Netw Spat Econ (2014) 14:647–667
DOI 10.1007/s11067-014-9264-4

Inferring Urban Land Use Using Large-Scale Social Media Check-in Data

Xianyuan Zhan · Satish V. Ukkusuri · Feng Zhu

Published online: 18 September 2014

Abstract Emerging location-based services in social media tools such as Foursquare and Twitter are providing an unprecedented amount of public-generated data on human movements and activities. This novel data source contains valuable information (e.g., geo-location, time and date, type of places) on human activities. While the data is tremendously beneficial in modeling human activity patterns, it is also greatly useful in inferring planning related variables such as a city's land use characteristics. This paper provides a comprehensive investigation on the possibility and validity of utilizing large-scale social media check-in data to infer land use types by applying the state-of-art data mining techniques. Two inference approaches are proposed and tested in this paper: the unsupervised clustering method and supervised learning method. The land use inference is conducted in a uniform grid level of 200 by 200 m. The methods are applied to a case study of New York City. The validation result confirms that the two approaches effectively infer different land use types given sufficient check-in data. The encouraging result demonstrates the potential of using social media check-in data in urban land use inference, and also reveals the hidden linkage between the human activity pattern and the underlying urban land use pattern.

Keywords Land use · Social media · Foursquare · Geo-location data · Data mining · Clustering algorithms · Supervised learning algorithms

X. Zhan · S. V. Ukkusuri (✉) · F. Zhu
Lyles School of Civil Engineering, Purdue University, 550 Stadium Mall Drive, West Lafayette, IN 47907, USA
e-mail: sukkusur@purdue.edu

X. Zhan
e-mail: zhanxianyuan@purdue.edu

F. Zhu
e-mail: zhu214@purdue.edu

1 Introduction

As an important component of transportation planning, land use analysis greatly facilitates planners' understanding of the influence of transportation movements on the use of space. The essential element in the land use analysis is to classify different facilities and geographical areas based on the type of use and functionality. The traditional approach to identify land use types involves checking the building permit data, on-site investigation, survey data, and questionnaires, which requires intensive labor resources, time, and money. Besides the traditional approach, one popular approach of land use identification is to use remote sensing or high spatial resolution satellite sensors (Barnsley and Barr 1996; Mesev 1998; Yang and Lo 2002; Sun et al. 2007). However, the approach may not be applicable to land use analysis in urban area, and the classification results are dependent on the image resolution, information extraction accuracy, and the sensor attributes (Moran et al. 1997; Schmit et al. 2006).

With the rapid development of mobile sensors, mobile phones and handheld devices, a new opportunity to use data from these tools is available. This is a novel way of land use detection based on human mobility and activity data, which can provide a powerful supplement to the traditional approach. Song et al. (2010) showed that human behavior can be captured and predicted by their mobility and activity pattern, as human behavior is characterized by circadian regularity patterns. This renders the data provided by ubiquitous technologies (e.g. GPS data, mobile phone data, and social media check-in data) a good source of human mobility and activity pattern analysis, as this data source contains invaluable information on human behavior, from which the mobility and activity pattern can be inferred. As shown in (González et al. 2008), there exists some simple reproducible mobility patterns for human behavior despite the individual variation of their travel histories. Over the past few years, location-based services (LBS) (e.g. Foursquare) in online social networks have provided an unprecedented amount of public-generated data on human movements and activities. Typically, location-based service activates the geo-location functionality of smartphones and allows users to "Check-in" to physical places. Thereby the human activity pattern can be revealed from the "check in" information of places produced by users. Moreover, the rich information of human mobility and activity contained in the "check-in" data offers tremendous promise to infer urban land use. The goal of this paper is to demonstrate the potential of using "check-in" data from social media tools to infer land use types.

The mutual influence between human behavior (in terms of human mobility and activity pattern) and land use classification, especially how land use can be inferred from the mobility and activity pattern has received some interest in recent literature. Soto and Frías-Martínez (2011a) separated different land use by using information contained in cell phone records. The fuzzy c-means clustering approach was applied to capture the inherent fuzziness of human behavior. Qi et al. (2011) inferred the regions' social functions by analyzing the temporal variation of the get-on and drop-offs of a city-scale taxi GPS data set. Agglomerative clustering method is applied to build up the relationship between the GPS data and the regions' social function. Preliminary results show a correlation accuracy of 94.44 %. However, the social functions are only composed of three types, namely, scenic spots, train/coach station, and entertainment districts. Making use of the call detail record (CDR) data from the cell phone network in Madrid, Soto and Frías-Martínez (2011b) computed the weekday-weekend daily aggregation

pattern, then performed unsupervised clustering to identify five clusters of similar activity patterns, which match with the land use types of industrial, commercial, nightlife, leisure, and residential in the city. The limitation lies in the lack of the validation analysis of the generated land use classification. Moreover, the inferred daily activity pattern may be biased as the CDR data are recorded using the location of the mobile phone tower instead of users' accurate geo-location.

Similarly, (Toole et al. 2012) applied the CDR data to infer dynamic land use in Boston metropolitan area. The data is more accurate as it contains the location information of the mobile phones, thus the spatial resolution is obtained at a resolution of 200 by 200 m. A supervised classification method is applied to infer land use types. The result shows that the daily distribution of activities has the potential of differentiating land use, especially if the residual activity pattern is used, where the circadian rhythm of life is eliminated from the absolute activity pattern. Moreover, (Yuan et al. 2012) integrated the data from two aspects: point of interest (POI) data in a region, and the human mobility data represented by GPS trajectory datasets from taxicabs. A topic-based inference model is applied to infer functions for each region, followed by an unsupervised learning clustering method to estimate region clusters. The result from the integration of two data sets outperforms that of solely POIs or GPS trajectory data, confirming the benefits of integrating the two datasets.

Currently, there is limited study on the use of social media check-in data to infer land use in urban areas. In virtue of the detailed categorical information of human behavior contained in the check-in data, the data provides a more effective and accurate way to investigate urban land use. Compared with CDR data set, GPS data set of vehicle trajectories, and traditional land use survey data, social media check-in data has some advantages:

1) Coverage: There has been an increase in the use of smartphones in recent years. In 2012, the U.S. saw a 55 % increase in smartphone subscriptions with about 98 million smartphone subscribers, representing nearly 42 % of all U.S. mobile users (ComScore 2012).
2) Accessibility: Most of the check-in information from location-based service website is shared by users and open to the public. Unlike the CDR data or GPS data of vehicle trajectories, which are privately owned and protected by agencies, the social media check-in information is more easily accessible.
3) Mobility feature: One of the most advantageous features of the check-in data, compared with CDR data or GPS data of vehicle trajectories is that it uses human individuals as mobile sensors, thus captures unique human mobility features.
4) Categorical activity information: Besides the information of physical location, social media check-in data also indicates the social intention (which is reflected in the activity categories) of the check-in, e.g. home, work, eating, entertainment, recreation, shopping, etc.

This paper is motivated by developing an efficient and accurate framework to apply the emerging check-in data in the area of land use inference. The detailed and categorical information on human behavior in the check-in data is useful in revealing

human mobility and activity patterns, which eventually contributes to the inference of land use. To infer land use, we have explored both unsupervised and supervised learning approaches. This paper contributes to the literature of land use inference in the following ways: (1) A new framework is proposed to infer land use by applying social media check-in data. (2) It reveals the underlying association between human activity pattern and urban land use pattern. (3) As of now, the proposed framework is a complementary technique to traditional land use inference approaches, especially to infer residential, commercial and recreational land use. (4) The dynamic change of land use over a period of time can be captured. (5) It is especially helpful for small cities that do not have or have limited land use information to infer the urban land use pattern.

The rest of the paper is organized as follows. Section 2 describes the source of check-in data and the process of retrieving the check-in data on-line. Section 3 discusses the data preprocessing methods. Section 4 presents the two approaches proposed for inferring land use types, and the empirical validation on New York City. Finally in Section 5, we provide concluding remarks and directions for future research.

2 Data Description

2.1 Data Collection

The dataset used in this research is collected from Twitter, where users can post short status messages up to 140 characters. The posted status message is referred to as "Tweets". "Tweets" supports the inclusion of geo-tags (latitude-longitude coordinates) as well as the third-party location-based services like Foursquare. Thus when Foursquare users "check-in" to a specific place, they can share the check-in information on Twitter in the form of a tweet. The format of a typical tweet with a "check-in" is shown as below:

> tweet (7913259124826)={189872633, ####, 7913259124826, Fri Jun 10 10:27:34+0000 2011, 40.7529422,-73.9780177, "I'm at Central Cafe & Deli (16 Vanderbilt Ave., New York) http://4sq.com/jMS87x"}

The original large-scale check-in dataset from a previous study by Cheng et al. (2011) is used in this research, while the venue category information of each check-in record is further collected to obtain the activity information. In the empirical case studies (Section 4), the processed data from New York City is extracted to perform the land use inference analysis. This process results in a dataset of more than 460,000 check-ins from 18,440 users for New York City.

2.2 Activity Category Classification

One of the major strengths of social media check-in data lies in the ability to identify individual activities. In each check-in record, an additional URL link (e.g. "http://4sq.com/jMS87x") is included to redirect the page to check-in venue in the location-based

service provider's website. This makes it possible to track the actual venue information by inquiry from the location-based service provider, e.g. Foursquare. By collecting the venue information of the data, we obtain 365 detailed venue categories. Furthermore, we classify each record into one of the nine activity categories: home, work, eating, entertainment, recreation, shopping, social service, education and travel related (Table 1).

3 Data Preprocessing

3.1 Data Preparation and Normalization

The raw check-in data only contains latitude-longitude coordinates and activity category information. In order to identify the land use types, a map of virtual grids is constructed by dividing the city map into square cells at the size of 200 by 200 m. The number of check-ins in each cell is aggregated by activity category, time period (8 time periods in a day) and day. Thus for cell (i, j) , we obtain a data tuple $T^{(i,j)}(c, t, D)$, in which $c \in \{1, 2, 3, \ldots 9\}$ denotes the activity category index representing home, work, eating, entertainment, recreation, shopping, social service, education, and travel related check-ins; $t \in \{1, 2, 3, \ldots 8\}$ denotes the time period in a day; and $D \in \{1, 2, 3, \ldots 7\}$ denotes 7 days in a week.

The size of the input information in the data tuple of each cell is quite large. It is unrealistic to use all the information to perform the land use inference. Hence we

Table 1 Activity category classification

Activity category (% of check-ins)	Type of visited locations
Home (2.90 %)	Home (private), Residential Building (Apartment/Condo)
Work (6.24 %)	Office, Co-working Space, Tech Startup, Design Studio
Eating (31.05 %)	Coffee Shop, Restaurant, Pizza, Burger, Bodega, Cafe, Diner, Sandwich, Bakery, Joint, Breakfast, Food, Bagel Shop, Steakhouse, Dessert Shop, etc.
Entertainment (20.27 %)	Pub, Entertainment, Venue, Nightclub, Bar, Theater, Club, Event Space, Stadium, Concert Hall, Dance Studio, Opera House, Casino, etc.
Recreation (9.80 %)	Park, Gym, Playground, Dog Run, Scenic Lookout, Zoo or Aquarium, Garden, Golf Course, Track, Field, Pool, Hiking Trail, Basketball Court, etc.
Shopping (10.75 %)	Supermarket, Store, Plaza, Pharmacy, Bookstore, Cosmetics Shop, Mall, Farmers Market, Boutique, Miscellaneous Shop, Automotive Shop, etc.
Social Service (7.83 %)	Hospital, Doctor's Office, Medical Center, Post Office, Church, Convention Center, Courthouse, City Hall, Police Station, Embassy or Consulate, etc.
Education (3.84 %)	Performing Arts Venue, University, College Academic Building, High School, Museum, Library, School, Student Center, Planetarium, etc.
Travel Related (7.31 %)	Highway or Road, Bridge, Bus Line, Taxi, Subway, Boat or Ferry, General Travel, Harbor or Marina, Ferry, etc.

further aggregate the data by weekdays and weekends, as the temporal patterns in weekday and weekend are significantly different:

$$Weekday:\ T_W^{(i,j)}(c,t,D_w=1)=\sum_{D=1}^{5}T^{(i,j)}(c,t,D),\ \forall c,t \tag{1}$$

$$Weekend:\ T_W^{(i,j)}(c,t,D_w=2)=\sum_{D=6}^{7}T^{(i,j)}(c,t,D),\ \forall c,t \tag{2}$$

As the information of a few activity categories are not useful or in some cases misleading for purposes of land use inference, we further shorten the activity categories from nine to seven after testing different combinations: home, work, eating, recreation, shopping and social service and travel related ($c_I \in \{1,2,3,5,6,7,9\}$). Entertainment, education related check-ins are found to be indecisive in land use inference, hence they are removed to reduce the dimension of the input data. Thus the raw input data tuple $T_R^{(i,j)}(c_I,t,D_w)$ for cell (i,j) obtained from the previous data aggregation process has 112 input scalar variables ($7\times 8\times 2=112$).

A two stage normalization process is applied to normalize the raw input data.

Stage 1: Due to the heterogeneity among check-ins, the number of check-ins for some activity categories (e.g. eating) is much larger than that of other activity categories (e.g. home). To normalize the difference among activity categories, each activity category's data is divided by the proportion of number of check-ins of each activity category over the overall number of check-ins, $p(c)$:

$$T_I^{(i,j)}(c_I,t,D_w)=\frac{T_R^{(i,j)}(c_I,t,D_w)}{p(c_I)},\ \forall c_I,t,D_w \tag{3}$$

Stage 2: The data tuple for cell (i,j), $T_I^{(i,j)}$ is further transformed into a 112 element vector $V_I^{(i,j)}$, and normalized to have zero mean and unit standard deviation. The transformation is shown as below:

$$V_N^{(i,j)}=\frac{V_I^{(i,j)}-E\left(V_I^{(i,j)}\right)}{\sqrt{Var\left(V_I^{(i,j)}\right)}} \tag{4}$$

The normalized input vectors $V_N^{(i,j)}$ are then used in the land use inference process.

3.2 Feature Selection

The input data used for classification have high dimensions (112 features). Data with high dimensionality often poses serious challenge in existing classification methods

(both unsupervised and supervised learning methods), which is known as "the curse of dimensionality" (Bishop 2006). A learning model tends to over-fit the data and degenerate performance by using a large number of features. Feature selection is one of the most frequently used techniques to reduce the dimensionality. It evaluates the relevance of features and removes the redundant ones. Feature selection usually leads to better learning performance, lower computational cost and better model interpretability (Alelyani et al. 2013).

In this study, we adopt the Laplacian Score (LS) technique for feature selection, which is a filter-based method independent from any classifier. The method is based on the intuition that data from the same class is more likely to be close to each other. The power of locality preserving (Laplacian score) is computed to evaluate the importance of each feature. Further, LS is very effective and efficient with respect to the dataset. The advantage of the LS is that it can be used for both unsupervised and supervised learning algorithms, thus it serves as an ideal base for the exploration of land use inference approaches presented in the later section. Detailed formulation and algorithm of LS can be found in He et al. (2006).

To apply the LS feature selection technique, we firstly compute the Laplacian Score for each of the 112 features, and then select the top 50 features with highest scores. The selected features are represented by thin vertical lines in Fig. 1.

4 Land Use Inference

Previous studies have applied both supervised learning algorithms (Toole et al. 2012) and unsupervised clustering algorithms (Soto and Frías-Martínez 2011a; Soto and Frías-Martínez 2011b; Yuan et al. 2012) to infer urban land use using large-scale geo-location data. Supervised learning algorithm requires actual land use data to train the model. For instance, Toole et al. (2012) implemented the Random Forest classification algorithm by utilizing the actual land use information. The major limitation of the supervised learning algorithm is the need of acquisition of ground truth land use information, which in some cases is not accessible or available.

On the contrary, the unsupervised clustering algorithms do not require ground truth information, thus it allows additional flexibility in land use inference. However, validating the land use inference can be a challenge in this case. In previous studies, Soto and Frías-Martínez (2011a) applied K-means clustering algorithm on mobile phone data in Madrid. Soto and Frías-Martínez (2011b) implemented the Fuzzy c-Means (FCM) algorithm to investigate land use types.

In this work, we explore the potential of using social media check-in data to infer urban land use by implementing both unsupervised and supervised learning algorithms. In real-world applications, these two methods can be used for different situations, such as:

1. For small cities with limited resources, the unsupervised clustering approach can be a useful tool to identify major land use patterns.
2. For megacities (e.g. New York City), the supervised classification method can be used as a complementary approach when combined with available ground truth data to generate more detailed classification results with higher accuracy. This can help the local agencies to keep track of the variations of land use pattern over time.

In the following sub-sections, we introduce the methodologies and inference results for each of these two approaches. The classification results are validated against the MapPLUTO data (NYCDCP 2013) provided by New York City Department of City Planning (NYCDCP). This data contains extensive land use and geographic data at the tax lot level in ESRI shape file format. To make the MapPLUTO data compatible with the inference results obtained at the same 200×200 m grid level. We re-generate an artificial 200×200 m grid reference and tag each cell with the dominant land use type over the land use layer. Moreover, since the land use information in New Jersey is not available, we only conduct the empirical validation based on the New York City land use information.

4.1 Clustering Inference Approach

To find the most appropriate unsupervised clustering algorithm for land use inference, various algorithms have been tested, including the standard hard partitioning methods, the fuzzy partitioning methods, and the more advanced clustering algorithms (PROCLUS and SUBCLU). In particular, the tested hard partitioning methods include K-means, K-medoid, and DTW (Dynamic Time Warping) based K-means. The tested fuzzy partitioning methods include FCM, Gustafson-Kessel Algorithm, and the Gath-Geva algorithm. For implementation, we developed Matlab codes using Fuzzy Clustering and Data Analysis toolbox (Balasko et al. 2005) to perform clustering analysis.

The standard fuzzy clustering algorithms do not work well in inferring land use, mainly due to the relatively high dimensionality of the input data. According to Winkler et al. (2011), the traditional fuzzy method like FCM (and similar algorithms) works quite well in the case of low dimensional data, but will fail in the case of high dimensional data. Similar results are also observed in Gustafson-Kessel Algorithm and Gath-Geva algorithm. On the other hand, the hard partitioning algorithms are found to have the best performance. Especially, compared with K-medoid and DTW based K-means, the result from K-means algorithm is found to have a much better match with the actual land used data provided in MapPLUTO.

In addition, to address the high dimensionality issue in the input data, two more advanced clustering algorithms for high dimensional data (namely, the projected clustering algorithm PROCLUS and the subspace clustering algorithm SUBCLU) are tested using the Weka based open subspace-clustering integration OpenSubspace (Müller et al. 2009). However, the former results in a large number of un-clustered cells while the latter one is computationally intractable due to the size of the land-use inference problem.

Among all the clustering algorithms tested, K-means algorithm demonstrated the best performance in the land use inference (more details are provided in Section 4.2). K-means algorithm is one of the most popular iterative descent clustering methods (Hastie et al. 2009). For a given cluster assignment C , the K-means algorithm minimize the within-cluster sum of squares, mathematically,

$$C^* = \min_C \sum_{k=1}^{K} \sum_{C(i)=k} \left\| x_i - \bar{x}_k \right\|^2 \tag{5}$$

Where $\overline{x}_k = (\overline{x}_{1k}, \overline{x}_{2k}, \ldots, \overline{x}_{pk})$ is the mean vector associated with the kth cluster. There are mainly two steps in the K-means algorithm:

Step 1: Assignment Step Given a current set of means $\{m_1, m_2, \ldots, m_K\}$, assign each observation to the closest (current) cluster mean, which is:

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} \|x_i - m_k\|^2 \tag{6}$$

Step 2: Updating Step Calculate the new mean and set it to be the centroid of the observations in the cluster.

$$m_i = \frac{1}{N_{C(i)}} \sum\nolimits_{x_i \in C(i)} x_i \tag{7}$$

Where $N_{C(i)}$ is the number of observations that have the label $C(i)$. Step 1 and Step 2 are iterated until the assignments converge to the preset threshold of convergence.

4.1.1 Optimal Number of Clusters

One limitation of K-means algorithm is that the number of clusters, K, must be provided as an input. The optimal number of clusters results in most compact and farthest separated clusters for the data, where the variance of members within the cluster is minimized (most compact) while the mean between different clusters is maximized (farthest separated).

There are many validity measures for the determination of optimal number of clusters in the literature, such as intra-cluster and inter-cluster ratio test (Ray and Turi 1999; Soto and Frías-Martínez 2011b), Davies-Bouldin index (Davies and Bouldin 1979), Dunn's index (Dunn 1973), Alternative Dunn Index (Abonyi and Feil 2007), and Xie and Beni's Index (XB) (Xie and Beni 1991). Here we choose Dunn's index (DI) and Xie and Beni's Index (XB) as the validation measures, as these measures are capable of identifying the most "compact and far separated" clusters.

For any partition $U \leftrightarrow C : C_1 \cup \cdots C_i \cdots \cup C_k$, where C_i denotes the centroid of the i^{th} cluster of the partition, and there are k clusters. It is important to note that the data applied in the validation has been normalized. The Dunn's index, denoted as $DI(U)$ is defined as:

$$DI(U) = \min_{1 \le i \le k} \left\{ \min_{\substack{1 \le j \le k \\ j \ne i}} \left\{ \frac{\min\limits_{x \in c_i, y \in c_j} \|x - y\|}{\max\limits_{X_n \in C_i} \|X_n - C_i\|} \right\} \right\} \tag{8}$$

where $\|x - y\|$ denotes the inter-cluster distance of observation x and y, while $\|X_n - C_i\|$ denotes the intra-cluster distance of observation X_n and centroid C_i, where observation X_n belongs to cluster C_i.

The Xie and Beni's Index, denoted as $XB(U)$, is defined as:

$$XB(U) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{N}\left(\mu_{ij}\right)^{m}\left\|X_j - C_i\right\|^2}{N \min_{1\le i,\, j\le k,\; j\ne i}\left\|C_i - C_j\right\|^2} \tag{9}$$

where μ_{ij} denotes the membership of data point j in cluster i, while m is an adjusted power parameter for the weight of the membership. As the K-mean algorithm is deterministic, we set $m = 1$; and $\mu_{ij} = 1$, if $X_j \in C_i$; 0, otherwise . N denotes the size of data set.

An ideal partition minimizes the intra-cluster distance (most compact) while maximizing the inter-cluster distance (farthest separated), thus the greater value of DI corresponds to a better performance of partition. For the case of XB, the smaller the value, the better the partition. Here we run tests from K=3 to K=9. Results for the DI and XB values are shown in Table 2:

As shown in Table 2, among the partition cases where $K \in \{3 \cdots 9\}$, $K = 5$ is the maximum value for the DI test, and the second smallest value for the XB test. Although in the XB test, the partition case $K = 3$ has the minimum value, only three types of land use may not be sufficient for urban land use identification. Moreover, the DI test suggests $K = 3$ is not the optimal number of clusters. Thus the most consistent optimal K-value for DI and XB validation measures is $K = 5$.

4.1.2 Inference Results

The clustering results are obtained by using K-means algorithm with five clusters. As K-means algorithm may converge to a local minimum, we run the randomly initialized K-means algorithm repeatedly for 100 times and select the result that yields the minimum objective function value.

Figure 1 presents the clustering centers for the identified five clusters, in which the thin vertical lines indicate all selected features that are used for clustering. Figure 2 shows the comparison between the geographical representation of the inferred and actual land use patterns. Cells with no check-in data are excluded in the land use

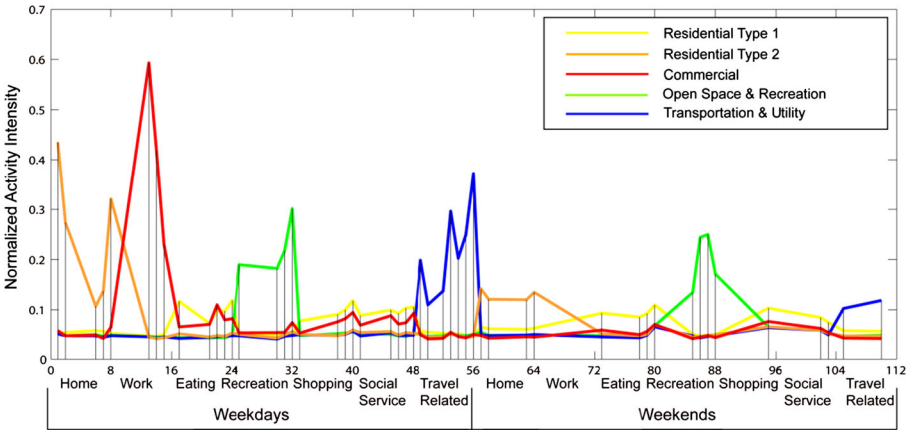

Fig. 1 Clustering centers for land use inference

Table 2 Validation results for the determination of *K*-value

K value	3	4	5	6	7	8	9
DI	3.334	3.451	3.609	3.251	3.012	2.995	2.946
XB	0.010	0.019	0.012	0.019	0.020	0.020	0.031

inference and left blank in Fig. 2. Most of the empty cells are located in the river and bay area, others scattered all over the whole map.

It is seen that the activity category information plays a crucial role in the identification of land use. For example, open space and recreation land use type (green) and the transportation and utility land use type (blue) are easily identified since they are largely determined by the recreation and travel related check-in activities. The commercial land use type is identified as the cluster represented in red, which are classified mainly based on work related activities, where a significant peak appears in weekday afternoon and night work activities. One interesting finding is that the clustering algorithm finds two different clusters corresponding to residential land use. The residential type 1 (yellow) corresponds to a "flat" activity pattern, where the normalized activity intensity remain low with a few minor peaks in eating, shopping and social service activity categories. This agrees with the fact that there are relatively fewer check-in activities in residential areas since there are no major attractions to check-in behaviors. On the other hand, the other cluster (residential type 2) is largely determined by late night home activities for both weekdays and weekends. Such activities definitely suggest residential land use type, but also represents more about nightlife characteristics in residential areas.

Since both residential type 1 and type 2 correspond to residential land use type, thus we merge these two clusters for further investigation. From the comparison in Fig. 2, we see that the inference result demonstrates a very similar land use pattern compared with the actual land use pattern. For land use type of "residential", the inferred result shows that the majority of Brooklyn, Queens and Staten Island are residential areas. In Manhattan Island, areas on both sides of Central Park are identified as residential area. For land use type of "commercial", the inferred result shows two large commercial

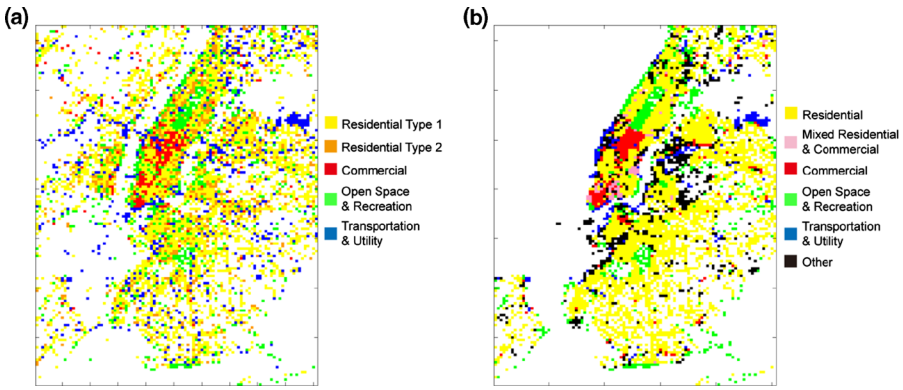

Fig. 2 Comparison of **a** inferred land use pattern using clustering and **b** actual land use pattern from MapPLUTO data (NYCDCP)

areas in Manhattan Island. One is located in the central part of Manhattan, areas between 8th Avenue to Park Avenue. The other one is located in the southern part of Manhattan, where Wall Street and other financial institutions are located. For the land use type of "open space and recreation", we see that the Central Park, areas along Henry Hudson Parkway, Battery Park and East River Park in Manhattan are correctly inferred as open space and recreation. In Brooklyn, the Prospect Park, Green Wood Cemetery, Marine Park are also correctly identified. For the land use type of "transportation and utility", we see that the LaGuardia airport is clearly identified, however, there is some tendency to over predict.

As there are very few industrial/manufacturing related check-ins, the industrial and manufacturing land use is not inferred. A major industrial/manufacturing area in the northwestern part of Queens is blank in the inference result. The mixed residential and commercial land use type is also not inferred. It is difficult to use clustering algorithm, especially hard partitioning algorithms to correctly classify the mixed residential and commercial land use since the boundary of determining residential, commercial and mixed land use is often quite vague. The lack of inference of the industrial/manufacturing and other land use is a potential drawback of using the unsupervised clustering approach.

4.1.3 Validation

To have a better understanding of the clustering results, we perform a quantitative evaluation of the inferred land use pattern against the ground truth information. Since land use type like the industrial/manufacturing land use is not inferred, we only perform the validation on land use types of residential (Res), commercial (Com), open space and recreation (Prk) and transportation and utility (Trp). The validation results are presented in Table 3 and the spatial distribution of the correct and incorrect inferred cells are presented in Fig. 3.

Table 3 presents the confusion matrix and corresponding proportions of each land use type. In the confusion matrix, each column represents the instances in the inferred land use type, while each row represents the instance in the actual land use

Table 3 Validation results of the inferred land use

Confusion Matrix (proportion)				
Actual\Predicted	Res	Com	Prk	Trp
Res	2115 (0.787)	63 (0.023)	238 (0.089)	272 (0.101)
Mix	67 (0.744)	5 (0.056)	12 (0.133)	6 (0.067)
Com	97 (0.422)	85 (0.370)	22 (0.096)	26 (0.113)
Prk	167 (0.337)	3 (0.006)	261 (0.526)	65 (0.131)
Trp	86 (0.393)	11 (0.050)	30 (0.137)	92 (0.420)
Other	399 (0.623)	40 (0.062)	92 (0.144)	110 (0.172)
Total number of instances	4364			
Overall accuracy	0.6560			

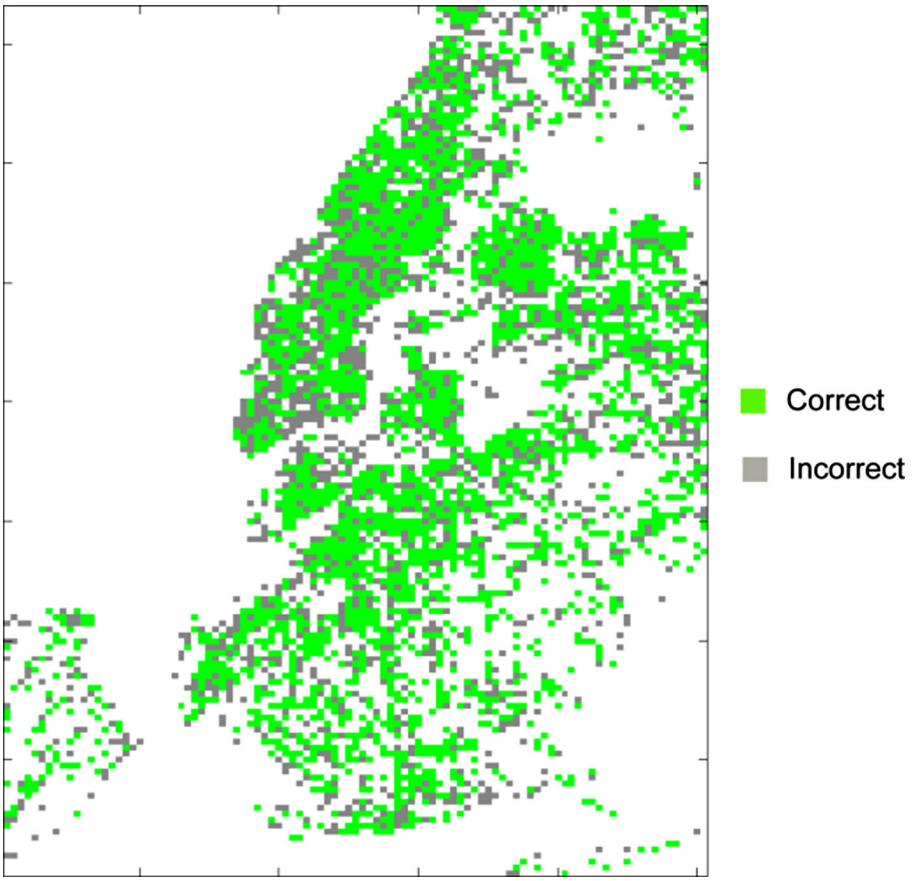

Fig. 3 Spatial distribution of correctly and incorrectly inferred cells

type. We observe an overall accuracy (defined as the fraction of correctly classified cells) of 65.60 %, which is better than a previous study using supervised learning method by Toole et al. (2012) where they use CDR data and obtain an accuracy of 54 %. This suggests the additional activity related information do contribute to more accurate prediction on urban land use, even using clustering algorithm without any prior knowledge. Also, compared with the dominant residential land use, which account for 62.5 % of all cells in the study region, it suggests that the clustering result is better than classifying all cells as residential land use type.

The results show that we achieve a 78.7 % accuracy in inferring residential land use type, which account for almost half of all labeled cells (2115 out of 4364). We also achieve reasonable accuracy in land use type of open space and recreation, which is 52.6 %. On the other hand, the inference accuracy for commercial and transportation and utility land use types are 37 and 42 %, which are relatively low. The accuracy of the inference results can be further improved by using a larger number of check-ins. Since home, work and travel related check-ins take very few proportions in the entire dataset, however, these three activity categories are important indicator for residential, commercial and travel related land use types. Also, as a grid reference system is

constructed over the actual land use layer, and land use type of each grid cell is determined by the dominant land use type (the one with highest area proportion). It is possible that this approximation might contribute to the difference as well.

4.2 Supervised Learning Approach

In this section, we investigate the performance of using supervised learning algorithms to infer urban land use. Different from the previous clustering approach, we use data with known labels to train the model. Supervised learning algorithms usually have better performance compared with clustering approach due to the introduction of extra information. However, the trade-off is that part of the ground truth information is needed for training.

Three widely used supervised learning classifiers are tested in this paper, namely, Naïve Bayes method, support vector machine (SVM) and random forest. The Naïve Bayes is a simple probabilistic classifier based on Bayes' Theorem, which assumes conditional independence among attributes. SVM is another popular method, which classifies data by maximizing the perpendicular distance between the decision boundaries and the closest data points (support vectors) (Bishop 2006). On the other hand, the random forest (Breiman 2001) approach is an ensemble method, which uses a combination of randomly generated decision tree classifiers to increase accuracy. The outline random forest algorithm can be summarized in two steps: (1) generate an ensemble of decision tree classifiers using a random selection of attributes, and the best split on the selected attributes is used to split the node in the decision tree; (2) during classification, each decision tree gives a classification ("vote"), and the classification having the most votes is returned. The random forest algorithm is proven in practice to be more robust to errors and scales well for dataset with large number of attributes. For detailed discussions about the theoretical background and other features of random forest, please refer to (Breiman 2001).

The same data with 50 selected features in previous section is used to test the performance of the three supervised learning algorithms. The input data contains a total of 5416 instances (cells with check-ins), with 4364 cells have labels. The remaining 1052 unlabeled cells are mainly located in New Jersey, where the detailed land use data was not readily available. We randomly split 50 % of the labeled cells for training, in which the ground truth land use types are provided, and the other 50 % labeled cells for testing.

We used Weka (Hall et al. 2009) to implement the Naïve Bayes and SVM classifiers. For random forest, the MATLAB implementation toolbox developed by Jaiantilal[1] is used to perform both testing and prediction. The radial basis kernel is used for SVM, and 100 random decision trees with the maximum of 25 features for each tree are used for random forest. The validation results on the test set for the three supervised learning algorithms are summarized in Table 4.

F-measure (also refer as F_1) is used to evaluate the classification accuracy on the test set, which is a commonly used accuracy measure in data mining. F_1 is computed as the harmonic mean of precision (percentage of tuples that are classified as positive are actually positive) and recall (percentage of positive tuple that are classified as

[1] https://code.google.com/p/randomforest-matlab/

Table 4 Testing results for supervised learning algorithms (*F*-measure)

Class	Naïve Bayes	SVM	Random forest
Residential	0.432	0.757	**0.771**
Mixed residential & commercial	**0.073**	0	0
Commercial	0.264	0	**0.46**
Open space & recreation	**0.468**	0	0.415
Transportation & utility	**0.181**	0	0.169
Other	0.074	0	**0.121**
Correctly classified instances (percentages)	690 (31.39 %)	1339 (60.92 %)	**1377** (**64.14**%)

positive), which is calculated as $F_1 = 2\ (precision \times recall)\ /(precision + recall)$. A higher F_1 score suggests a better classification result. Table 4 shows that that the random forest classifier outperforms the other two supervised learning algorithms in terms of classification accuracy. The random forest classifier provides a result that has the highest F_1 score for residential, commercial and other land use types. For the other two tested methods, Naïve Bayes classifier finds a very poor classification results with an accuracy of only 31.39 %, while the SVM tends to classify all cells as the residential land use type. Thus we base our later analysis only on random forest algorithm. It is observed that the overall accuracy of random forest on the test set is 64.14 %, which is similar to the clustering approach (65.60 %). However, it should be noted that these two accuracy values are not directly comparable, since supervised learning algorithm predicts more land use types that match exactly with the classifications in the ground truth data, while clustering approach is only able to find a subset of them. Hence the supervised learning approach may be more useful for practitioners.

4.2.1 Inference Result

To evaluate the predictive power of the random forest classifier, we use the previously trained model to predict the land use types of the entire input dataset. Although it will be more preferable to use a different set of data to evaluate the inference results, we perform our validation on the same dataset that combines all of the training, testing and unlabeled cell data because of the limited amount of data. In this test, the land use label of each cell is assumed unknown and will be predicted from the model. The prediction results are then evaluated against the ground truth. It should be noted that the evaluating accuracy of the test will be inflated due to the involvement of the training data. However, this can be perceived as the prediction accuracy if additional ground truth information is provided along with the input data in land use inference.

Figure 4 shows the comparison between the predicted land use pattern and the actual land use pattern. The result shows a much better match compared with the previous clustering approach. Furthermore, all the six land use types are classified while in the clustering results, we can only identify four land use types. Figure 5 provides a better illustration of the correct and incorrect cells in the study region. It is observed that the majority of the cells are predicted correctly with a few incorrectly classified cells

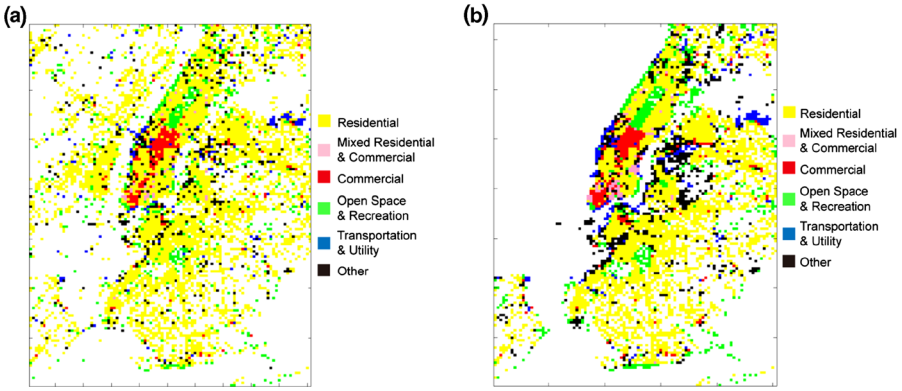

Fig. 4 Comparison of **a** predicted land use pattern using random forest classifier and **b** actual land use pattern from MapPLUTO data (NYCDCP)

scattered across the region. Table 5 presents the confusion matrix of the prediction results and the accurately predicted instance proportions of each land use type. All of the land use types are successfully predicted with reasonable accuracy. For the

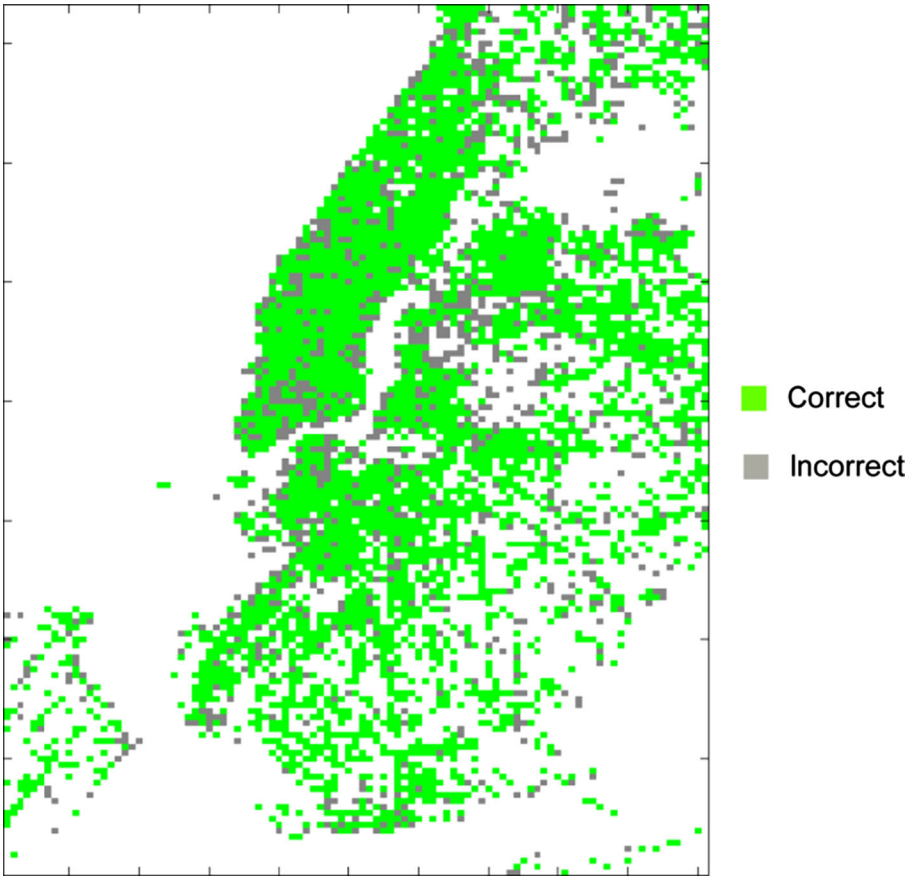

Fig. 5 Spatial distribution of correctly and incorrectly predicted cells using all data

Table 5 Confusion matrix of the Random Forest predictions on all data

Confusion Matrix (proportion)						
Actual\Predicted	Res	Mix	Com	Prk	Trp	Other
Res	2521 (0.938)	1 (0.000)	19 (0.007)	79 (0.029)	13 (0.005)	55 (0.021)
Mix	37 (0.411)	46 (0.511)	1 (0.011)	3 (0.033)	0 (0)	3 (0.033)
Com	63 (0.274)	0 (0)	158 (0.687)	1 (0.004)	2 (0.009)	6 (0.026)
Prk	163 (0.329)	0 (0)	3 (0.006)	314 (0.633)	5 (0..010)	11 (0.022)
Trp	94 (0.429)	0 (0)	0 (0)	17 (0.078	97 (0.443)	11 (0.050)
Other	297 (0.463)	1 (0.002)	7 (0.011)	26 (0.041)	12 (0.019)	298 (0.465)
Total number of instances		4364				
Overall accuracy		0.7869				

residential land use type, the prediction accuracy is as high as 94 %. For commercial and open space and recreation land use, more than 60 % of cells are correctly predicted. For the rest land use types, all of them have an accuracy more than 44 %. The overall accuracy of the prediction result is 78.69 %. As mentioned earlier, this accuracy is not directly comparable to the accuracy of 65.6 % in the clustering inference approach, due to different classifications. Moreover, the way that training data are used for both training and testing inflates the overall accuracy level of the prediction results. However, the results suggest that, if we incorporate part of the ground truth information (in this case 50 %) with social media check-in data, we can obtain a powerful tool to predict the land use pattern of the whole area with good accuracy.

4.2.2 Information and Accuracy Trade-Off

Since the introduction of extra ground truth information for training will always help to improve the classification accuracy in the supervised learning method. A valuable question to investigate is: what is the trade-off between the prediction accuracy and the amount of extra ground truth information provided? In actual implementation of the supervised learning approach, the ground truth information might be only partially available. How to utilize the available information to make the most accurate inference is an important topic.

We conducted an empirical test on the prediction accuracy with different levels of ground truth information provided. We randomly select 10~80 % labeled data as the training set, and the rest as the test set. The accuracies are evaluated both against the test data and the entire labeled set. Figure 6 shows the results from the tests. It is observed that including more data does not improve the accuracy much when validated against the test data. This suggests that introducing more training data does not significantly improve the predictive power for new data using the random forest algorithm. The results show that even 30 % training data is sufficient to achieve a reasonable accuracy on classifying the test data. However, the prediction accuracy for the entire dataset does improve when more ground truth information is included. It is observed that a prediction accuracy of more than 70 % on the overall dataset is achieved by

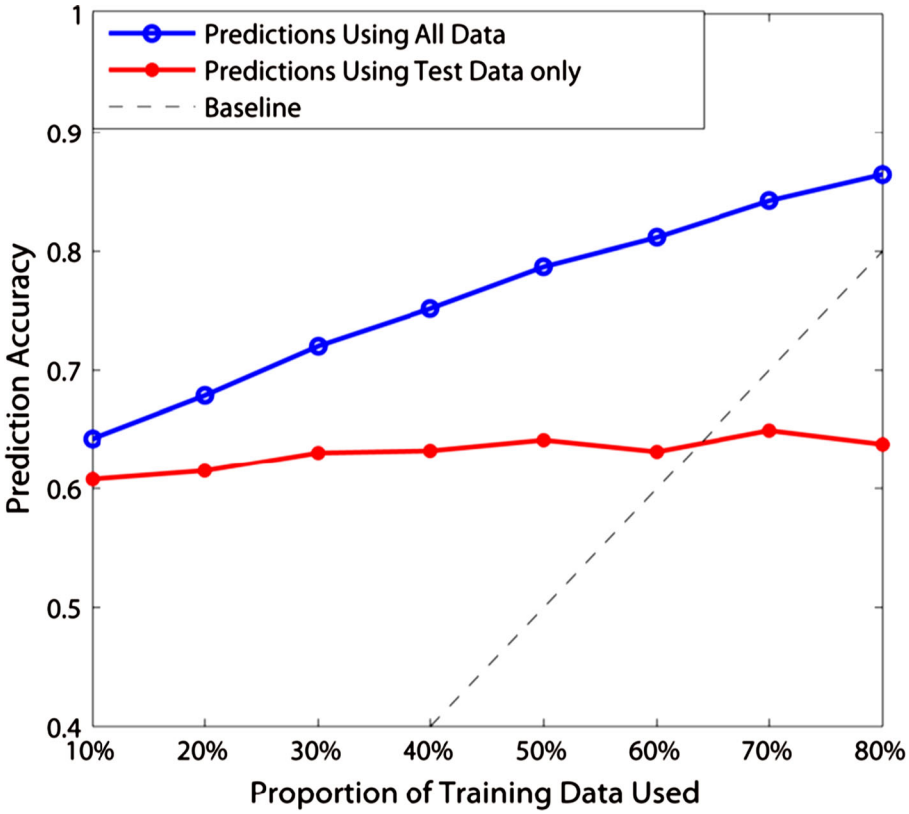

Fig. 6 Trade-off between land use prediction accuracy and the level of information provided

incorporating only 30 % labeled data as training set. By incorporating more than 50 % of training data, an accuracy level of 80 % is obtained. However, the prediction becomes inefficient when more than 70 % of ground truth information is used. As the gap between the base line (percentage of accurate data used) and the overall prediction accuracy quickly decreases, the benefit becomes smaller to introduce extra ground truth information. This result shows the potential of using the combination of limited ground truth information and social media check-in data in land use inference. High quality inference results can be obtained when only 30 % of ground truth information is provided. The rich human activity pattern contained in the social media check-in data can be an ideal complementary information source to reveal the whole picture of a city's land use pattern.

4.3 Comparison Between Unsupervised and Supervised Approach

In previous sections, we explored both the unsupervised clustering approach and the supervised learning approach of urban land use inference. Both approaches have their own advantages and limitations.

For unsupervised clustering approach, the major advantage is that no ground truth information is needed. However, limitations also exist, such as (a) extra effort in identifying land use type from the clustering results; (b) may not infer all the desired

land use types; (3) relatively lower accuracy compared with supervised learning algorithm with extra ground truth information provided.

For supervised learning approach, the requirement for all or part of the ground truth information is a major drawback, as the information may not be available in certain urban areas. However, the high accuracy in predicting land use types makes it a particularly helpful tool for local planning agencies to track the variations of urban land use patterns over time. The dynamic changes of urban land use pattern is well known to play an key role in the urban planning decision making (Marchal 2005; Pfaffenbichler et al. 2008) and determining the degree of urban sprawl (Sun et al. 2007). Given the check-in data of an enough long observation period, we can break the time period into sub-periods and perform the same technique to check the variation of land use over time. Compared with the static traditional land use inference approach, this new approach offers an interesting and reliable view of the dynamic changes of land use, and greatly enhances our understanding of the changes in an urban area. It will show greatest value for urban planners in applying the proposed approach at regional level, where there is a mixed level of jurisdictions and availability of georeferenced land use data. In these cases, the ground truth land use data that are partially available can be used as training data, and the proposed approach can be applied to predict the land use pattern for the entire region. Another advantage of supervised learning approach is that the predictions of land use types are more detailed, and practitioners can have full control on the type of land use to be classified by introducing proper ground truth data (training labels).

In spite of the limitations in both of the unsupervised and supervised learning techniques, the proposed modeling framework provides an innovative approach to infer urban land use from social media data. Moreover, the accuracy of the inferred results can be greatly improved by collecting more number of check-ins in the future.

5 Conclusion

In this paper, we propose a novel land use inference framework by making use of the social media check-in data. Two different land use inference approaches are proposed, namely, an unsupervised clustering approach using K-means algorithm and a supervised learning approach using random forest classifier. The unsupervised clustering approach is more suitable for small cities where the actual land use data is not available; while the supervised learning approach can be used to readily track the dynamic changes of urban land use, or predict urban land use types for the entire area when only part of the ground truth data is available.

Both of these two approaches are tested using half a million social media check-in data in New York City. The land use information in MapPLUTO data provided by New York City Department of City Planning is used as the ground truth for validation. The numerical evaluations show that both approaches provide land use inference results with reasonable accuracy at the grid level of 200 by 200 m. For unsupervised clustering approach, the results achieve an overall accuracy of 65.6 %. For supervised learning approach, by introducing 50 % of the ground truth information, the result attains an overall accuracy of 78.69 %. The encouraging result indicates the existence of the inherent linkage between the human urban activity pattern (as revealed in the social

media check-in data) and the underlying land use structure. The land use inference accuracy can be further improved by using more data, as a lot of cells have only a limited number of data records in the current dataset (only 1 year's data is available). All of these results have demonstrated the potential of using social media data as a complementary data source in maintaining urban land use information.

There are some limitations for this framework. First, due to the limitation of data, certain land use type (e.g., industrial/manufacturing) is currently not inferable. Second, the spatial distribution of social media check-in data is highly heterogeneous. That is, data is mostly concentrated with big cities, and small cities and rural area have very few data. Future research can be done to further investigate the possibility of inferring land use types with scarce check-in data. Furthermore, it will be meaningful to investigate the transferability of the model among different urban areas. This will be helpful to answer the important research question that if the inherent association between human activity pattern and urban land use pattern is universal across different urban areas.

References

Abonyi J, Feil B (2007) Cluster analysis for data mining and system identification. Springer, London

Alelyani S, Tang J, Liu H (2013) Feature selection for clustering: A review. Data Clust Algorithm Appl, CRC Press

Balasko B, Abonyi J, Feil B (2005) Fuzzy clustering and data analysis toolbox. http://www.abonyilab.com/software-and-data/fclusttoolbox

Barnsley MJ, Barr SL (1996) Inferring urban land use from satellite sensor images using kernel-based spatial reclassification. Photogramm Eng Remote Sens 62(8):949–958

Bishop CM (2006) Pattern recognition and machine learning (Information Science and Statistics), 1st edn. Springer-Verlag New York, Inc, Secaucus

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Cheng Z et al. (2011) Exploring millions of footprints in location sharing services. AAAI ICWSM, 2010(Cholera)

ComScore, Inc (2012) 2012 mobile future in focus. *ComScore, Inc.* https://snaphop.com/2012-mobile-marketing-statistics/

Davies D, Bouldin D (1979) A cluster separation measure. IEEE Trans Pattern Anal Mach Intell 1(2):224–227

Dunn JC (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. J Cybernet 3(3):32–57

González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explor Newsl 11(1):10–18

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, (Springer Series in Statistics), 2nd edn. Springer, New York

He X, Cai D, Niyogi P (2006) Laplacian score for feature selection. Adv Neural Inf Process Syst 18:507

Marchal F (2005) A trip generation method for time-dependent Large-Scale Simulations of Transport and Land-Use. Netw Spat Econ 5:179–192

Mesev V (1998) The use of census data in urban image classification. Photogramm Eng Remote Sens 5:431–438

Moran MS, Inoue Y, Barnes EM (1997) Opportunities and limitations for image-based remote sensing in precision crop management. Remote Sens Environ 61(3):319–346

Müller E, Günnemann S, Assent I, Seidl T (2009) Evaluating clustering in subspace projections of high dimensional data. In Proc. 35th International Conference on Very Large Data Bases (VLDB 2009), Lyon, France

New York City Department of City Planning (NYCDCP) (2013) MapPluto. http://www.nyc.gov/html/dcp/html/bytes/dwn_pluto_mappluto.shtml#mappluto

Pfaffenbichler P, Emberger G, Shepherd S (2008) The integrated dynamic land use and transport model MARS. Netw Spat Econ 8(2–3):183–200

Qi G, Li X, Li S, Pan G, Wang Z, Zhang D (2011) Measuring social functions of city regions from large-scale taxi behaviors. In the proceeding of Ninth Annual IEEE International Conference on Pervasive Computing and Communications, PerCOM, 384–388

Ray S, Turi RH (1999) Determination of number of clusters in k-means clustering and application in colour image segmentation. In ICAPRDT

Schmit C, Rounsevell MDA, La Jeunesse I (2006) The limitations of spatial land use data in environmental analysis. Environ Sci Pol 9(2):174–188

Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. Science (New York, NY) 327(5968):1018–1021

Soto V, Frias-Martinez E (2011a) Robust land use characterization of urban landscapes using cell phone data. In 1st Workshop on Pervasive Urban Applications, in conjunction with 9th Int. Conf. Pervasive Computing, June 2011

Soto V, Frías-Martínez E (2011b) Automated land use identification using cell-phone records. In Proceedings of the 3rd ACM International Workshop on MobiArch - HotPlanet'11, 17. ACM Press, New York

Sun H, Forsythe W, Waters N (2007) Modeling urban land use change and Urban Sprawl: Calgary, Alberta, Canada. Netw Spat Econ 7(4):353–376

Toole JL, Ulm M, González MC, Bauer D (2012) Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp'12, 1. ACM Press, New York

Winkler R, Klawonn F, Kruse R (2011) Fuzzy c-means in high dimensional spaces. Int J Fuzzy Syst Appl (IJFSA) 1(1):1–16

Xie XL, Beni G (1991) A validity measure for fuzzy clustering. IEEE Trans pattern anal mach intell 13(8): 841–847

Yang X, Lo CP (2002) Using a time series of satellite imagery to detect land use and land cover changes in the Atlanta, Georgia Metropolitan Area. Int J Remote Sens 23(9):1775–1798

Yuan J, Yu Z, Xing X (2012) Discovering regions of different functions in a City using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'12, 186. ACM Press, New York