrating remote sensing and social media data. *International Journal of Geographical Information Science*. 2017; **31**(8):1675-1696
- [32] Liu X, Kang C, Gong L, Liu Y. Incorporating spatial interaction patterns in classifying and understanding urban land use. *International Journal of Geographical Information Science*. 2016;**30**(2): 334-350
- [33] Long Y, Thill J-C. Combining smart card data and household travel survey to analyze jobs–housing relationships in Beijing. *Computers, Environment and Urban Systems*. 2015;**53**:19-35
- [34] Zhou Y, Thill J-C, Xu Y, Fang Z. Variability in individual home-work activity patterns. *Journal of Transport Geography*. 2021;**90**
- [35] Shen Y, Karimi K. Urban function connectivity: Characterisation of functional urban streets with social media check-in data. *Cities*. 2016; **55**:9-21
- [36] Ye M, Yin P, Lee WC, Lee DL. 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*. Beijing. 2011. pp. 325-334.
- [37] Sheng C, Zheng Y, Hsu W, Lee ML, Xie X. Answering top-k similar region queries. In: *International Conference on Database Systems for Advanced Applications*. Berlin, Heidelberg: Springer; 2010. pp. 186-201
- [38] Khoroshevsky F, Lerner B. Human mobility-pattern discovery and next-place prediction from GPS data. In: Schwenker F, Scherer S, editors. *Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction (MPRSS)*. Berlin: Springer; 2017
- [39] Ben Zion E, Lerner B. Identifying and predicting social lifestyles in people's trajectories by neural networks. *EPJ Data Science*. 2018;**7**(45):1-27
- [40] Zhao Z, Shaw SL, Xu Y, Lu F, Chen J, Yin L. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*. 2016; **30**(9):1738-1762
- [41] Trasarti R, Olteanu-Raimond AM, Nanni M, Couronné T, Furletti B, Giannotti F, et al. Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. *Telecommunications Policy*. 2015;**39**(3): 347-362
- [42] Toole JL, Ulm M, González MC, Bauer D. Inferring land use from mobile phone activity. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*. ACM, Beijing. 2012. pp. 1-8
- [43] Wang H, Calabrese F, Di Lorenzo G, Ratti C. Transportation mode inference from anonymized and aggregated mobile phone call detail records. In: *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference*. Funchal, Portugal. 2010. pp. 318-323

- [44] Isaacman S, Becker R, Caceres R, Kobourov S. Identifying important places in people's lives from cellular network data. In: International Conference on Pervasive Computing. 2011. pp. 133-151
- [45] Calabrese F, Ferrari L, Blondel VD. Urban sensing using mobile phone network data: A survey of research. *ACM Computing Surveys*. 2015;**47**(2):25
- [46] Breiman L. Random forests. *Machine Learning*. 2001;**45**(1):5-32
- [47] Bezdek JC, Ehrlich R, Full W. FCM: The fuzzy c-means clustering algorithm. *Computers & Geosciences*. 1984;**10**(2-3, 203):191
- [48] Nugraha AT, Waterson B, Blainey S, Nash F. On the consistency of urban cellular automata models based on hexagonal and square cells. *Environment and Planning B: Urban Analytics and City Science*. 2021;**48**:845-860
- [49] Leyk S, Balk D, Jones B, et al. The heterogeneity and change in the urban structure of metropolitan areas in the United States, 1990–2010. *Sci Data*. 2019;**6**:321
- [50] Triguero I, García S, Herrera F. Self-labeled techniques for semi-supervised learning: Taxonomy, software and empirical study. *Knowledge and Information Systems*. 2015;**42**(2): 245-284
- [51] Dudani SA. The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*. 1976;**4**:325-327
- [52] Mehta S, Shen X, Gou J, Niu D. A new nearest centroid neighbor classifier based on K local means using harmonic mean distance. *Information*. 2018;**9**(9):234
- [53] Ghosh AK. On optimum choice of k in nearest neighbor classification. *Computational Statistics & Data Analysis*. 2006;**50**(11):3113-3123
- [54] Fortmann-Roe S. Understanding the bias-variance tradeoff. 2012. Available online at <http://scott.fortmann-roe.com/docs/BiasVariance.html>. [Accessed 9 November, 2018]
- [55] Duda RO, Hart PE, Stork DG. *Pattern Classification*. John Wiley & Sons; 2012