



Article

# Social Sensing for Urban Land Use Identification

Adindha Surya Anugraha , Hone-Jay Chu \* and Muhammad Zeeshan Ali

Department of Geomatics, National Cheng Kung University, Tainan City 701401, Taiwan;  
ssqc.libz@sinarmas-agri.com (A.S.A.); p68077064@mail.ncku.edu.tw (M.Z.A.)

\* Correspondence: honejaychu@geomatrics.ncku.edu.tw

Received: 30 July 2020; Accepted: 13 September 2020; Published: 15 September 2020



**Abstract:** The utilization of urban land use maps can reveal the patterns of human behavior through the extraction of the socioeconomic and demographic characteristics of urban land use. Remote sensing that holds detailed and abundant information on spectral, textual, contextual, and spatial configurations is crucial to obtaining land use maps that reveal changes in the urban environment. However, social sensing is essential to revealing the socioeconomic and demographic characteristics of urban land use. This data mining approach is related to data cleaning/outlier removal and machine learning, and is used to achieve land use classification from remote and social sensing data. In bicycle and taxi density maps, the daytime destination and nighttime origin density reflects work-related land uses, including commercial and industrial areas. By contrast, the nighttime destination and daytime origin density pattern captures the pattern of residential areas. The accuracy assessment of land use classified maps shows that the integration of remote and social sensing, using the decision tree and random forest methods, yields accuracies of 83% and 86%, respectively. Thus, this approach facilitates an accurate urban land use classification. Urban land use identification can aid policy makers in linking human activities to the socioeconomic consequences of different urban land uses.

**Keywords:** urban land use map; human behavior; remote sensing; social sensing; decision tree; random forest; accuracy assessment

## 1. Introduction

Urban land use carries crucial information for human activity for urban planning and economic analysis, as well as hazard and pollution management [1]. Issues associated with land use have attracted considerable interest from communities because of the central role of land use as a cause and a consequence of human activity. The demand for the utilization of urban land use maps by urban authorities, researchers, and citizens has steadily increased, especially in fast developing regions where the timely acquisition of up-to-date land use information is crucial [2].

Land use is a human–environment system consisting of the relationship between human activities and socioeconomic, environmental, and demographic characteristic components of urban land. Uncovering land uses from only remotely sensed imagery, however, is rather difficult [3,4]. The spectral reflectance is not directly related to socioeconomic features and objects. Social sensing broadly refers to a set of sensing and data collections where data are collected from humans or devices on their behalf [4]. Social sensing data in cities such as traffic information and crowd movement have become an important means of analyzing urban development issues. For urban areas, social sensing intensity is usually high in urban centers and human activities are relatively dense. However, social sensing density and activity intensity in suburbs are relatively low. Such social sensing data include bike trajectories, taxi trajectories, smart card records in public transportation systems, social media or social networking data, and so on [4–6]. Previous studies have used social sensing data for mapping dynamic urban land use patterns [7,8], and traffic and transportation systems [5,6]. Human mobility can be

detected from location-aware devices such as GPS (global positioning system) receivers and mobile phones for collecting large volumes of individual trajectory data. Using social media locations [5], mobile telephone positioning [6] and GPS tracking data from taxi drivers [7] can aid understanding of the dynamics of human spatial mobility. In this study, bicycle and taxi information (pick-up and drop-off) are used to investigate user behavior and human patterns in a city. Data from remote sensing (satellite image), and social sensing (bicycle trips and taxi routes) were collected to estimate urban land uses. Especially from the social sensing data, the spatial and temporal distribution of human activities, combined with the current telemetry classification, can be used to understand the spatial and temporal distribution of land use and human activities. Using solely remote sensing data may not achieve high classification accuracy in an urban area, e.g., commercial and residential area. Previous studies have considered social sensing in land use classification [4,9,10]. In the current study, social sensing uncertainty removal or data cleaning is considered. Social sensing with data cleaning provides highly reliable and detailed information on human activities in urban structures and thus facilitates the assignment of socioeconomic functions to different zones.

This study collects sensing data on bicycle information, taxi routes and remote sensing to estimate urban land use using data cleaning and machine learning. According to the social sensing data, the spatial and temporal distribution of human activities enrich the understanding of the spatial and temporal distribution of human activities. Eventually, the machine learning of remote sensing and social sensing ensures a highly accurate urban land use classification.

## 2. Materials

### 2.1. Study Area

The study area consists of the districts of Manhattan, Brooklyn, and Queens in New York City, United States of America (USA). These districts are selected because of their characteristics that are suitable for urban land use classification. Manhattan is not only the economic and administrative center of the city but also the most densely populated district in New York City. Brooklyn has an ever-changing landscape and a spike in real estate development. Queens offers an ethnically diverse urban area. The study area covers a square area of approximately 139 km<sup>2</sup> with residential areas, office areas, industrial areas, entertainment areas, open spaces, and rivers. Brooklyn borders the borough of Queens, and both districts are separated from Manhattan by the East River. Manhattan comprises many office and entertainment buildings because it is the administration center and is the most densely populated district in the city. Central Park, which is an iconic open space, is located in Manhattan. Brooklyn has many residential buildings because of a spike in real estate development. Many industrial buildings are also built in Brooklyn to meet human needs. Queens has a development balance in terms of building type because it is an ethnically diverse urban area.

### 2.2. Datasets

Two datasets are used in this study: remote sensing data and social sensing data. Both datasets are openly accessible but have different sources, meanings, and information. Sentinel remotely sensed imagery belongs to the remote sensing data, whereas bike and taxi data belong to the social sensing data.

#### 2.2.1. Remote Sensing Data

The sentinel-2A remote sensed imagery of New York on 7 June 2016 was used as the remote sensing data source. Thirteen spectral bands can be found in the sentinel-2A imagery, and they range from the visible and the near-infrared to the shortwave infrared at spatial resolutions ranging from 10 m to 60 m on the ground. The sentinel-2A imagery has spatial resolutions of 10, 20, and 60 m. Each band in the different spatial resolutions of the sentinel-2A imagery has a different function (e.g., the bands for the 10 m spatial resolution are used for basic land cover classification, the bands for the 20 m spatial resolution are used to enhance the retrieval of geophysical parameters, and the bands for the 60 m

spatial resolution are used for atmospheric correction and cirrus-cloud screening). The red, green, blue, and NIR bands (10 m spatial resolution) from Level L1C image are used in this study.

### 2.2.2. Social Sensing Data

Bike data and taxi data are chosen as the social sensing data in this study. Social sensing data are used to understand human behavior in the study area for the benefit of urban authorities and citizens. The bike and taxi datasets are openly accessible, and the retrieved data cover one month. The whole month of June in 2016 was selected as the study period. The bike dataset for this study was retrieved from the CitiBike Bikeshare System of New York City.

Each bike trip record contains the following fields:

- Trip ID: the unique ID of the trip
- Bike ID: the unique ID of the bike
- Departure Station ID: the unique ID of the station where people rent the bikes
- Departure Station Coordinate: the coordinates of the station where people rent the bikes
- Arrival Station ID: the unique ID of the station where people return the bikes
- Arrival Station Coordinate: the coordinates of the station where people return the bikes
- Departure Time: the time when a corresponding bike is rented by a person from a dock at the departure station
- Arrival Time: the time when a corresponding bike is returned by a person at a dock of the arrival station

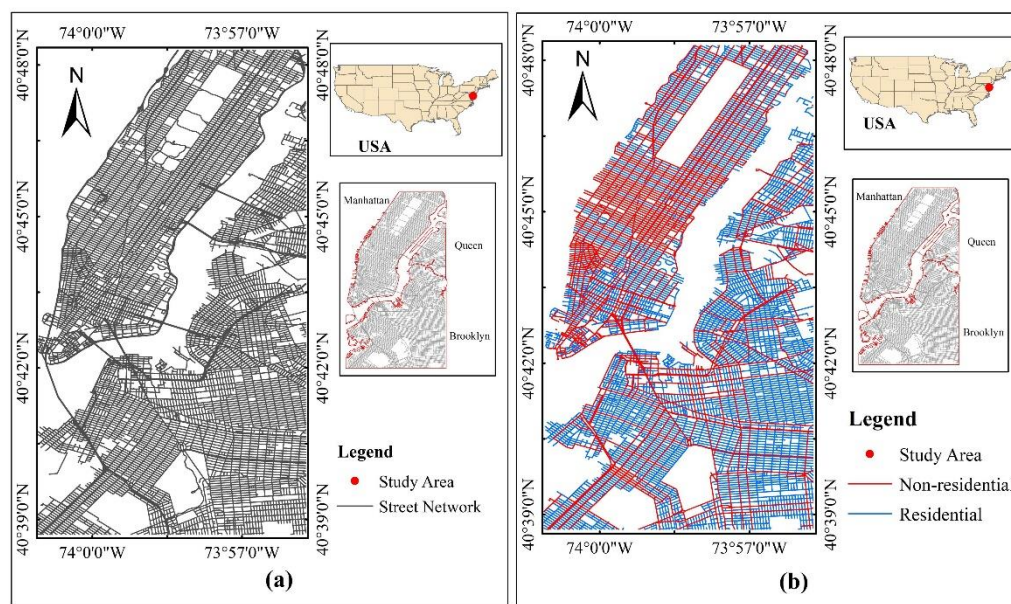
The taxi dataset for this study was retrieved from the New York City Taxi and Limousine Commission.

Each taxi trip record contains the following fields:

- Trip ID: the unique ID of the trip
- Pick-Up Coordinate: the coordinates where the taxi picks up the passenger
- Drop-Off Coordinate: the coordinates where the taxi drops off the passenger
- Pick-Up Time: the time when a corresponding taxi picks up the passenger
- Drop-Off Time: the time when a corresponding taxi drops off the passenger

### 2.2.3. OSM Map

OpenStreetMap (OSM) is an ongoing project that has been generating spatial and thematic content on a planetary scale since 2004 through millions of volunteer contributors within an open source environment [11]. In this study, street network and point of interest (POI) data obtained from OSM that are openly accessible and licensed under the Open Data Commons Open Database License are used, because OSM provides a more detailed representation of land use/land cover than remote sensing does [12]. Figure 1 shows the street network for two kinds of streets in OSM. The roads in residential areas (highway = residential) are defined as residential streets, and the primary and secondary roads (highway = primary or secondary in OSM tags) are defined as nonresidential streets in this study.



**Figure 1.** Street network of OpenStreetMap (OSM): (a) overall street network; (b) residential and nonresidential streets in study area.

#### 2.2.4. Class Definition

Land use and land cover are key factors in understanding important issues such as climate change, natural resource management, and urban and regional planning [13]. A land cover map describes the physical and biological cover of an area (e.g., grassland or water), whereas a land use map reveals human activity (e.g., infrastructure) [14–16]. Defining classes is required to generate a land cover and land use map. The detailed definitions of land use categories and building categories are provided in Table 1. The six major categories of land cover and land use are water, open space, industrial, residential, office, and entertainment.

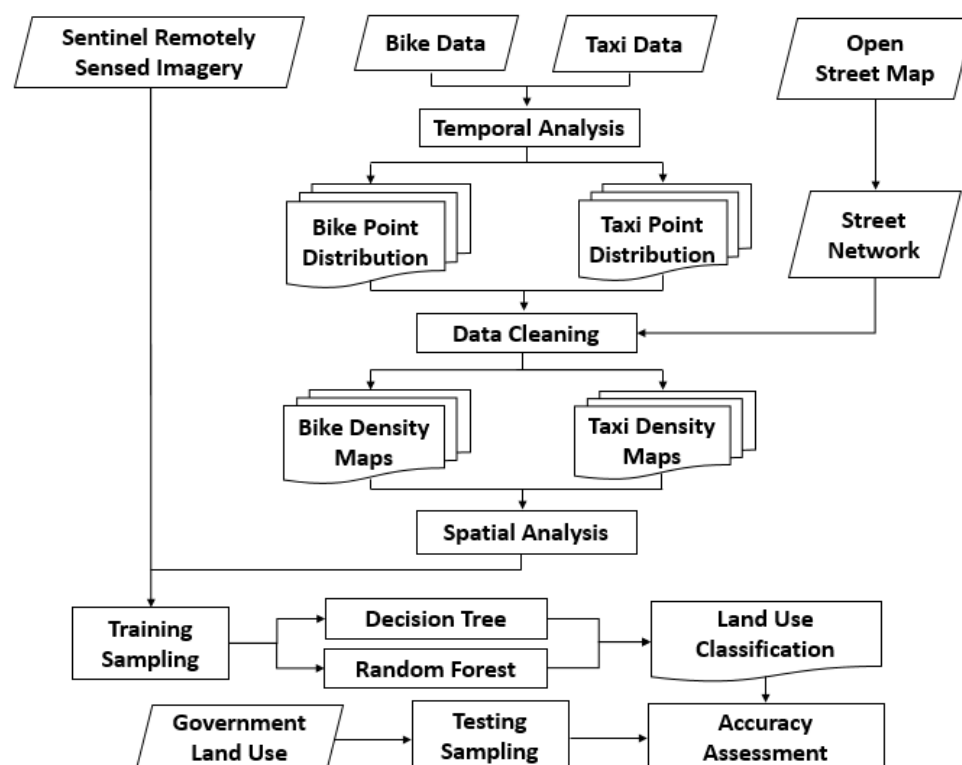
**Table 1.** Definitions of land use with land use and building categories.

Land Use Definition	Land Use Categories	Building Categories
Residential	One Family Dwelling Two Family Dwelling Walk Up Apartment	One Family Dwelling Two Family Dwelling Walk Up Apartment
	Elevator Apartment	Elevator Apartment Condominium
Office	Office	Office
Entertainment	Commercial	Hotel Theater Store Building
	Public Facilities	Churches Public Assembly and Cultural
Industrial	Industrial and Manufacturing	Warehouse Factory and industrial building
Open space	Open space and Outdoor Recreation	Outdoor Recreation Parks

### 3. Method

The data mining framework, i.e., temporal and spatial analysis, data cleaning, and machine learning, adopted in this study is presented in Figure 2. The bike and taxi datasets were processed to

understand when and where people are in the city and thereby discover human behavior patterns in the city. Temporal analysis was performed by extracting the bike and taxi datasets to reveal when people are in the city. Moreover, temporal analysis can reveal the peak times when people ride bikes and taxis in one day. Human behavior was divided into two, namely, human behavior during weekdays and human behavior during weekends. To visualize the average bike and taxi usage in one day, we divided a month's worth of bike and taxi usage into four, on the basis of the number of weeks in a month. The average bike and taxi data were split into 24 h to investigate the peak times when people ride bikes and taxis in one day during weekdays and weekends. The peak times showing when people ride bikes and taxis the most in one day were identified as Wednesday and Sunday in this study. The bike and taxi data on Wednesday, which represents a weekday; and on Sunday, which represents a weekend, were selected to generate bike and taxi usage. Meanwhile, spatial analysis was carried out by analyzing the bike and taxi density maps to reveal where people are in the city. The density maps of bikes and taxis need to be generated using kernel density estimation (pixel size = 10 m, search radius = 100 m). The coordinate location information shows the exact coordinates of the places where people take and return bikes in the bike station map and where taxi drivers pick up and drop off passengers. The bike and taxi density maps can be obtained from people's usage of bikes and taxis in the city. The origin density map indicates the location where people depart the stations on their bikes or ride taxis from pick-up locations. The destination density map indicates the location of people who return their bikes at the arrival stations or alight taxis in their drop-off locations.



**Figure 2.** Flowchart of sensing and identifying land uses, including temporal and spatial analysis, data cleaning, and machine learning (decision tree and random forest).

As the bike and taxi point distribution contains uncertain factors and outliers, data cleaning was performed to remove them. The outliers contained in the density maps were removed using the street network. The process of data cleaning involved the OSM dataset. After data cleaning, effective bike and taxi density maps for land use were generated. All of these steps were used to reveal the human behavior patterns in the city. These steps were used to classify land use through the integration of remote sensing and social sensing data. Eventually, the decision tree and random forest methods

were used in land use classification from the training samples. Training sampling was carried out by overlaying the data from the remote sensed imagery and the bike and taxi density maps. The remotely sensed imagery and bike and taxi density maps, e.g., the origins and destinations at the first and second peaks, were selected as the features for classification. For accuracy assessment, testing sampling was performed on the basis of MapPLUTO, which the New York government uses as ground truth or reference map. The land cover and land use categories included water, open space, industrial, residential, office, and entertainment.

### 3.1. Data Cleaning

Data cleaning was performed to remove the outliers from the bike and taxi point distribution. The outliers, e.g., data points far away from the commercial and residential streets, were removed. As revealed in this study, most people in the study area tend to ride bikes and taxis during weekdays to go to their offices from their homes. By contrast, most people ride bikes and taxis during weekends to go to entertainment areas from their homes. Under this phenomenon, outliers emerge because some people ride bikes and taxis during weekdays to go to places other than their offices. Likewise, some people ride bikes and taxis during weekends to go to places other than entertainment areas. Therefore, OSM was used to overcome these kinds of outliers. Two kinds of streets obtained from OSM, namely, nonresidential street and residential street, were used to remove the outliers. The residential street is located in a residential area, whereas the nonresidential street is located close to an office area. On the basis of the observation of OSM data in the study area, entertainment points are also located along the nonresidential streets.

### 3.2. Decision Tree

The decision tree belongs to supervised classification often used in land use classification because it is an efficient algorithm for classifying large datasets. The decision tree can handle data measurement on different scales without any assumptions regarding the distribution of data frequency [17]. It is a top-down classification strategy based on automatically selected rules that partition a set of given entities into small classes [18]. Several decision tree learning algorithms have been proposed, and they include classification and regression tree (CART) [19], iterative dichotomizer version 3 (ID3) [20], and C4.5 (an industrial version of ID3). These algorithms have different ways of quantifying distinction and different criteria, of which the entities of the input dataset might be independent. CART processed using R programming environment (rpart package) was used in this study.

### 3.3. Random Forest

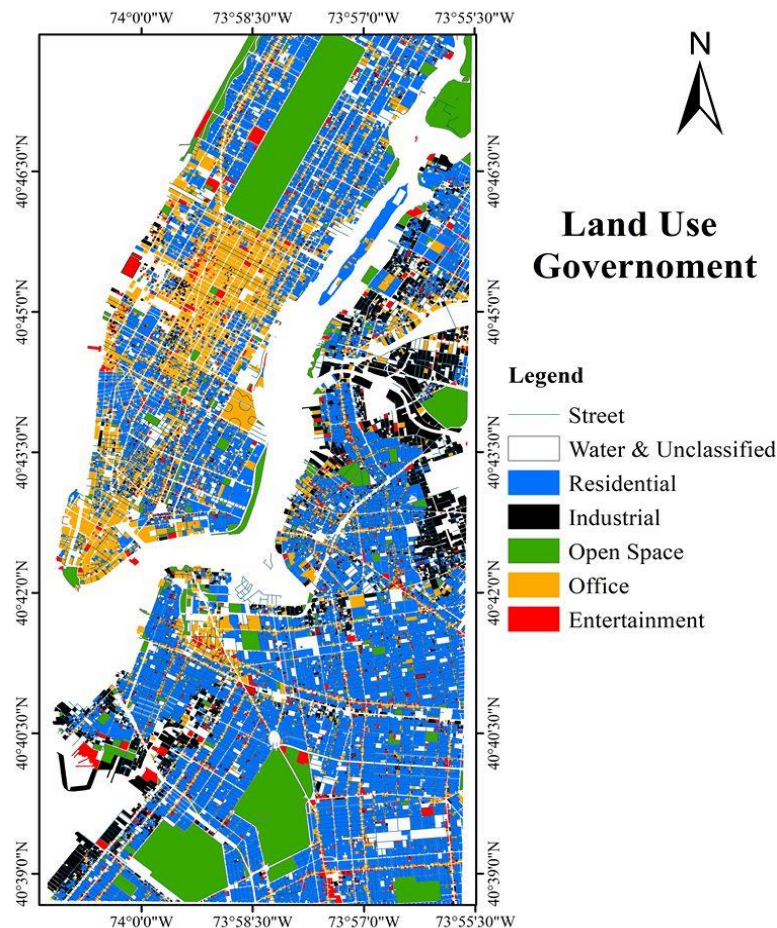
The random forest is an ensemble learning method that uses multiple independently constructed decision trees, each of which uses a unique bootstrap sample of the data [21]. The random forest algorithm has two levels of randomization in each tree [22]. The first is bootstrap aggregation, in which a random subset consisting of two-thirds of the data is sampled with the replacement of the training data and the remaining portion of the data. The second level of randomization is at each node of the individual decision tree. The randomization identifies the best split among the variables. The classification results can be achieved using a majority vote from all the decision trees, and the vote for each tree carries the same weight [23]. In this study, the random forest model is processed using R programming environment (randomForest package). The default parameter setting, i.e., 500 ntrees (number of trees), is used.

### 3.4. Training and Testing Sampling

The New York government has an openly accessible land use map called MapPLUTO obtained from the Department of City Planning of New York City (NYC Planning). MapPLUTO was created in 2016 and contains extensive land use and geographic data in shapefile and geodatabase formats. MapPLUTO also merges the data from PLUTO, which contains extensive land use and geographic data



at the tax lot level in a comma-separated value (CSV) file format and is maintained by city agencies with the Department of Finance's Digital Tax Map (NYC, 2018). Therefore, the training and testing samples in this work were performed based on the MapPLUTO, which serves as the reference map. The 700 training samples are used to construct models of decision tree and random forest. The 300 testing samples were applied to measure the accuracy of the classified maps. Figure 3 shows MapPLUTO as the reference map.



**Figure 3.** Government land use: MapPLUTO.

### 3.5. Accuracy Assessment

Accuracy assessment is the most important step in any classification method because it reveals how accurate the classification results are [24]. In this study, accuracy assessment was performed to investigate how accurate the model is. Specifically, the percentage of the testing sample from the reference map that was correctly classified by the model was estimated. The confusion matrix method was selected as the accuracy assessment tool in this study. In general, the confusion matrix identifies four types of accuracies, namely, overall accuracy, producer's accuracy, user's accuracy, and kappa coefficient. Overall accuracy denotes the total percentage of the reference or actual data correctly classified by the model. User's accuracy denotes the percentage of the model that is actually present on the reference shown by the map maker or from the user's point of view. Producer's accuracy denotes the percentage of the reference or actual map that is correctly classified by the model shown by the map maker or from the producer's point of view [25]. The confusion matrix of accuracy assessment was established to investigate how good the model was by using the testing samples from MapPLUTO.

## 4. Results and Discussion

### 4.1. Temporal Analysis of Bike and Taxi Data

Human behavior patterns can be determined through temporal analysis, which was performed in this study to reveal when people are in the city on the basis of bike and taxi usage.

#### 4.1.1. Weekday Time

Figure 4 shows the average bike and taxi usage on a weekday. The pattern shows the average number of people who rode bikes and taxis in a 24-h period during weekdays in the whole month of June 2016. Interesting insights could be gleaned from the results. For example, the average bike and taxi usage on a weekday has two peak times, namely, 08:00–09:00 and 18:00–19:00. The former refers to the period when people leave the house to go to the office. The latter refers to the period when people leave the office to go home. Another interesting insight is the similarity of the graphs of bike and taxi usage for weekdays. Both graphs show a decline in usage from 00:00–01:00 to 03:00–04:00, indicating the limited use of bikes and taxis at those times. Both graphs also show an increase in usage from 04:00–05:00 to 08:00–09:00, indicating people's regular use of bikes and taxis to travel to the workplace. Another decline is noted from 09:00–10:00 to 12:00–13:00, indicating periods when people are working. An increase in usage is noted from 13:00–14:00 to 18:00–19:00, indicating people's lunch break and departure from the office, respectively.

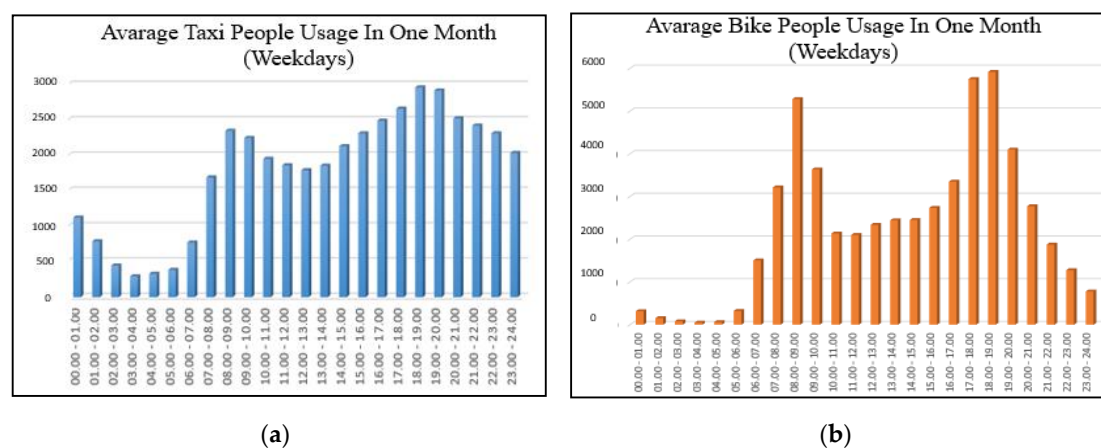


Figure 4. Average usage on weekdays for one month: (a) bike; (b) taxi.

#### 4.1.2. Weekend Time

Figure 5 shows the average bike and taxi usage during weekends. The results present the average number of people who rode bikes and taxis in a 24-h period during weekends in the whole month of June 2016. Interesting insights can also be gleaned from the results. First, the average bike and taxi usage during weekends indicate different peak times. Bike usage data show one peak time at 17:00–18:00 while taxi usage data show one peak time at 18:00–19:00. This result indicates a one-hour difference in people's use of bikes and taxis, which tends to yield a similar pattern. Weekend peak times refer to the period when most people ride bikes and taxis to go to entertainment areas from their home and vice versa.

The graphs of bike and taxi usage during weekends are similar. The weekend graphs show declines in bike and taxi usage from 00:00–01:00 to 05:00–06:00 and from 06:00–07:00, respectively. This result indicates that people seldom ride bikes and taxis during those times. In addition, both graphs show increases in bike and taxi usage until the peak times of 17:00–18:00 and 18:00–19:00, respectively. This result indicates that people ride bikes and taxis at those times to go to entertainment areas. The graphs also show a decline until midnight, indicating the period when people leave entertainment areas to go home.



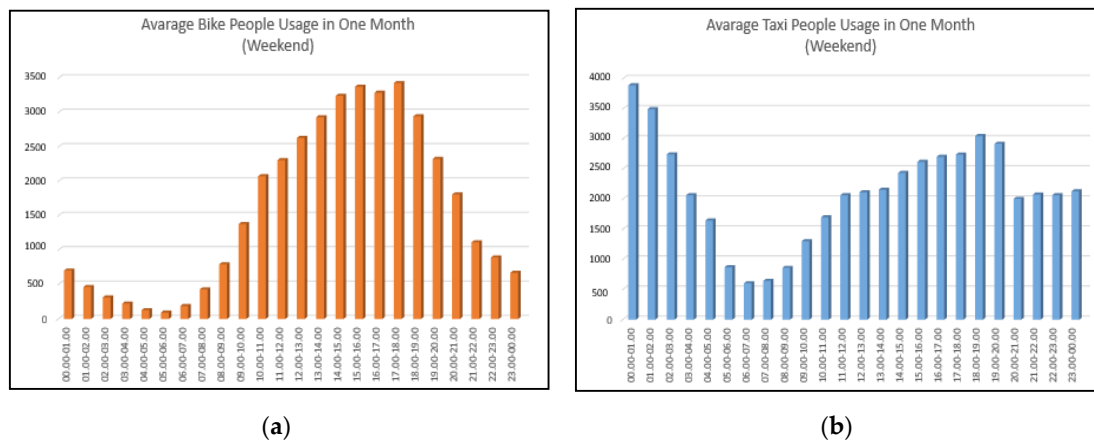


Figure 5. Average usage on weekends in one month: (a) bike; (b) taxi.

#### 4.2. Spatial Analysis of Bike and Taxi Density Maps

The bike and taxi density maps for those peak times were obtained to reveal human behavior in the city by visualizing the areas where people ride bikes and taxis. The bike and taxi density maps were divided into two, namely, origin density map and destination map.

##### 4.2.1. Weekday Time

Bike and taxi usage on weekdays showed two peak times. The first peak time is 08:00–09:00, and the second peak time is 18:00–19:00. The former indicates the period when people ride bikes and taxis the most to go to work from their homes, whereas the latter indicates the period when people ride bikes and taxis the most to go home. Therefore, the bike and taxi density maps in the origin and destination areas at the first peak time (08:00–09:00) and at the second peak time (18:00–19:00) during weekdays were generated to reveal human behavior in the city. Figures 6 and 7 show the bike and taxi density maps during the first and second weekday peak times in the origin and destination areas. The bike and taxi density hotspots at the first weekday peak time (08:00–09:00) are residential areas in the origin density map and office areas in the destination density map. The second weekday peak time (18:00–19:00) occurs in office areas in the origin density map and in residential areas in the destination density map.

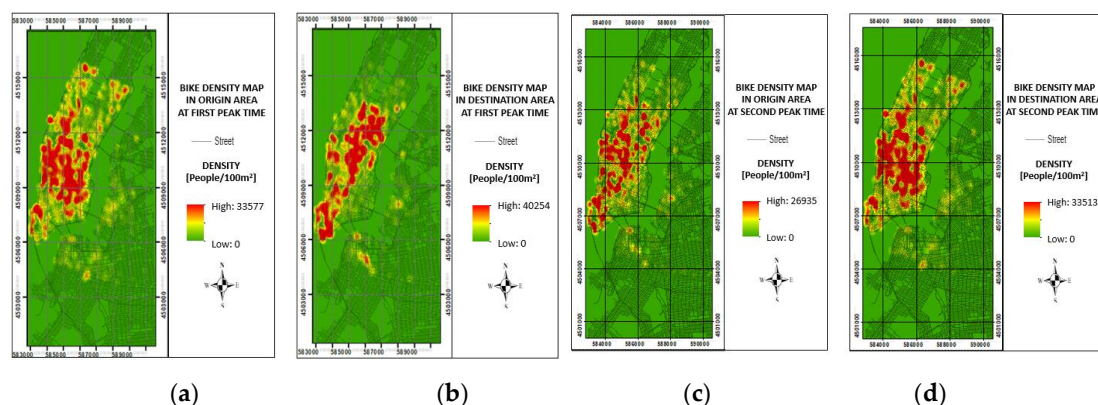
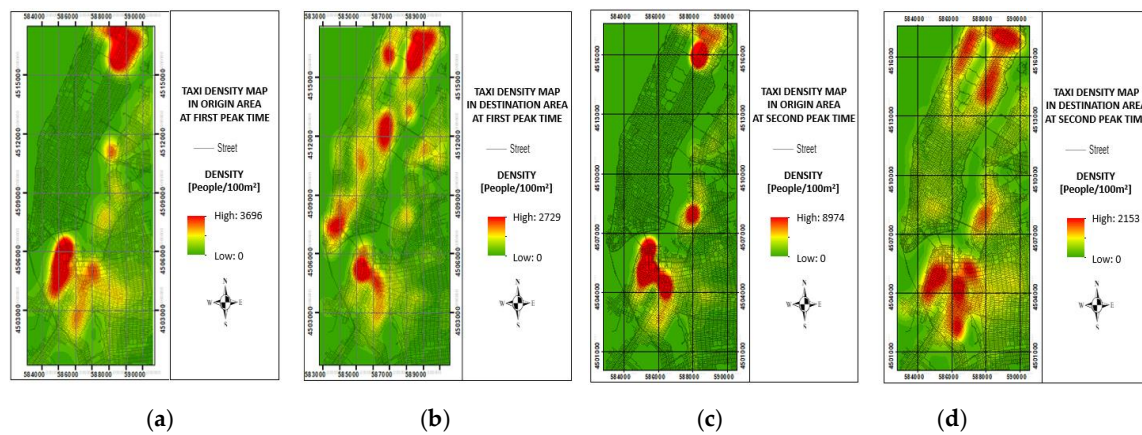


Figure 6. Bike density maps during weekday peak times: (a) origin density map at 08:00–09:00; (b) destination density map at 08:00–09:00; (c) origin density map at 18:00–19:00; (d) destination density map at 18:00–19:00.



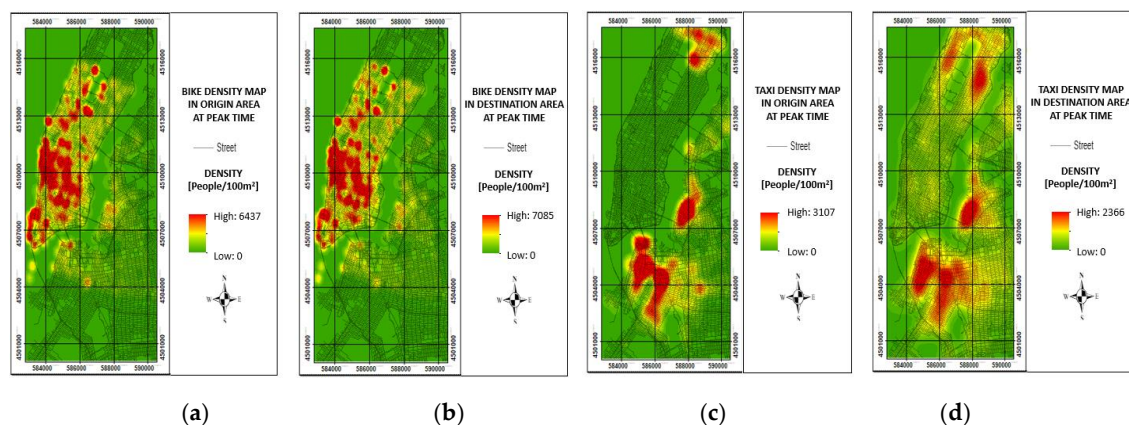
**Figure 7.** Taxi density maps during weekday peak times: (a) origin density map at 08:00–09:00; (b) destination density map at 08:00–09:00; (c) origin density map at 18:00–19:00; (d) destination density map at 18:00–19:00.

#### 4.2.2. Weekend Time

During weekends, bike usage shows one peak time at 17:00–18:00, whereas taxi usage shows one peak time at 18:00–19:00. This result indicates a one-hour difference in peak times and a similar pattern. Weekend peak times refer to the periods when people ride bikes and taxis the most to go to entertainment areas from their home and vice versa. The origin and destination density maps of bike usage during the weekend peak time of 17:00–18:00 and taxi usage during the weekend peak time of 18:00–19:00 were generated to reveal human behavior in the city.

In Figure 8, the locations of where people are during the weekend peak times of bike and taxi usage in the area of origin and destination area are residential and entertainment areas. Moreover, the locations of where people are during the second weekend peak time of bike usage in the area of origin and destination area reflect entertainment and residential areas.

Generally, people in the area of origin are located in a residential area during the first peak time on a weekday, and their destination area is an office area. During the second peak time on a weekday, people originate their journeys in an office area, and their destination area is a residential area. The weekend peak times refer to the periods when people ride bikes and taxis the most to go to entertainment areas from their homes and vice versa. On the basis of bike and taxi usage, human behavior in a city can be identified and analyzed. Bike and taxi usage hotspots in traffic source-sink areas are related to traffic intensities and land use patterns [4,10].



**Figure 8.** Bike and taxi density maps during weekend peak times: (a) bike origin density map at 17:00–18:00; (b) bike destination density map at 17:00–18:00; (c) taxi origin density map at 18:00–19:00; (d) taxi destination density map at 18:00–19:00.

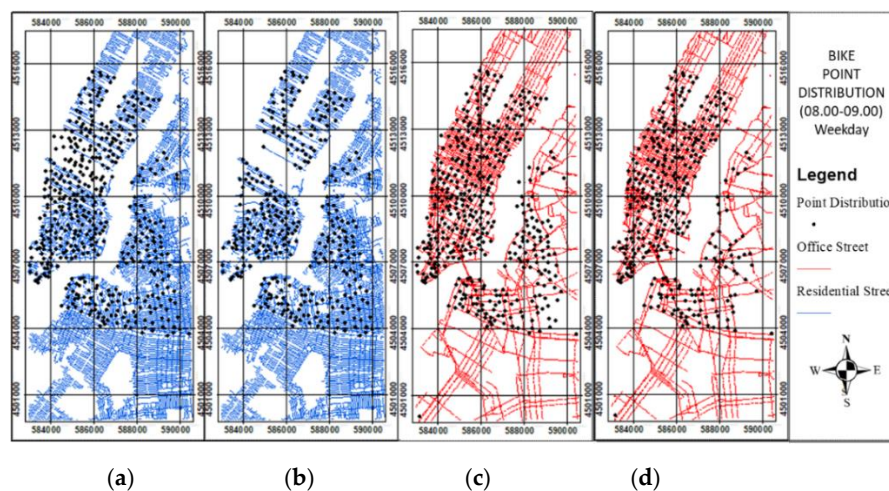
### 4.3. Effects of Data Cleaning on Point Distribution

The bike and taxi point distribution show the location where people ride bikes and taxis. Data cleaning was performed to remove the outliers from the bike and taxi point distribution by using the OSM office and residential streets.

#### 4.3.1. Weekday Time

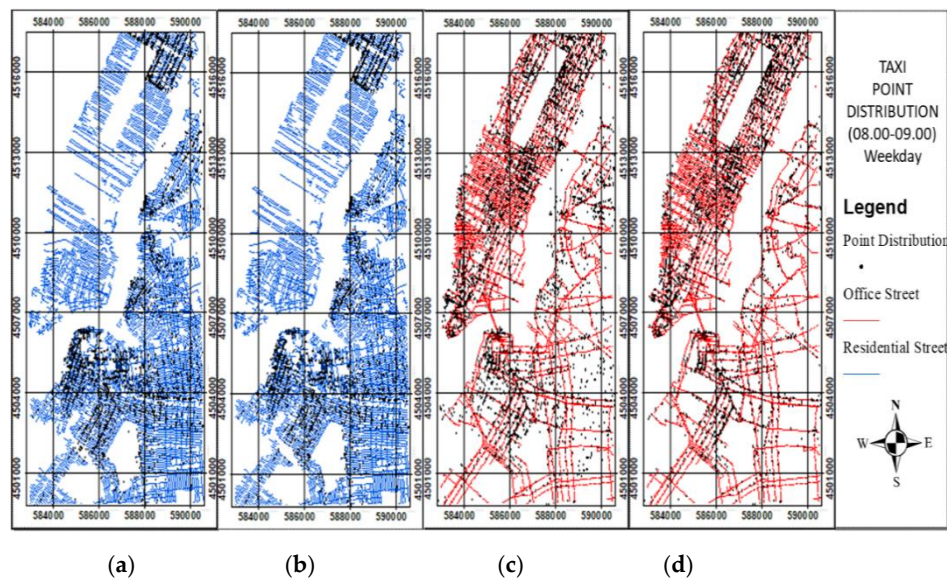
The bike and taxi usage on weekdays show the same peak times at 08:00–09:00 and 18:00–19:00. The former refers to the period when people ride bikes and taxis to leave the house and go to the office. The latter refers to the period when people ride bikes and taxis to leave the office and go home.

The data cleaning procedure was employed to remove the outliers of the point distributions in the origin and destination locations by utilizing the residential and nonresidential streets. At the peak time of 08:00–09:00, the point distributions in the origin and destination locations show a residential area located in a residential street and an office area located in a nonresidential street. Therefore, at the peak time of 08:00–09:00, the outliers of the point distribution in the origin location not located in a residential street, and in the destination location not located in a nonresidential street, should be removed (Figures 9 and 10). At the peak time of 18:00–19:00, the point distribution in the origin location shows a residential area in a nonresidential street, whereas that in the destination location shows a residential area in a residential street. Therefore, at the peak time of 18:00–19:00, the outliers of the point distribution in the origin location not located in a nonresidential street, and in the destination location not located in a residential street, should be removed. The bike and taxi point distributions at those same peak times before and after data cleaning were generated (Figures 11 and 12).

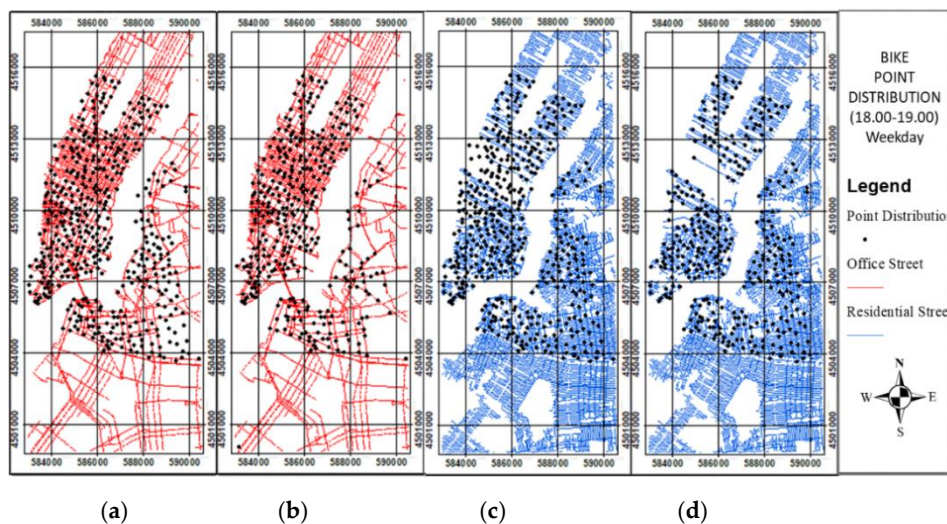


**Figure 9.** Bike point distribution during the first weekday peak time (08:00–09:00): (a) origin before data cleaning; (b) origin after data cleaning. (c) destination before data cleaning; (d) destination after data cleaning.

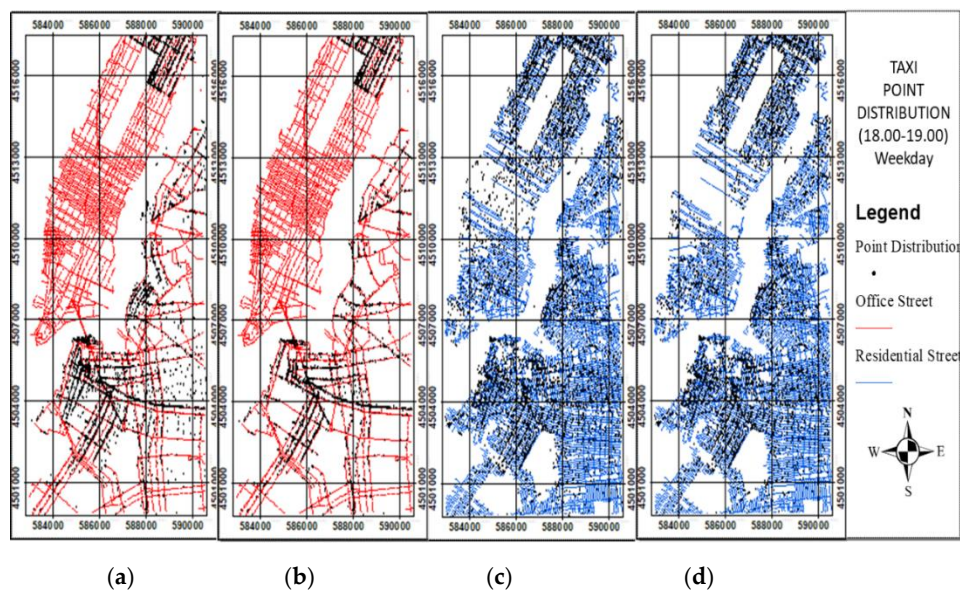




**Figure 10.** Taxi point distribution at the first weekday peak time (08:00–09:00): (a) origin before data cleaning; (b) origin after data cleaning. (c) destination before data cleaning; (d) destination after data cleaning.



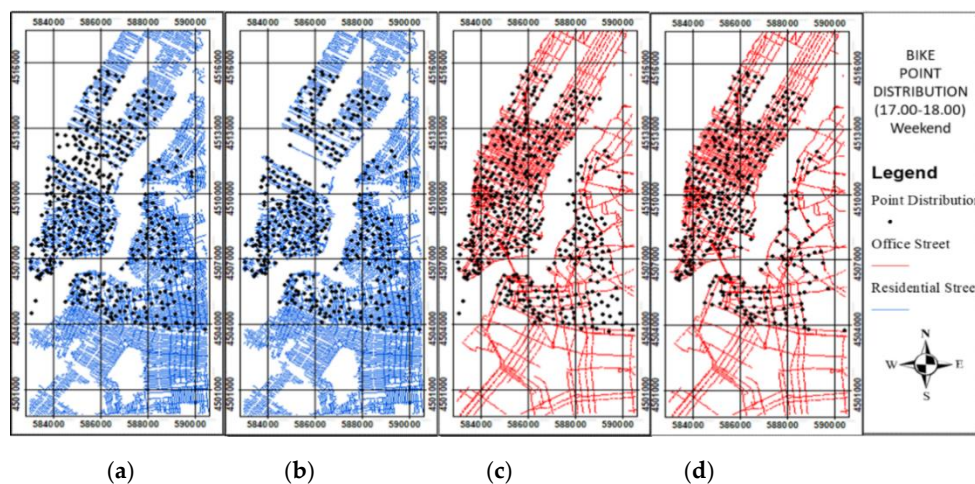
**Figure 11.** Bike point distribution at the second weekday peak time (18:00–19:00): (a) origin before data cleaning; (b) origin after data cleaning. (c) destination before data cleaning; (d) destination after data cleaning.



**Figure 12.** Taxi point distribution at the second weekday peak time (18:00–19:00): (a) origin before data cleaning; (b) origin after data cleaning. (c) destination before data cleaning; (d) destination after data cleaning.

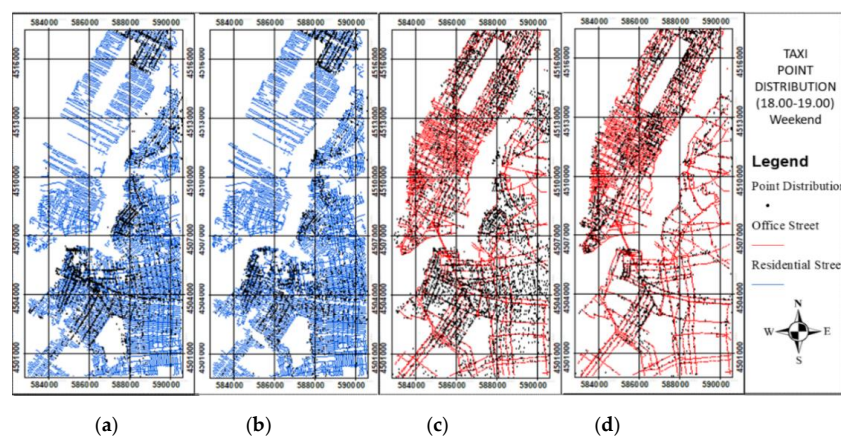
#### 4.3.2. Weekend Time

The weekend peak times of bike and taxi usage are 17:00–18:00 and 18:00–19:00, respectively. During these weekend peak times, most people ride bikes and taxis to leave the house and go to entertainment areas. The data cleaning procedure was implemented to remove the outliers of point distribution in the origin and destination locations by utilizing the residential and nonresidential streets. During the peak time of bike and taxi usage, the point distribution in the origin location shows a residential area located in a residential street, whereas that in the destination location shows an entertainment area located in a nonresidential street. Therefore, at those peak times of bike and taxi usage, the outliers of point distributions in the origin location not in a residential street and those in the destination location not in an entertainment area in a nonresidential street should be removed. The bike and taxi point distributions at the different peak times before and after data cleaning are shown in Figures 13 and 14, respectively.



**Figure 13.** Bike point distribution at the weekend peak time (17:00–18:00): (a) origin before data cleaning; (b) origin after data cleaning (c) destination before data cleaning; (d) destination after data cleaning.





**Figure 14.** Taxi point distribution at the weekend peak time (18:00–19:00): (a) origin before data cleaning; (b) origin after data cleaning (c) destination before data cleaning; (d) destination after data cleaning.

#### 4.4. Accuracy Assessment of Land Use Model

Table 2 shows the different overall accuracy and kappa coefficient values of the urban land use models for the six classes. Considering only remote sensing information, the overall accuracy of the decision tree was 69%, whereas the overall accuracy of the random forest was 63%, respectively. Using remote and social data, the overall accuracy of the decision tree increases to 78%, whereas the overall accuracy of the random forest increases to 81%. After data cleaning, the overall accuracy of the decision tree reaches 83%, whereas the overall accuracy of the random forest reaches 86%. The best kappa coefficients of the decision tree classified map and random forest classified maps were 0.80 and 0.82, respectively. The results show that the effective data cleaning of remote and social sensing data ensures an accurate urban land use classification. The results also concluded that for the six-class case, the random forest classified map had better accuracy than the decision tree classified map.

**Table 2.** Comparison of accuracy assessment.

Classification Method	Different Ways	Accuracy Assessment	
		Overall Accuracy	Kappa Coefficient
Decision Tree	Use RS Only	69%	0.63
	Integration of RS and SS	78%	0.73
	Without Data Cleaning		
	Integration of RS and SS	83%	0.80
	With Data Cleaning		
Random Forest	Use RS Only	63%	0.55
	Integration of RS and SS	81%	0.76
	Without Data Cleaning		
	Integration of RS and SS	86%	0.82
	With Data Cleaning		

(RS: remote sensing; SS: social sensing).

Tables A1–A6 respectively show the confusion matrix of the decision tree and random forest classified maps resulting from the integration of remote and social sensing with the use of data cleaning. The result implies the dramatic improvement with the integration of social and remote sensing data with data cleaning relative to the sole use of remote sensing data. With the sole application of remote sensing, the user and producer accuracies are low in entertainment, industrial, commercial, and residential areas. In urban land use classification, using a combination of remote and social sensing data with data cleaning is superior to using remote sensing only. Take the decision tree as

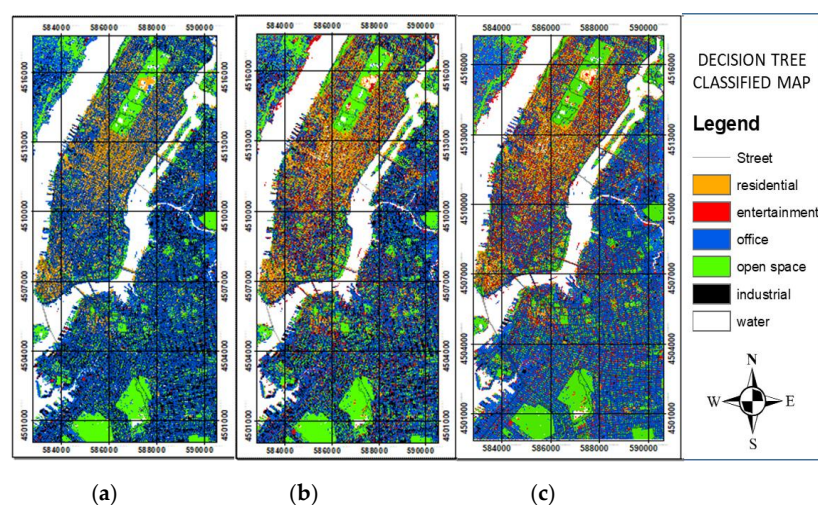
an example; user accuracy (commission error) improves in office, residential, and industrial areas (18%, 25%, and 21% respectively), and producer accuracy (omission error) improves in residential and entertainment areas (34% and 40%, respectively).

#### 4.5. Urban Land Use Map

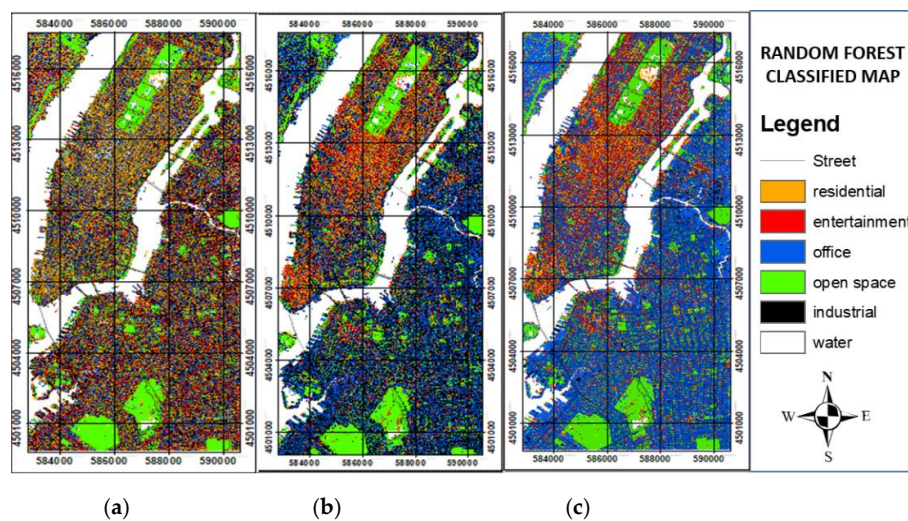
The land use classified map is generated in three ways, namely, using remote sensing data only, using integration of remote sensing and social sensing data without data cleaning, and using integration of both data with data cleaning. Figures 15 and 16 show the comparison results for the six categories of the land use classified map from the three aforementioned methods based on decision tree and random forest classification. All the land use classified map results reveal that the distribution of industrial areas in Brooklyn is random for the six classes of decision tree and random forest land use classified maps that used the integration of remote and social sensing data without data cleaning. The distribution of offices in the random forest classified map is more random than that in the decision tree classified map when only remote sensing is used. Random forest obviously requires social sensing data to generate office areas. The random forest classified map based on the integration of remote sensing and social sensing with data cleaning has higher density of office and residential areas in Central Manhattan than that of the decision tree classified map.

Through additional social sensing data, model performance can be improved. The combination of remote and social sensing data is a promising way to map detailed urban land use in a large metropolitan area [9]. Bike and taxi density maps are applied to reliably estimate the intensity, duration, and frequency of human activity patterns [26]. However, a certain bias or uncertainty may exist toward the use of social data to detect the urban land use types of a city [26]. For instance, people who use bikes or taxis on weekdays do not always travel to office areas from their homes, e.g., going to educational places from home. Nevertheless, most people use bikes and taxis to leave their homes and go to office areas on weekdays. On the basis of the given rule, outliers that do not belong to residential and office areas are removed in the density maps. After data cleaning, the urban land use patterns using both models show a 5% improvement in overall accuracy in comparison with those obtained without data cleaning.

Sensing-based land use identification can help policy makers link land use intensity and human activity to the socioeconomic consequences of different urban land uses within a landscape. On the basis of the proposed approach, we will conduct further research to identify the patterns of urban land use classes, such as compactness of a specified land use type, degree of urban change, and expansion rate [27].



**Figure 15.** Six-class decision tree classified map using (a) remote sensing only; (b) integration of remote and social sensing (without data cleaning); (c) integration of remote and social sensing (with data cleaning).



**Figure 16.** Six-class random forest classified map using (a) remote sensing only; (b) integration of remote and social sensing (without data cleaning); (c) integration of remote and social sensing (with data cleaning).

## 5. Conclusions

This study focused on considering dynamic patterns associated with various human activities and behaviors for urban land use classification. This study successfully generated the density maps to show where most people are located in the city at peak times. The bike and taxi density maps of the origin and destination areas on weekdays reflect the location of residential and office areas, whereas on weekends, they reflect the location of residential and entertainment areas. The density map patterns are highlighted for understanding human behavior in a city. The daytime destination and nighttime origin density reflect work-related land uses, including commercial and industrial areas. By contrast, the nighttime destination and daytime origin density pattern capture the pattern of residential areas. Human behavior toward taxi and bike usage during weekdays and weekends offers interesting insights. During weekdays, bike and taxi usage show the same two peak times, and the patterns are identical due to the similarity of the graphs. These peak times occur at 08:00–09:00 and 18:00–19:00. The same peak times indicate that most people who ride bikes or taxis go to work and return home. The bike and taxi usage on weekends reflects a different pattern. During weekends, bike and taxi usage shows different peak times, and the graph patterns are similar, although the similarity is not as close as that for the graph of bike and taxi usage on weekdays. These weekend peak times refer to the time when most people ride bikes and taxis to go to entertainment areas from their home and vice versa.

On the basis of the remote sensing and social sensing data, the machine learning methods successfully generate a land use classified map of an urban area in New York, USA, with six categories, namely, water, open space, industrial, residential, office, and entertainment classes. The accuracy assessment of the confusion matrix is used to investigate the accuracy of the decision tree and random forest classified maps. The overall accuracies of the decision tree and random forest classified maps for the six classes reach 83% and 86%, respectively. The land use classified maps show high accuracy for the integration of remote sensing and social sensing data with data cleaning. The accuracy assessment indicates the effective mining of remote and social sensing data and ensures an accurate urban land use classification. In future work, area-based accuracy assessment and other information for data cleaning can be considered.

**Author Contributions:** Conceptualization, Hone-Jay Chu and Adindha Surya Anugraha; Methodology, Adindha Surya Anugraha; Formal Analysis, Adindha Surya Anugraha and Muhammad Zeeshan Ali; Data Curation, Adindha Surya Anugraha; Writing-Original Draft Preparation, Adindha Surya Anugraha; Writing-Review & Editing, Hone-Jay Chu; Visualization, Muhammad Zeeshan Ali and Adindha Surya Anugraha; Supervision,

Hone-Jay Chu; Project Administration, Hone-Jay Chu; Funding Acquisition, Hone-Jay Chu. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by MOST, Taiwan, grant number 107-2119-M-006-024- and The APC was funded by MOST, Taiwan.

**Acknowledgments:** The authors would like to thank the editors and anonymous reviewers for providing suggestions of paper improvement. Moreover, the data that support the findings of this study are openly available in Citi Bike NYC and TLC Trip Record Data. The authors also thank these data providers.

**Conflicts of Interest:** We confirm that the authors face no conflict of interest.

## Appendix A

**Table A1.** Confusion matrix of decision tree classified map using remote sensing only.

Actual Model	OpenSpace	Water	Industrial	Office	Entertainment	Residential	User's Accuracy
OpenSpace	44	3	3	0	0	4	81
Water	4	45	1	1	0	0	88
Industrial	0	0	34	0	7	8	69
Office	1	0	4	41	16	9	58
Entertainment	0	0	2	0	18	4	75
Residential	1	2	6	8	9	25	49
Producer's accuracy	88	90	68	82	36	50	

**Table A2.** Confusion matrix of decision tree classified map using integration of remote and social sensing (without data cleaning).

Actual Model	OpenSpace	Water	Industrial	Office	Entertainment	Residential	User's Accuracy
OpenSpace	46	3	0	0	1	7	81
Water	4	45	0	1	0	0	90
Industrial	0	0	34	0	2	9	76
Office	0	0	4	40	6	2	77
Entertainment	0	0	5	4	36	0	80
Residential	0	2	7	5	5	32	63
Producer's accuracy	92	90	68	80	72	64	

**Table A3.** Confusion matrix of decision tree classified map using integration of remote and social sensing (with data cleaning).

Actual Model	OpenSpace	Water	Industrial	Office	Entertainment	Residential	User's Accuracy
OpenSpace	46	5	0	0	0	0	90
Water	4	45	0	0	0	0	92
Industrial	0	0	39	1	2	1	91
Office	0	0	4	40	5	4	75
Entertainment	0	0	2	4	38	3	81
Residential	0	0	5	5	5	42	74
Producer's accuracy	92	90	78	80	76	84	

## Appendix B

**Table A4.** Confusion matrix of random forest classified map using remote sensing only.

Actual Model	OpenSpace	Water	Industrial	Office	Entertainment	Residential	User's Accuracy
OpenSpace	46	2	0	0	0	0	96
Water	4	45	0	1	0	0	90
Industrial	0	1	24	1	2	5	73
Office	0	2	12	40	32	12	41
Entertainment	0	0	7	8	14	14	33
Residential	0	0	7	0	2	19	68
Producer's accuracy	92	90	48	80	28	38	



**Table A5.** Confusion matrix of random forest classified map using integration of remote and social sensing (without data cleaning).

Actual Model	OpenSpace	Water	Industrial	Office	Entertainment	Residential	User's Accuracy
OpenSpace	46	2	0	0	0	0	96
Water	4	45	0	0	0	0	92
Industrial	0	2	36	4	2	8	69
Office	0	1	4	43	8	5	70
Entertainment	0	0	2	3	38	3	83
Residential	0	0	8	0	2	34	77
Producer's accuracy	92	90	72	86	76	68	

**Table A6.** Confusion matrix of decision tree classified map using integration of remote and social sensing (with data cleaning).

Actual Model	OpenSpace	Water	Industrial	Office	Entertainment	Residential	User's Accuracy
OpenSpace	46	2	0	0	0	0	96
Water	4	45	0	0	0	0	92
Industrial	0	0	37	1	1	3	88
Office	0	0	6	43	6	1	77
Entertainment	0	0	2	4	41	1	85
Residential	0	3	5	2	2	45	79
Producer's accuracy	92	90	74	86	82	90	

## References

- Jensen, J.R.; Cowen, D.C. Remote sensing of urban/suburban infrastructure and socio-economic attributes. In *The Map Reader*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2011; pp. 153–163.
- Hu, S.; Wang, L. Automated urban land-use classification with remote sensing. *Int. J. Remote Sens.* **2012**, *34*, 790–803. [\[CrossRef\]](#)
- Herold, M.; Liu, X.; Clarke, K.C. Spatial Metrics and Image Texture for Mapping Urban Land Use. *Photogramm. Eng. Remote Sens.* **2003**, *69*, 991–1001. [\[CrossRef\]](#)
- Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L. Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Ann. Assoc. Am. Geogr.* **2015**, *105*, 512–530. [\[CrossRef\]](#)
- D'Andrea, E.; Ducange, P.; Lazzerini, B.; Marcelloni, F. Real-Time Detection of Traffic From Twitter Stream Analysis. *IEEE Trans. Intell. Transp. Syst.* **2015**, *16*, 1–15. [\[CrossRef\]](#)
- Shin, D.; Aliaga, D.; Tuncer, B.; Arisona, S.M.; Kim, S.; Zünd, D.; Schmitt, G. Urban sensing: Using smartphones for transportation mode classification. *Comput. Environ. Urban Syst.* **2015**, *53*, 76–86. [\[CrossRef\]](#)
- Pan, G.; Qi, G.; Wu, Z.; Zhang, D.; Li, S. Land-Use Classification Using Taxi GPS Traces. *IEEE Trans. Intell. Transp. Syst.* **2013**, *14*, 113–123. [\[CrossRef\]](#)
- Jiang, S.; Alves, A.; Rodrigues, F.; Ferreira, J.; Pereira, F.C. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Comput. Environ. Urban Syst.* **2015**, *53*, 36–46. [\[CrossRef\]](#)
- Hu, T.; Yang, J.; Li, X.; Gong, P. Mapping Urban Land Use by Using Landsat Images and Open Social Data. *Remote Sens.* **2016**, *8*, 151. [\[CrossRef\]](#)
- Anugraha, A.S.; Chu, H.-J. Land Use Classification from Combined Use of Remote Sensing and Social Sensing Data. *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.* **2018**, *XLII-4*, 33–39. [\[CrossRef\]](#)
- Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environ. Plan. B Urban Anal. City Sci.* **2010**, *37*, 682–703. [\[CrossRef\]](#)
- Schultz, M.; Voss, J.; Auer, M.; Carter, S.; Zipf, A. Open land cover from OpenStreetMap and remote sensing. *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *63*, 206–213. [\[CrossRef\]](#)
- Jones, K. *Importance of Land Cover and Biophysical Data in Landscape-Based Environmental Assessment*; North America Land Cover Summit; Association of American Geographers: Washington, DC, USA, 2008; pp. 215–249.
- Thenkabail, P.S.; Schull, M.; Turrall, H. Ganges and Indus river basin land use/land cover (LULC) and irrigated area mapping using continuous streams of MODIS data. *Remote Sens. Environ.* **2005**, *95*, 317–341. [\[CrossRef\]](#)
- Kasetkasem, T.; Arora, M.K.; Varshney, P.K. Super-resolution land cover mapping using a Markov random field based approach. *Remote Sens. Environ.* **2005**, *96*, 302–314. [\[CrossRef\]](#)



16. Lambin, E.F.; Turner, B.L.; Geist, H.J.; Agbola, S.B.; Angelsen, A.; Bruce, J.W.; George, P. The causes of land-use and land-cover change: Moving beyond the myths. *Glob. Environ. Chang.* **2001**, *11*, 261–269. [\[CrossRef\]](#)
17. Pal, M.; Mather, P. An assessment of the effectiveness of decision tree methods for land cover classification. *Remote Sens. Environ.* **2003**, *86*, 554–565. [\[CrossRef\]](#)
18. Quinlan, J. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81–106. [\[CrossRef\]](#)
19. Breiman, L.; Friedman, J.H.; Olshen, R.A.; Stone, C.J. *Classification and Regression Trees*; Wadsworth & Brooks/Cole Advanced Books & Software: Monterey, CA, USA, 1984.
20. Mitchell, D.; Tom, M. *Machine Learning*; McGraw-Hill: New York, NY, USA, 1997.
21. Liaw, A.; Wiener, M. Classification and regression by randomforest. *R News* **2002**, *2*, 18–22.
22. Woznicki, S.A.; Baynes, J.; Panlasigui, S.; Mehaffey, M.; Neale, A. Development of a spatially complete floodplain map of the conterminous United States using random forest. *Sci. Total Environ.* **2019**, *647*, 942–953. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Zhang, M.; Chen, F.; Tian, B.; Liang, D. Multi-temporal SAR image classification of coastal plain wetlands using a new feature selection method and random forests. *Remote Sens. Lett.* **2018**, *10*, 312–321. [\[CrossRef\]](#)
24. Chu, H.-J.; Wang, C.-K.; Kong, S.-J.; Chen, K.-C. Integration of full-waveform LiDAR and hyperspectral data to enhance tea and areca classification. *GIScience Remote Sens.* **2016**, *53*, 542–559. [\[CrossRef\]](#)
25. Olofsson, P.; Foody, G.M.; Stehman, S.V.; Woodcock, C.E. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sens. Environ.* **2013**, *129*, 122–131. [\[CrossRef\]](#)
26. Wang, Y.-D.; Wang, T.; Tsou, M.-H.; Li, H.; Jiang, W.; Guo, F. Mapping Dynamic Urban Land Use Patterns with Crowdsourced Geo-Tagged Social Media (Sina-Weibo) and Commercial Points of Interest Collections in Beijing, China. *Sustainability* **2016**, *8*, 1202. [\[CrossRef\]](#)
27. Jiao, L. Urban land density function: A new method to characterize urban expansion. *Landsc. Urban Plan.* **2015**, *139*, 26–39. [\[CrossRef\]](#)



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).