LAND USE CLASSIFICATION FROM COMBINED USE OF REMOTE SENSING AND SOCIAL SENSING DATA

Adindha Surya Anugraha[1], Hone-Jay Chu[1] *

[1]Dept. of Geomatics, National Cheng Kung University, Taiwan –
adindha.surya@gmail.com, honejaychu@gmail.com

Commission IV, ICWG III/IVb and WG IV/4

KEY WORDS: Human Behavior**,** Social Sensing, Remote Sensing, Decision Tree, Geographic Information System, Accuracy Assessment

ABSTRACT:

Large amounts of data can be sensed and analyzed to discover patterns of human behavior in cities for the benefit of urban authorities and citizens, especially in the areas of traffic forecasting, urban planning, and social science. In New York, USA, social sensing, remote sensing, and urban land use information support the discovery of patterns of human behavior. This research uses two types of openly accessible data, namely, social sensing data and remote sensing data. Bike and taxi data are examples of social sensing data, whereas sentinel remote sensed imagery is an example of remote sensing data. This research aims to sense and analyze the patterns of human behavior and to classify land use from the combination of remote sensing data and social sensing data. A decision tree is used for land use classification. Bike and taxi density maps are generated to show the locations of people around the city during the two peak times. On the basis of a geographic information system, the maps also reflect the residential and office areas in the city. The overall accuracy of land use classification after the consideration of social sensing data is 85.3%. The accuracy assessment shows that the combination of remote sensing data and social sensing data facilitates accurate urban land use classification.

1. INTRODUCTION

Urban land use information is crucial to urban planning, economic analysis, hazard and pollution analysis, and environmental conservation (Jensen et al., 2011). In the past decades, the demand for urban land use maps for utilization by urban authorities, researchers, and citizens, gas steadily increased. The timely acquisition of up-to-date land use information is equally important because the urban environment has been changing at a great pace, especially in rapidly developing regions (Hu et al., 2013). Therefore, the product of remote sensing, namely, remote sensed imagery, has become a major data source for mapping land use to obtain land use information.

Remote sensed imagery provides abundant and detailed information on the spectral, textural, contextual, and spatial configuration of urban land cover (Herold et al., 2003). However, remote sensed imagery is unable to examine the socioeconomic and demographic characteristics of urban land. Liu et al. (2015) showed that social sensing data (e.g., social media, mobile phones, digital maps, and GPS trajectories) can reveal the socioeconomic and demographic characteristics of urban land. Social sensing supports the discovery of patterns of human behavior in cities. Therefore, the combination of remote sensing data and social sensing data is expected to provide insights into urban landscape patterns and thereby facilitate accurate urban land use classification.

Nowadays, large amounts of data can be sensed and analyzed to discover patterns of human behavior in cities. Understanding the patterns of human behavior can benefit urban authorities and citizens. Understanding human behavior is important in traffic forecasting, urban planning, and social science. Traffic forecasting, in particular, requires an understanding of where and when people are in a city; urban planning is for blueprinting new bicycle paths and roads; and social science involves the study of how people move out around a city (Latour, 2007).

In the current work, the patterns of human behavior in a city are identified to understand where and when people are in the city. Many people today tend to choose public transportation over private vehicles to avoid traffic congestion. By using public transportation, human behavior in a city can be easily sensed and analyzed due to the availability of public transportation data provided by urban authorities. Therefore, public transportation data that belong to social sensing data are needed to fulfill the traffic forecasting purpose.

The emergence of bike sharing systems provides a way for people to promote green public transportation and healthy lifestyles (DeMaio, 2009). Bike system data are often made openly accessible for the benefit of citizens and researchers (Chen et al., 2015). With such data, people can easily find the nearest bike stations and check the availability of docks and bikes. Researchers can also extract huge information from bike system data to understand human behavior in the city.

Taxis are a widely used mode of public transportation in many cities. In metropolitan areas, such as New York, London, and Beijing, a large number of taxis move around the streets to

* Corresponding author

transport people from and to urban cores, business areas, tourist attractions, transportation hubs, and residential areas (Chu et al., 2014). Nowadays, GPS devices installed on taxis can record the moving path of taxis in real time as a series of positions with periodic intervals. At each position, the GPS provides information such as time, speed, geographic coordinates of latitude and longitude, speed, and occupancy status of taxis. Massive taxi data contain abundant information about a city and its citizens, and thus, they have been widely used in urban computing (Dohuki et al., 2017), especially in the discovery of patterns of human behavior in a city.

In urban areas, office and residential areas are the two most frequented places by people who use public transportation on weekdays. That is, people generally leave their homes (residential areas) and go to work (office areas) in the morning and return to their homes from their offices at night. Therefore, with the use of bike and taxi data, human behavior in a city can be sensed and analyzed to determine the periods in which people go to work and return home. Moreover, using bike and taxi data can reflect the residential and office areas in a city by generating bike and taxi density maps.

This research has two aims. The first aim is to sense and analyze the patterns of human behavior in a city. The second aim is to classify land use using the combination of remote sensing data and social sensing data.

2. METHODS

This section describes the study area, the source of remote sensing and social sensing data, and the methodological framework to map urban land use (Figure 1). This study integrates remote sensing and social sensing information to classify land use on the basis of a decision tree.

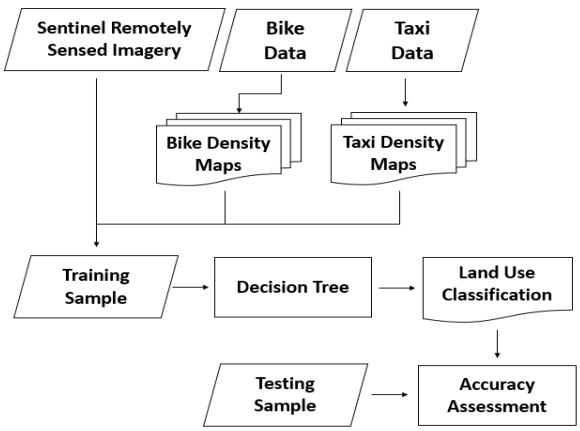


Figure 1. Methodological framework of urban land use classification

2.1 Study Area

New York City in the United States of America (USA), which comprises the districts of Manhattan, Brooklyn, and Queens, is selected as the study area (Figure 2). The study area covers a square area of approximately 139 km^2. It comprises rivers, industrial areas, parks, residential areas, and office areas. Brooklyn borders the borough of Queens, and both districts are separated from Manhattan by the East River. Manhattan is the most densely populated district of New York City, and it has become the economic and administrative center. Brooklyn is characterized by a spike in real estate development and an ever-changing landscape. Queens is an ethnically diverse urban area. Therefore, the districts of Manhattan, Brooklyn, and Queens are suitable for urban land use classification.

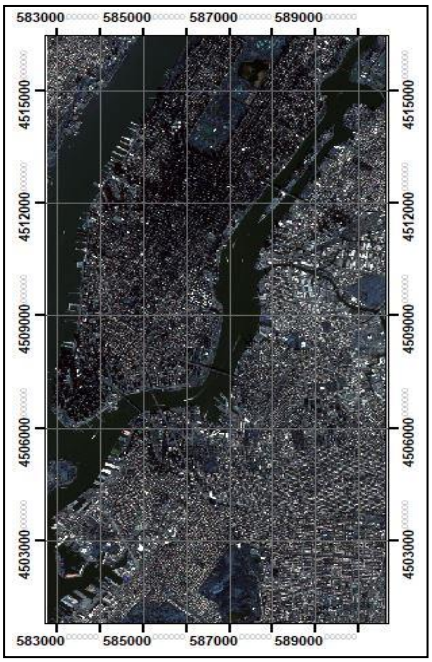


Figure 2. Study area of New York City overlaid on a Sentinel true-color image

2.2 Remote Sensing Data

In this research, sentinel-2A remote sensed imagery of New York in 2016 was used as the remote sensing data source. Sentinel-2A imagery has the span of 13 spectral bands ranging from the visible and the near-infrared to the shortwave infrared at different spatial resolutions in the range of 10–60 m on the ground; thus, such imagery takes global land monitoring to an unprecedented level (Satellite Imaging Corporation, 2017). Sentinel-2A imagery has three types of spatial resolution, namely, 10, 20, and 60 m. The bands for a spatial resolution of 10 m are used for basic land cover classification. The bands for a 20 m spatial resolution are used to enhance the classification of land cover and the retrieval of geophysical parameters. The bands for a spatial resolution of 60 m are used for atmospheric correction and cirrus-cloud screening. Among these bands, the bands for the 10 m resolution are mainly used for classification, with the labels being bands red, green, blue, and NIR. Table 1 shows the specifications of the Sentinel-2A satellite sensor.

Band	Band	Sentinel-2	
		Central Wavelength (μm)	Spatial Resolution (m)
Band 1	Coastal Aerosol	0.443	60
Band 2	Blue	0.490	10
Band 3	Green	0.560	10
Band 4	Red	0.665	10
Band 5	Vegetation Red Edge	0.705	20
Band 6	Vegetation Red Edge	0.740	20
Band 7	Vegetation Red Edge	0.783	20
Band 8a	NIR	0.842	10
Band 8	Vegetation Red Edge	0.865	20
Band 9	Water Vapour	0.945	60
Band 10	SWR - Cirrus	1.380	60
Band 11	SWR	1.610	20
Band 12	SWR	2.190	20

Table 1 Specifications of Sentinel-2A satellite sensor

2.3 Social Sensing Data

This research presents two types of social sensing data, namely, bike data and taxi data. Social sensing data are used to sense and analyze human behavior in New York, USA. Understanding human behavior can benefit urban authorities and citizens.

2.3.1 Bike Data: The current data of many bike sharing systems are openly accessible. The bike dataset used in this research originates from the open accessible CitiBike Bikeshare System of New York. CitiBike consists of more than 750 bike stations and 12,000 bikes. Each station is equipped with a specific number of docks and bikes. CitiBike has become the nation’s largest bike sharing program. Figure 3 shows the coverage of the CitiBike Bikeshare stations in New York urban areas.

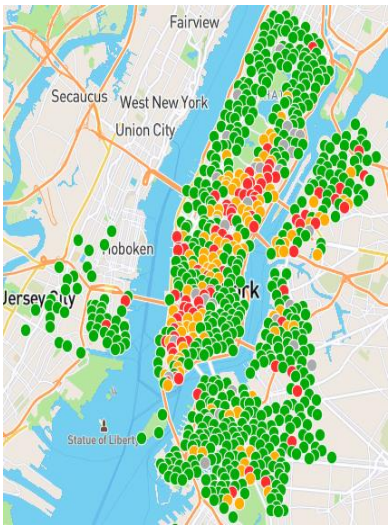


Figure 3. Station map of the CitiBike Bikeshare system in New York urban areas ©CitiBike Image Copyright 2018.

In this work, we retrieve the dataset from the CitiBike Bikeshare system, including a month’s worth of bike trip records obtained every Wednesday in June 2016. In summary, the datasets contain 45,431 recorded trips. Each bike trip record contains the following fields:

- **Trip ID:** the unique trip ID
- **Departure Station ID:** the unique ID of the station where people rent bikes for travel.
- **Arrival Station ID:** the unique ID of the station where people return the bikes after use.
- **Departure Time:** the time when a corresponding bike is rented by people from a dock of the departure station.
- **Arrival Time:** the time when a corresponding bike is returned by people to a dock of the departure station.
- **User Type:** the type of people who rent bikes, including customers and subscribers. Customers refer to riders who rent bikes with temporal keys. Subscribers are those who are registered in the system and probably use bikes regularly.

2.3.2 Taxi Data: Taxis are the mostly widely used mode of public transportation in many cities. Massive taxi datasets are available because of the large number of taxis that move along the streets to transport people around the city. Taxi datasets record consecutive samples at an interval of a few seconds in a given time period. Data are recorded by GPS devices installed on taxis.

The openly accessible taxi dataset used in this work is obtained from New York City Taxi and Limousine Commission (TLC). New York TLC provides taxi record data for green and yellow taxis in New York. We retrieve the dataset from New York City TLC, including a month’s worth of green taxi trip records obtained every Wednesday in June 2016. In summary, the datasets contain 21,272 recorded trips. Each green taxi trip record contains the following fields:

- **Trip ID:** the unique trip ID
- **Pick-up Coordinate:** the coordinate (longitude and latitude) where the taxi picks up the passenger.
- **Drop-off Coordinate:** the coordinate (longitude and latitude) where the taxi drops off the passenger.
- **Pick-up Time:** the time when the corresponding taxi picks up the passenger.
- **Drop-off Time:** the time when the corresponding taxi drops off the passenger.
- **Passenger Count:** the number of passengers in a one-way trip.
- **Trip Distance:** the distance traveled by the taxi in a one-way trip.
- **Fare Amount:** the fee paid by the passenger in a one-way trip.
- **Payment Type:** the type of taxi payment, including credit card and cash payments.

2.4 Density Map

The density map shows where points are concentrated in a given area, and the different colors of the points depend on their density. In this research, a density map is required to visualize the areas where people use bikes and taxis. The density map is divided into two, namely, origin density map and destination density map. The origin density map indicates where people use bikes or taxis in the areas of origin. The area of origin is the area where people depart the stations on their bikes or ride taxis from pick-up locations. The destination density map indicates where people alight their bikes or taxis upon arrival to their target destinations. The destination area is the area where people return their bikes at the arrival station or alight taxis in their drop-off locations.

The density map can be obtained from people’s usage of bikes and taxis in the city using a geographic information system (GIS). GIS is a powerful tool for spatial analysis. First, GIS plots the location points of people who use bikes or taxis in the area of origin and plots the location points of people who return the bikes or alight the taxis in the destination areas. Second, GIS generates density maps using kernel density on the basis of the location points of people’s usage of bikes and taxis in the origin and destination areas.

2.5 Decision Tree

The decision tree is used in land use classification, which belongs to supervised classification. The decision tree is defined as a classification procedure that recursively partitions a dataset into smaller subdivisions on the basis of a set of tests defined at each branch or node in the tree (Friedl et al., 1997). Then, the partitioning process continues until no further splits can be made. The goal of the decision tree is to create a training model that can predict the classes or values of target variables by learning decision rules inferred from prior data/training data. Figure 4 shows that each circle is a node at which tests (T) are applied to recursively split the data into smaller groups. The labels of A, B, and C at each leaf node refer to the class labels assigned to the observations.

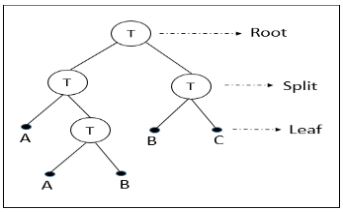


Figure 4. Decision tree classifier

2.6 Training and Testing Sample

A training sample is a set of samples that are used to construct a model. A testing sample is a set of samples that are used to estimate the accuracy of the model. A testing sample is often called an actual or ground truth sample due to its purpose of estimating model accuracy. In this research, training samples are selected from remote sensing and social sensing data. Different from the training sample, the testing sample comprises actual or ground truth data from the New York Government’s land use records that serve as reference map. Therefore, this research uses 700 training samples and 300 testing samples.

2.7 Accuracy Assessment

The accuracy assessment is performed to investigate how good the model is by estimating the percentage of the testing sample from the reference map that is correctly classified by the model. Accuracy assessment is the most important step in any classification method. In this research, the confusion matrix is selected for accuracy assessment. The confusion matrix has four types of accuracy, namely, kappa coefficient, overall accuracy, producer’s accuracy, and user’s accuracy. The kappa coefficient denotes the agreement between the classification and the reference data and is used to correct any chance agreement between classes (Cohen, 1960). Overall accuracy denotes the total percentage of the reference or actual map that is correctly classified by the model. User’s accuracy means the accuracy of the map from the user’s (not the map maker) point of view. It indicates the percentage of the model that is actually present on the actual or reference. Producer’s accuracy means the accuracy of the map from the map maker’s (the producer) point of view. It indicates the percentage of the actual map or reference that is correctly classified by the model.

3. RESULTS AND DISCUSSION

To sense and analyze the discovered patterns of human behavior in the city and to classify land use using the combination of remote sensing data and social sensing data, we describe the results of the aforementioned research aims.

3.1 Human Behavior Results

To investigate human behavior for the purpose of traffic forecasting, we identify the patterns of human behavior in New York on the basis of the bike and taxi usage data obtained every Wednesday for the whole month of June in 2016. Figures 5 and 6 show the avarage bike and taxi usage during the study period. Note that a day is split into 24 hours.

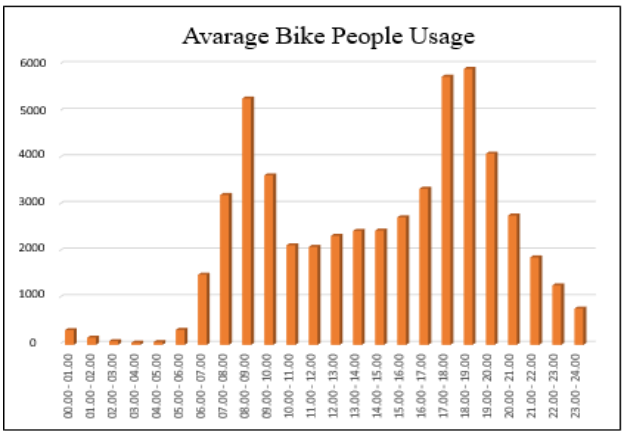


Figure 5. Avarage bike usage of people in one month

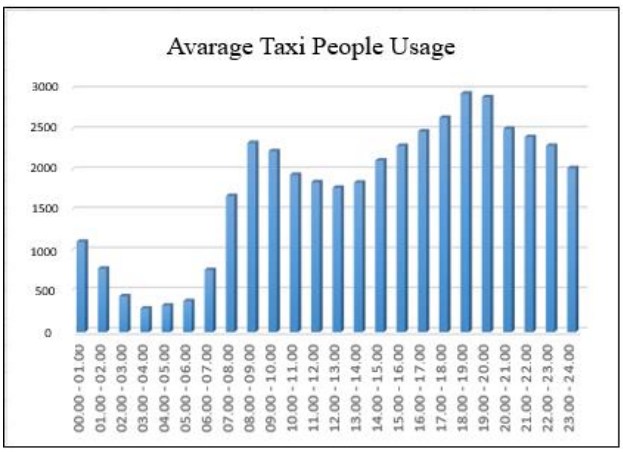


Figure 6. Avarage taxi usage of people in one month

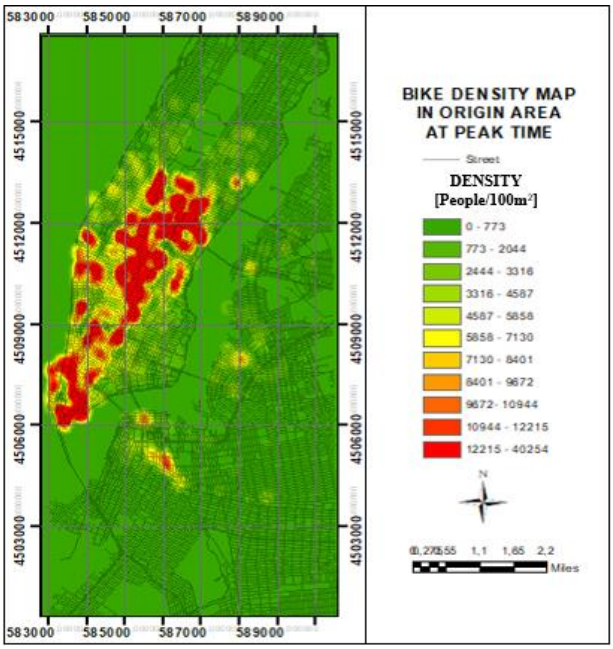
The figures provide interesting insights into the patterns of human behavior toward average taxi and bike usage in New York. Both figures show similar graphs. At 00.00–01.00, the graph shows a decline until 03.00–04.00. This outcome indicates that people seldom use bikes or taxis. At 04.00–05.00, the graph shows an increase until 08.00–09.00. This outcome indicates that people regularly use bikes or taxis to go to work. At 09.00–10.00, the graph shows a decline until 12.00–13.00. This outcome indicates that people are working. At 13.00–14.00, the graph shows an increase until 18.00–19.00. This outcome denotes lunch time and the time to return home.

Interestingly, both figures show the same peak times at 08.00–09.00 and 18.00–19.00. The peak time refers to the time when most people use bikes and taxis to go to work (08.00–09.00) and return home (18.00–19.00). Thus, the two peak times for bike and taxi usage are 08.00–09.00 and 18.00–19.00. These

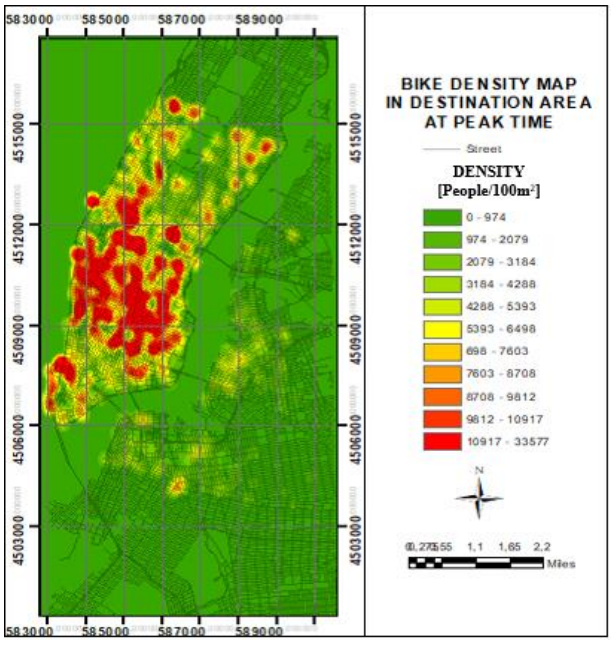
interesting patterns describe human behavior and help realize traffic forecasting.

3.2 Density Map Results

To realize the traffic forecasting purpose, this research uses GIS in generating bike and taxi density maps at peak times. The density maps are divided into two, namely, origin density map and destination density map.

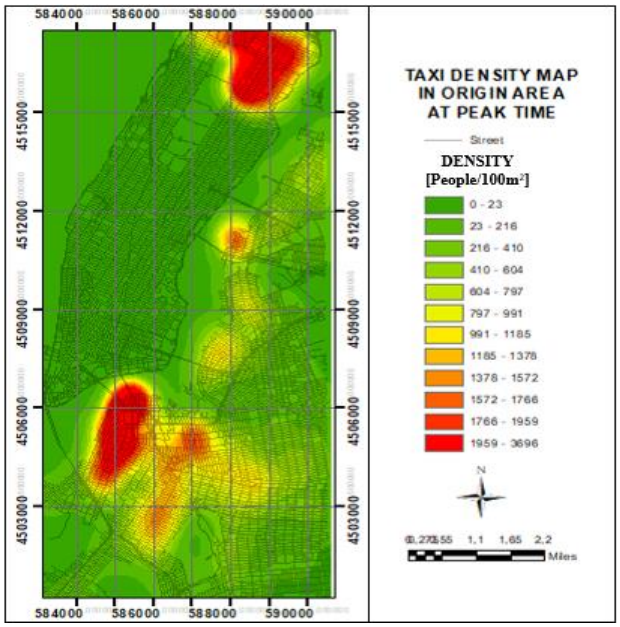


(a)

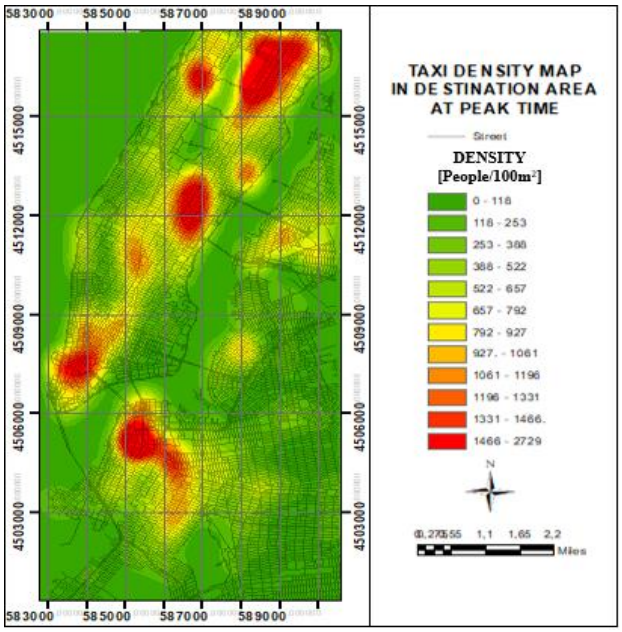


(b)

Figure 7. Bike density map at peak times: (a) in area of origin; (b) in destination area



(a)



(b)

Figure 8. Taxi density map at peak times: (a) in area of origin; (b) in destination area

Figures 7 and 8 show the density maps of bikes and taxis at peak times in the origin and destination areas. These density maps show human behavior by indicating where most people are located in the city at a given peak time. Thus, they help realize traffic forecasting. All density maps show a peak time of 08.00–09.00 in the origin and destination areas. The area of origin is the area where people depart the stations on their bikes or ride taxis from pick-up locations. The destination area is the area where people return their bikes at the arrival station or alight taxis in their drop-off locations.

The location where people are during peak time in the area of origin reflects the residential area due to people go to work via bikes or taxis upon leaving their houses in residential areas. The location where people are during peak time in the destination area reflects the office area, where people go to work via bike or taxi. The New York Government prohibits green taxis from picking up passengers within the three-quarter of Manhattan District area. Thus, the three-quarter mile area of Manhattan District in Figure 8a presents a low density of people who use taxis. Alternatively, people in Manhattan can ride yellow taxis appointed by the Government. However, yellow taxis are not allowed to pick up passengers in Brooklyn and Queens according to the rules promulgated by the New York City government.

3.3 Accuracy Assessment Comparison

In remote sensed imagery, residential and office buildings, especially those in the study area of this research, are difficult to distinguish. To overcome this problem, we apply social sensing data. Specific density maps of bikes and taxis at peak times in the origin and destination areas that reflect office and residential areas are selected as the training sample overlaid with the Sentinel imagery. The office and residential buildings in the coverage area of those density maps are likewise selected.

In this research, land use classification based on the combined use of remote sensing and social sensing data is compared with land use classification based only on remote sensing data. Table 2 shows that the overall accuracy and kappa coefficient of land use classification from remote sensing only reach 69.66% and 0.615, respectively. The result describes whether using the combination of remote sensing data and social sensing data is better than using remote sensing data only for urban land use classification in New York. The overall accuracy and kappa coefficient reach 85.3% and 0.815, respectively.

	Remote Sensing	Remote Sensing & Social Sensing
Overall Accuracy	69.66%	85.33%
Kappa Coeficient	0.615	0.815

Table 2. Comparison of the accuracy assessment results

3.4 Land Use Classification Result and Confusion Matrix

The decision tree that belongs to supervised classification is used to perform land use classification in New York, USA. The decision tree partitions the data into subsets. The partitioning process starts with a binary split and continues until no further splits can be made. The goal of the decision tree is to create a training model that can be used to predict the classes or values of target variables by learning decision rules inferred from prior data/training data.

Figure 9 shows the result of the decision tree classification based on remote and social sensing data. Five classes reflect water, park, industrial, residential, and office building areas.

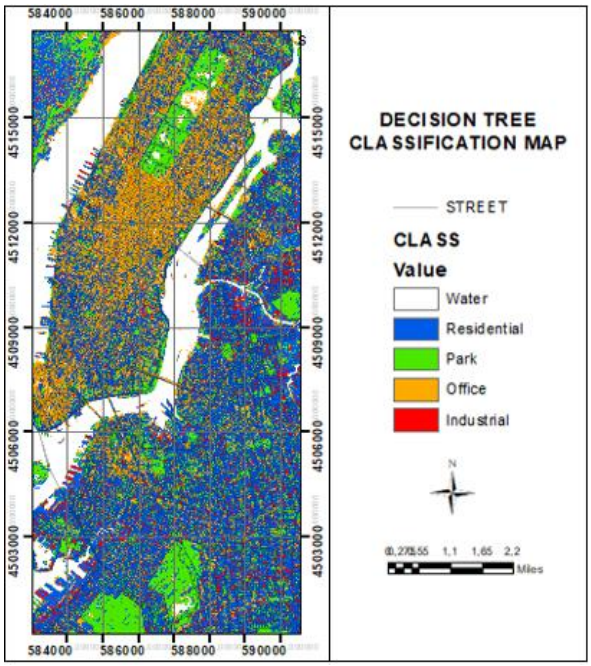


Figure 9. Decision tree classification based on social and remote sensing data

Confusion matrix of accuracy assessment is performed to investigate how good the model is by estimating the percentage of testing sample from the reference map that is correctly classified by the model. Table 3 shows the confusion matrix using remote sensing only. Table 4 shows the confusion matrix using combination semote sensing and social sensing data.

Actual / Model	Water	Park	Industrial	Office	Residential	User's Accuracy
Water	67	10	0	8	2	77.01 %
Park	2	65	0	3	2	90.28 %
Industrial	2	0	39	8	11	65.00 %
Office	2	0	5	18	15	45.00 %
Residential	2	0	6	13	20	48.78 %
Producer's Accuracy	89.33 %	86.67 %	78.00 %	36.00 %	40.00 %	**69.66 %**

Table 3. Confusion matrix using remote sensing data only

Actual / Model	Water	Park	Industrial	Office	Residential	User's Accuracy
Water	67	10	0	0	0	87.01 %
Park	2	65	0	0	0	97.01 %
Industrial	2	0	39	3	1	86.67 %
Office	2	0	6	43	5	76.79 %
Residential	2	0	5	6	42	76.36 %
Producer's Accuracy	89.33 %	86.67 %	78.00 %	86.00 %	84.00 %	**85.33 %**

Table 4. Confusion matrix using combination remote sensing and social sensing

4. CONCLUSION

Extracted bike and taxi data are successfully used to sense and analyze human behavior in New York, USA. The results help us understand when and where people are in the city. The results offer interesting insights into the patterns of human behavior toward taxi and bike usage by people in New York. The patterns are identical due to the similarities of the graphs of the bike and taxi usage patterns. The two peak times of bike and taxi usage identified in this work are 08.00–09.00 and 18.00–19.00. The two peak times show that most people who use bikes or taxis aim to fulfill their purpose of going to work and returning home. Density maps are successfully generated by GIS to show where most people are located in the city at peak times. The density maps show the locations of people who use bikes or taxis in the origin and destination areas. The density maps of the origin and destination areas also reflect the locations of residential and office areas. Therefore, these density map patterns highlight human behavior and help realize traffic forecasting in the city.

Based on social and remote sensing data, the decision tree classification successfully generates land use classification in the urban area of New York, USA. The decision tree classification shows five classes, namely, water, park, industrial, residential, and office areas. To investigate how well the decision tree result is, we perform an accuracy assessment of the confusion matrix. The number of 300 testing samples is selected from the reference map to measure the accuracy of the decision tree classification result. The overall accuracy of the decision tree classification result reaches 85.3%, and the kappa coefficient reaches 0.815. The comparison of land use classification using remote sensing only and using a combination of remote sensing and social sensing data is provided. The result of the comparison shows that using a combination of remote sensing and social sensing data is better than using remote sensing only for urban land use classification. Furthermore, the overall accuracy and kappa coefficient in this research indicates that the combination of remote sensing and social sensing data ensures an accurate urban land use classification.

REFERENCES

Latour, B., 2007. Beware, your imagination leaves digital traces, Column for Times Higher Education Supplement, 6th of April 2007. http://www.brunolatour.fr/poparticles/poparticle/P-129-THES-GB.doc

DeMaio., P. 2009. “Bike-Sharing: History, Impacts, Models of Provision, and Future,” *Journal of Public Transportation*, Vol. 12,no. 4, pp. 41–56.

Chen, L., Yang,D., Jakubowicz, J., Pan,G., Zhang,D., Li, S., 2015. Sensing the Pulse of Urban Activity Centers Leveraging Bike Sharing Open Data. In: *IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing*.

Chu, D. A. Sheets, Y. Zhao, Y. Wu, J. Yang, M. Zheng, and G. Chen., 2014. Visualizing hidden themes of taxi movement with semantic transformation. In: *Pacific Visualization Symposium (PacificVis), IEEE*, pp.137–144.

Dohuki, S., Kamw, R., Zhao, Y., Ma, C., Wu, Y., Yang,J., Ye,X., Wang,F., Li,X., Chen,W., 2017. SemanticTraj: A New Approach to Interacting with Massive Taxi Trajectories. In: *IEE Transactions on Visualization and Computer Graphics*, Vol.23, no.1.

Jensen, J.R.; Cowen, D.C., 2011. Remote sensing of urban/suburban infrastructure and socio-economic attributes. In The Map Reader; JohnWiley & Sons, Ltd.: Hoboken, NJ, USA; pp. 153–163.

Hu, S.; Wang, L., 2013. Automated urban land-use classification with remote sensing. *Int. J. Remote Sens*, 34, 790–803.

Herold, M.; Liu, X.; Clarke, K.C., 2003. Spatial metrics and image texture for mapping urban land use. *Photogramm. Eng. Remote Sens*. 69, 991–1001.

Liu, Y.; Liu, X.; Gao, S.; Gong, L.; Kang, C.; Zhi, Y.; Chi, G.; Shi, L., 2015. Social sensing: A new approach to understanding our socioeconomic environments. *Ann. Assoc. Am. Geogr*. 105, 512–530.

Satellite Imaging Corporation, 2017. Sentinel-2A Satellite Sensor (10m). https://www.satimagingcorp.com/satellite-sensors/other-satellite-sensors/sentinel-2a/ (1 June 2018).

CitiBike.2018. https://www.citibikenyc.com/ (1 June 2018).

Cohen, J. A Coefficient of agreement for nominal scales. *Educ. Psychol. Meas* 1960, 20, 37–46.

Friedl, M, A., Brodley, C, E., 1997. Decision Tree Classification of Land Cover from Remotely Sensed Data. In: *Remote Sensing Environment*. Elsevier, Vol.61, issue 3, pp. 399-409.