The Journal of Transport and Land Use jtlu.org
Vol. 16 No. 1 [2023] pp. 19–41

Using traffic data to identify land-use characteristics based on ensemble learning approaches

Jiahui Zhao
jiahuizhao@seu.edu.cn

Zhibin Li
lizhibin@seu.edu.cn

Pan Liu
liupan@seu.edu.cn

Abstract: The land-use identification process, which involves quantifying the types and intensity of human activities at a regional level, is a critical investigation step for ongoing land-use planning. One limitation of land-use identification practices is that they are based on theoretical-driven models using survey and socioeconomic data, which are often considered costly and time consuming. Another limitation is that most of these identification methods cannot incorporate the effect of daily human activity, resulting in some significant spatial heterogeneity being ignored. In this context, a novel land-use identification framework is proposed to quantify land-use characteristics using traffic-flow and traffic-events data. Regarding the identification models, two widely used Ensemble learning methods: Random Forest and Adaboost, are introduced to classify the land-use type and fit the land-use density. The case study collected the transit vehicle positions, traffic events, and geo-tagged data at the regional level in the San Francisco Bay Area, California. The results demonstrated that this framework with Ensemble learning was significantly accurate at identifying land-use characteristics in both the type classification and density regression tasks. The result averages improved 12.63%, 12.84%, 11.05%, 5.44%, 12.84% for Area Under ROC Curve (AUC), Classification Accuracy (CA), F-Measure (F1), Precision, and Recall, respectively, in classification tasks and 56.81%, 21.20%, 47.29% for Mean Squared Error (MSE), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE), respectively, in regression tasks than other models. The Random Forest model performs better in labels with high regularity, such as education, residence, and work activities. Apart from the accuracy, the correlation analysis of the error term also showed that the result was consistent with people's common sense of land-use characteristics, demonstrating the interpretability of the proposed framework.

Keywords: Land-use identification, ensemble learning, human activity data

Article history:
Received: April 18, 2022
Received in revised form: September 9, 2022
Accepted: November 26, 2022
Available online: January 17, 2023

http://dx.doi.org/10.5198/jtlu.2023.2218
ISSN: 1938-7849

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies.

1 Introduction

Land-use planning is concerned with organizing and regulating land-use to maximize the efficiency of land resources. Before conducting land-use optimization, it is critical to investigate the current situation, that is, identify the types and densities of human activities (e.g., residence, office, industry) occurring on a specific piece of land. Government-oriented planning, where the government dominates the manner and scale of future land-use that emphasizes control and longer-term goals, often causes mismatches between human activities and planning (Wu, Shan, & Choguill, 2021). It is critical to make identifications based on transportation-related data to effectively and reliably identify land-use features and fully grasp the relationship between transportation and long-term human activities.

Many researchers applied identification practices through land-use related to physical and socio-demographic factors. Numerous identification methods were carried out through land-use development regarding physical and socio-demographic characteristics. These socioeconomic data were derived from Censuses, Urban Statistical Yearbook, Urban Construction Yearbook, and Annual Statistical Bulletin (Jedwab et al., 2021; Li et al., 2020). The majority of studies in this field have analyzed the relationship between urban land-use and population densities or income, frequently expressed as various density gradients and density indices (Kasanko et al., 2006), and the compactness/degree of sprawl in urban areas (Li et al., 2020; Mendonça et al., 2020) pointed out a long-term bidirectional causal relationship between urban construction land and economic and population growth. One limitation of these statistical-based identification methods is that conducting surveys on socioeconomic factors such as household income at the acceptable sub-district level is costly and time-consuming. Another limitation is that most of these identification methods cannot incorporate the effect of daily human activity, resulting in some significant spatial heterogeneity being ignored. For instance, areas where considerable exit ridership happens in the morning and significant enter ridership occurs in the evening are more likely to be residential areas; and areas with frequent ridership happening during weekends are more likely to be commercial areas.

Previous studies utilized theoretical-driven models such as the hierarchical mixture model, clustering model, and spatially weighted regression model to evaluate land-use features. (Jia et al., 2018) proposed a method using kernel density estimation (KDE) with spatial clustering to evaluate POI data for land-use features in Suzhou Industrial Park (SIP), China. (Xu & Yang, 2019) used a geographically weighted regression model to evaluate the urban land-use plan considering transit accessibility. (Duan et al., 2021) used the multiple linear regression model to analyze the land-use characteristics of residential blocks under the guidance of walking trips. However, these models are based on simplified assumptions and incapable of capturing the complex relationship between varieties and land-use characteristics.

The growing availability of open-source urban data and the sophistication of machine learning techniques may aid in solving the limitations mentioned above. Regarding data sources, the proliferation of information technologies has dramatically increased the data resources used to track human activities in real time. (Krause & Zhang, 2019) used the land use/POI dataset merging with the GPS dataset to predict travel destinations. (Fekih et al., 2022) explored large-scale 2 G and 3 G cellular signalling data combined with land-use data to identify dynamic travel demand patterns. (Moeckel et al., 2020) proposed integrating a land-use model with a travel demand model to adapt to new trends, such as household relocation, telework and autonomous vehicles.

Recently, Machine learning methods have been widely used in transportation due to their accuracy and reliability. Ensemble learning strongly performs traffic-related tasks (Pavlyuk, 2020). By utilizing several basic units, Ensemble learning methods can be used to handle complex regression or classification problems more effectively than conventional land statistic techniques that primarily rely on simplifying assumptions. (Mendonça et al., 2020) addressed the train–test gap in traffic classification combined with the sub-flow model with Ensemble learning. (Xiao, 2019) used SVM and KNN Ensemble learning for traffic incident detection with improved robustness. (Zheng et al., 2021) proposed a joint temporal-spatial ensemble model for short-term traffic prediction since the accurate prediction of short-term traffic flow facilitates timely traffic management and rapid response. Most of these methods are applied in ITS-related tasks, such as traffic volume prediction, traffic condition, and accident estimation. In contrast, none of the Ensemble learning models has been used to study land-use characteristics.

This paper explores the identification based on two types of human activity data: traffic volume and traffic events. Given the strong connection between land-use features and traffic data (flows, events) based on previous studies, transportation-related data can infer urban land-use characteristics. (Wang & Debbage, 2021) investigated the relationship between traffic volume and land use. They explored temporal variations of pick-ups and drop-offs with various land-use types (commercial, industrial, residential, institutional, and recreational) and land-use intensity using a seven-day taxi trajectory data set collected in Shanghai. Also, many researchers have verified the correlation between land-use characteristics and traffic events/accidents. (Wang et al., 2020) looked into the interactions between roadway/land-use characteristics, traffic enforcement cameras, and the risk of regional injury/PDO crashes.. (Liu et al., 2012) used a spatial metric-based approach and panel regression to quantify the relationships between urban land use and traffic congestion. This study can identify land-use characteristics through traffic volume and event data.

Based on the limitations mentioned above and research opportunities, a novel Ensemble learning-based land-use identification framework/model is proposed to quantify land-use types and density at the Travel Analysis Zones (TAZ) level. The findings might guide ongoing land-use planning or serve as a dynamic measurement for human activities in the urban area. The specific contributions are shown as follows.

- To the best of the authors' knowledge, this is the first study to propose a quantitative method for identifying land-use characteristics using traffic flow and events data.
- In both the land-use types classification and land-use intensity regression experiments, the findings indicate that this framework with Ensemble learning is substantially more accurate at identifying land-use features than other approaches. The correlation analysis of the error term demonstrates that the result is compatible with people's common sense about land-use features, showing the interpretability of the proposed framework.

The rest of this paper is organized as follows. Section 2 describes the data. Section 3 discusses methodology. Section 4 analyses the results. The conclusions and future work are presented in section 5.

2 Data description

Quantitative analysis and identification of land-use characteristics are carried out in this study using dynamic traffic data collected from San Francisco in the United States. San Francisco has been at the forefront of the pack when implementing the open data policy. Government websites provide access to a plethora of information sources. The study period is one year beginning in the year 2020. All of the data in this paper has been aggregated into TAZ, which are the fundamental spatial units used by the Transportation Authority for transportation analyses. There are 891 TAZ in the city of San Francisco (see Figure 1). The shapefile for TAZ boundaries is obtained from the Transportation network companies (TNC) of San Francisco.

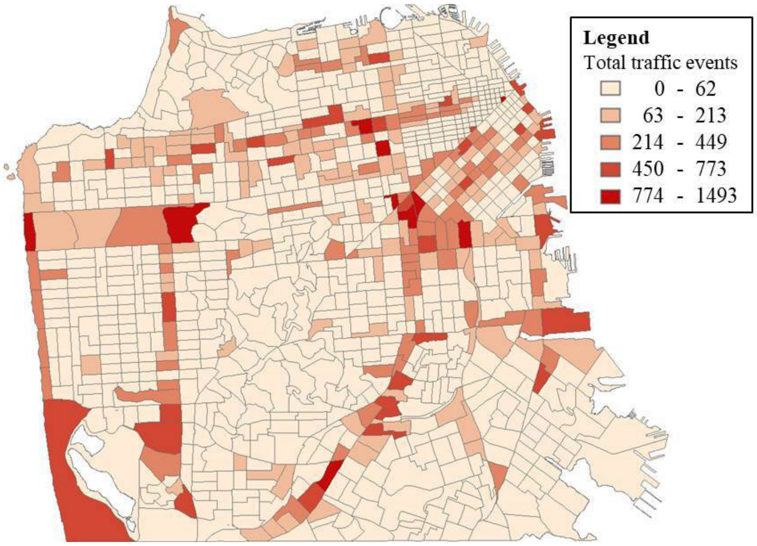


Figure 1. The TAZ boundaries and distribution of traffic events in the city of San Francisco in 2020

Data are collected in five categories: traffic events, transit vehicle data, traffic road network data, land-use data, and social demographic data. The article's structure divides the data into two sections: input data - dynamic traffic data, and output data - land-use characteristic data. Additionally, ArcGIS 10.2 aggregates the five data types into the corresponding TAZ.

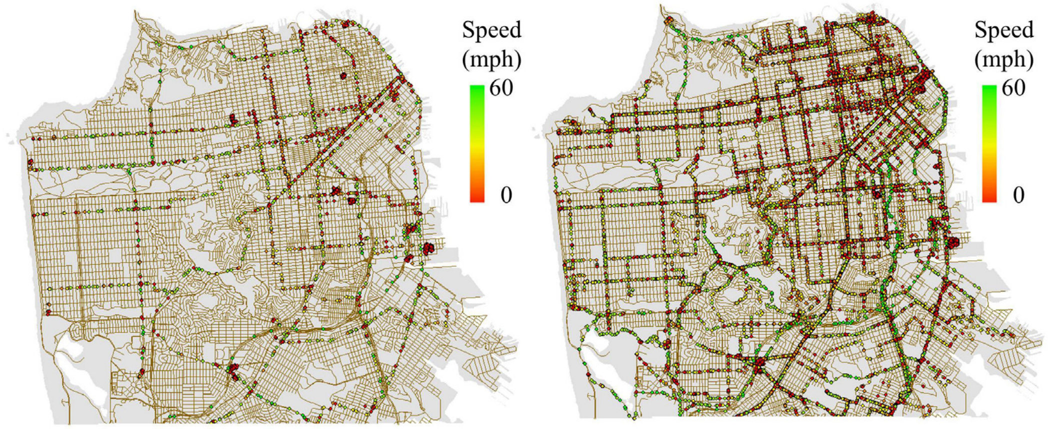
2.1 Input data description

(a) Traffic event data

Traffic event records are extracted from the LSTW dataset (Large-Scale Traffic and Weather Events Dataset). The traffic event is a spatiotemporal entity, where such an entity is associated with location and time. The variables in the traffic event dataset include temporal information, geolocation information, traffic event types, severity in terms of five severity categories, and influence distance. More specifically, traffic event types can be classified into seven categories: accident, broken-vehicle, congestion, construction, event, lane-blocked, and flow-incident. There were 181995 traffic events in the city of San Francisco during the study period, as shown in Figure 1.

(b) Transit vehicle data

The San Francisco Municipal Transportation Agency (SFMTA) provides information regarding transit vehicles. It contains taxi vehicle information for the entire city, including location date and time, vehicle identification, location latitude, longitude, heading, and speed. During the study period, 403068529 transit vehicle records were in the city of San Francisco. Figure 2 depicts the position and velocity of transit vehicles in San Francisco at midnight and in the afternoon of 2020.



7st, Apr., 2010 3:00 - 4:00

7st, Apr., 2010 17:00 - 18:00

Figure 2. Transit vehicle location and speed in the city of San Francisco

2.2 Output data description

Previous researchers collected the venue type or Point of Interest (POI) information and classified it into several activity categories. (Bao et al., 2017) identified a total of 328 detailed venue categories and divided them into seven activity categories: working, eating, entertainment, recreation, shopping, social, and education. Previous researchers defined these categories through subjectivity, and there was no strict criterion to classify them.

This paper collected two land-use datasets for building targets.

Dataset 1: The first land-use dataset consists of five different land-use classifications (residence, office, industry, commercial, and transit). This land-use data is based on the zoning map of San Francisco as of April 2020 and is subject to change, as illustrated in Figure 3.

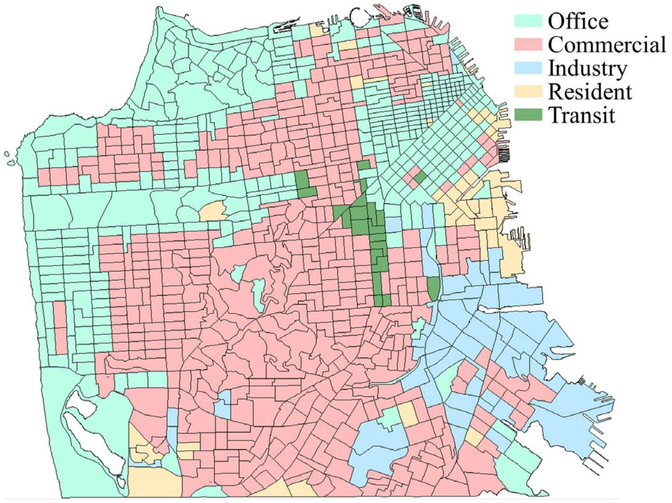


Figure 3. Spatial distribution of land-use type

Dataset 2: The second land-use dataset contains eight distinct land-use categories and their respective relative intensities, which can be used to solve regression problems. The land-use categories were developed using data from various municipal and commercial databases. There are eight categories: CIE (cultural, institutional, and educational), MED (medical), MIPS (office: management, information, and professional services), PDR (industrial: production, distribution, and repair), RES (residential), RETAIL (retail, and entertainment), VISITOR (retail, and entertainment) (hotels and visitor services) and ROAD (road).

The road data set includes road network, highway network, traffic signal data, traffic stop data, and intersection data. This road network, highway network, and intersection data is based on the City's GIS base map. Stop sign information is taken from SSD Shops Reports and parsed via python code. The SFMTA signal data is imported from the Signals Access database. The shapefile of the road network, highway network, traffic signal data, traffic stop data, and intersection data contain CNN (centerline network node) id numbers for each record. The road network contains geolocation information, speed limit, street type, and length. The Highway network contains geolocation information and highway length. Traffic signal, stop, and intersection data have geolocation information.

3 Methodology

3.1 Network architecture

This workflow aims to leverage dynamic traffic data to identify and quantify land-use information at the TAZ level, with the data collected from various sources. It is necessary to convert raw traffic event and transit vehicle data into traffic data that has been time segmented and spatially zoned in a matrix data format. It helps Ensemble learning ingest the data to complete the workflow in the first step. The workflow output is intended for classification and regression to establish the types and intensity of land-use information at the TAZ scale. Figure 4 depicts a diagram of the network architecture.

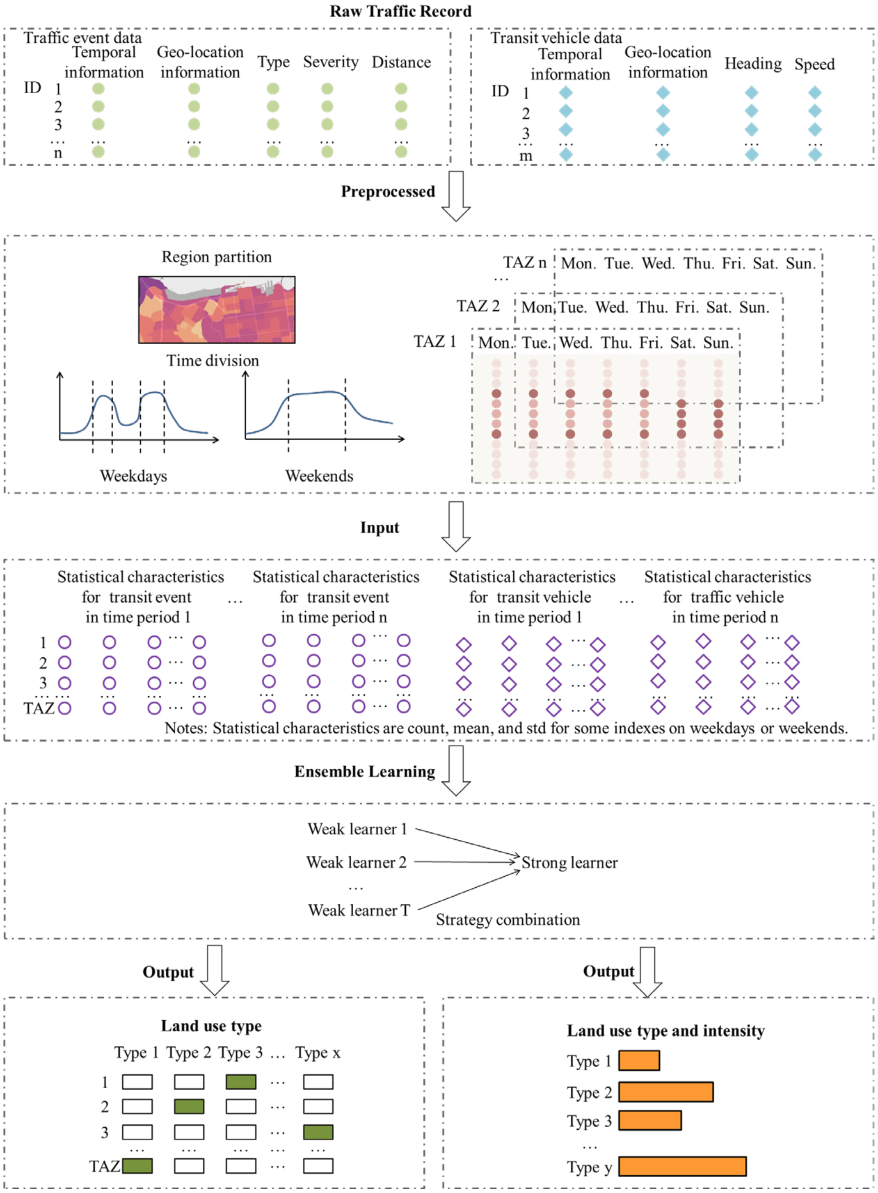

Raw Traffic Record
Traffic event data
Temporal information
Geo-location information
Type
Severity
Distance
ID 1 2 3 … n
Transit vehicle data
Temporal information
Geo-location information
Heading
Speed
ID 1 2 3 … m
Preprocessed
Region partition
Time division
Weekdays
Weekends
TAZ n
TAZ 2
TAZ 1
Mon. Tue. Wed. Thu. Fri. Sat. Sun.
Input
Statistical characteristics for transit event in time period 1
Statistical characteristics for transit event in time period n
Statistical characteristics for transit vehicle in time period 1
Statistical characteristics for traffic vehicle in time period n
1 2 3 … TAZ
Notes: Statistical characteristics are count, mean, and std for some indexes on weekdays or weekends.
Ensemble Learning
Weak learner 1
Weak learner 2
…
Weak learner T
Strong learner
Strategy combination
Output
Output
Land use type
Type 1 Type 2 Type 3 … Type x
1 2 3 … TAZ
Land use type and intensity
Type 1
Type 2
Type 3
…
Type y

Figure 4. Network architecture

3.2 Pre-processing data

3.2.1 Feature extraction methods

This section describes the pre-processing data workflow to facilitate comprehension of the proposed data feature extraction method's role and operation.

(a) For traffic event and transit vehicle data, a region partition is used to map longitude and latitude to their corresponding TAZ. This section discusses points' longitude and latitude information with the appropriate TAZ partition.

(b) Time division computes data characterization by dividing traffic data by peak period, low period, and plat hump period. Weekdays are divided into four sections, while weekends are divided into three. This paper calculates counts, percent categories of each type, percent categories of each severity, and the mean and standard deviation of influence distance for each segment.

(c) This paper calculates counts, the mean, and the standard deviation of speed for each segment. These statistical characteristics are aggregated for each part and used as model input data.

3.2.2 Land-use labelling methods

This section estimates a set of labels to describe the land-use type and intensity based on existing land-use planning data for training the ensemble model. The land zones are separated by some geographical elements (such as roads, railways, and rivers). The dimension of the land-use block is smaller than that of TAZ. Therefore, the indexes of multiple blocks are accumulated in the same TAZ to obtain the TAZ's land-use characteristics and intensity. This process is suitable for both land-use dataset 1 and land-use dataset 2.

In terms of the road index, this paper compiles data on the road network, highway network, traffic signal network, traffic stop network, and intersection network. For each TAZ, we calculate the road length, the ratio of each road type, highway length, the intersection density, the percentage of traffic signals to intersections, and the proportion of traffic stops to crossroads. This paper used a comprehensive evaluation to merge the sub-indicators into a road index. Principal component analysis (PCA) is preferred since it is an objective method for integrating the sub-indicators. This technique has been widely employed in land-use planning and transportation, which tries to extract the essential components from a collection of potentially relevant indicators. The mathematical procedure described in this article is mimicked by the description (Zhang et al., 2021).

The land-use intensity quantifies multiple land-use types in each TAZ. This paper calculates the land-use intensity in each TAZ as follows:

$$\text{LUI of type } x = \frac{\sum_{i=1}^{n} \text{type } x_i * \text{area}_i}{\sum_{i=1}^{n} * \text{area}_i}$$

where n is the number of block in each TAZ, the type x contains eight categories: CIE, MED, MIPS, PDR, RES, RETAIL, VISITOR and ROAD.

3.3 Ensemble learning

In this study, land-use variables are evaluated and statistically identified using ensemble learning. Ensemble learning involves creating and combining numerous machine learners to perform the learning tasks. Classification difficulties, regression challenges, feature selection problems, and abnormal point detection problems, among others, can all be integrated using ensemble learning.

Ensemble learning has two advantages: cure underfitting and cure overfitting. First, integrating models helps prevent underfitting. It combines all the weaker models $g(x)$ and uses collective wisdom to obtain a better model G. Integration is equivalent to feature conversion to convey a complex learning model. Second, integrating models helps prevent overfitting. It combines all the models $g(x)$ to get a more modest model G, thus avoiding some extreme situations, including the occurrence of overfitting.

3.3.1 Random Forest

Random forest (RF) is an ensemble method based on the decision of classification or regression trees developed by Breiman (2001). The radiofrequency algorithm is an enhancement to the bagging technique. To begin, RF employs a weak learner in the form of the cart decision tree. Secondly, RF increases the formation of the decision tree based on the decision tree. RF randomly selects some sample features from nodes and then chooses the optimum element from these randomly selected sample features to partition the decision tree into left and right subtrees. This model significantly strengthens the model's generalizability.

Input:

For a traffic input dataset $D=\{(x_i,y_i)\}_{i=1}^{m}$, where $x_i \in x$ and $y_i \in y$. The number of weak learning is T.

Process:

(1) For $t=1,2,\ldots,T$, the training set is sampled times t, m times in total, and the sampling set D_t containing m samples is obtained.

(2) The sampling set D_t is used to train the t^{th} decision tree model $G_t(x)$. When training the nodes of the decision tree model, select a part of the sample features from all the sample features on the node, and select an optimal quality from these randomly selected parts to make the decision tree's left and right subtree divisions.

(3) For the classification algorithm, the category or one of the categories with the most votes cast by weak learners T is the final category. For the regression algorithm, the regression results obtained by weak learners T are arithmetically averaged, and the value obtained is the final model output

Output:

Get a strong learner $G(x)$ to classify the land-use type and regress the intensity of land-use types.

3.3.2 AdaBoost algorithm

Adaptive Boosting (AdaBoost) is an ensemble machine learning model. It is a tree-based model introduced by Freund & Schapire (1997). The mechanism of the Boosting algorithm is: (a) Train a base learner from the initial training set; (b) Adjust the distribution of training samples according to the performance of the base learner; (c) Train the next base learner based on the adjusted sample distribution; (d) Repeat until the number of base learners reaches the pre-specified value; (e) Finally, all base learners are combined according to the ensemble strategy to obtain the final strong learner. Here, this paper introduces the AdaBoost algorithm for classification and regression.

(a) AdaBoost classification algorithm

Input:

For a traffic input dataset $D=\{(x_i, y_i)\}_{i=1}^{m}$, where $x_i \in x \subseteq R^n$ and $y_i \in y=\{+1,-1\}$. The learning algorithm is $\mathfrak{L}$. The number of learning is T.

Process:

(1) The algorithm initializes the distribution D (or weight) as follows:

$D_1=(\omega_{11},\ldots,\omega_{1i},\ldots,\omega_{1m}), \omega_{1i}=\frac{1}{m}, i=1,2,\ldots,m$

(2) Then for $t=1,2,\ldots,T$,

(2.1) AdaBoost algorithm builds weak models or learners $g_t=\mathfrak{L}(D,D_t)$ from the training dataset using D_t.

(2.2) Calculate the classification error rate E_t of the base learner g_t on the training data set.

$$\varepsilon_t=P(g_t(x)\neq y)= \sum_{i}^{m} w_{ti} I(h_t(x_i)\neq y_i)$$

(2.3) The weight coefficient α_t of the learners g_t is calculated as follows:

$$\alpha_t= \frac{1}{2} \cdot \ln \frac{1-\varepsilon_t}{\varepsilon_t}$$

(2.4) The distributions for the next iteration were updated as follows:

$$D_{t+1}= w_{t+1,1},\ldots,(w_{t+1,i},\ldots,w_{t+1,m})$$

$$w_{t+1,i}= \frac{w_{ti}}{Z_t} \exp(-\alpha_t y_i g_t(x_i))=\begin{cases} \frac{w_{ti}}{Z_t} e^{-\alpha t}, g_t(x_i)=y_i \\ \frac{w_{ti}}{Z_t} e^{\alpha t}, g_t(x_i)\neq y_i \end{cases}$$

where Z_t is the normalization factor

$$Z_t= \sum_{i=1}^{m} w_{ti} \exp(-\alpha_t y_i g_t(x_i))$$

(3) Construct a linear ensemble of the base learner to get the final strong learner $G(x)$,

$$G(x)= \sum_{t=1}^{T} \alpha_t g_t(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t g_t(x)\right)$$

Output:
Get strong learner $G(x)$ to classify the land-use type.

(b) AdaBoost regression algorithm

Input:
For a traffic input dataset $D=\{(x_i,y_i)\}_{i=1}^{m}$, where $x_i \in x \subseteq R^n$ and $y_i \in y$. Learning algorithm is $\mathfrak{L}$. The number of learning is T.

Process:
(1) The algorithm initializes the distribution D (or weight) as follows:
$D_1=(\omega_{11},\ldots,\omega_{1i},\ldots,\omega_{1m}), \omega_{1i}=\frac{1}{m}, i=1,2,\ldots, m$

(2) Then for t=1,2,…,T,

(2.1) AdaBoost algorithm builds weak models or learners $g_t=\mathfrak{L}(D,D_t)$ from the training dataset using D_t.

(2.2) Calculate the maximum error of samples on the training set.
$E_t= max\,|y_i-g_t(x_i)|, i=1,2\ldots m$

(2.3) Calculate the relative error of each sample.
If it is a linear error, then

$$e_{ti} = \frac{|y_i - g_t(x_i)|}{E_t}$$

If it is a square error, then

$$e_{ti} = \frac{(y_i - g_t(x_i))^2}{E_t^2}$$

If it is an exponential error, then

$$e_{ti} = 1 - \exp\left(\frac{-|y_i - g_t(x_i)|}{E_t} \right)$$

(2.4) E_t is a weighted error of the learners g_t and is given as follows:

$$E_t = \sum_{i=1}^{m} w_{ti}\, e_{ti}$$

(2.5) The weight coefficient α_t of the learners g_t is calculated as follows:

$$\alpha_t = \frac{\varepsilon_t}{1-\varepsilon_t}$$

(2.6) The distributions for the next iteration were updated as follows:

$$D_{t+1} = (w_{t+1,1}, \ldots, w_{t+1,i}, \ldots, w_{t+1,m})$$

$$w_{t+1,i} = \frac{w_{ti}}{Z_t}\, \alpha_t^{1-e_{ti}}$$

where Z_t is the normalization factor

$$Z_t = \sum_{i=1}^{m} w_{ti}\, \alpha_t^{1-e_{ti}}$$

(3) Construct a linear ensemble of the base learner to get the final strong learner $G(x)$,

$$G(x) = \sum_{t=1}^{T} \alpha_t g_t(x), = \sum_{t=1}^{T} \left(\ln \frac{1}{\alpha_t} \right) h(x)$$

where $h(x)$ is the median of $\alpha_t g_t(x)$, $t=1,2,\ldots,T$.

Output:
Get a strong learner $G(x)$ to regress the intensity of land-use types.

4 Results of data analysis

4.1 Baseline models

Several commonly used classification and regression models are employed as baseline models in this study to demonstrate the performance of the proposed Ensemble learning methods.

(a) The Support Vector Machine (SVM) (Phillips & Abdulla, 2021) is a machine learning method based on statistical theory, which improves the generalization ability of learning by seeking the minimum structured risk. It has better robustness. SVM is tested by using MATLAB 2020b, and the parameters are set as default.

(b) The k-Nearest Neighbors (KNN) was proposed by Cover and Hart in 1968, which is a relatively theoretically mature classification algorithm (Zhao et al., 2021). When an unknown sample needs to be predicted, it is determined by the K neighbors closest to the model. KNN can be used for both classification problems and regression problems. When performing classification prediction, use the K neighbors with the most significant number of categories (or the most weighted) as the prediction result; when performing regression prediction, use the average (or weighted average) of the K neighbors as the prediction result. KNN is tested by using Python 3.8.

(c) Deep neural networks (DNN) can mainly solve three problems: second classification, multi-classification, and regression (Zhu et al., 2021). The internal structure of the DNN includes an input layer, an output layer, and several hidden layers explicitly; all layers are fully connected. The input layer has neurons, and weights and biases, respectively, of the corresponding hidden layers. DNN is tested by Keras using Tensorflow.

4.2 Evaluation measurement

AUC, CA, F1, Precision, Recall, and ROC are used to evaluate the performance of classification models in this work. The percentage of correctly categorized examples is referred to as classification accuracy. Precision is defined as the fraction of true positives among positive instances. The recall is the proportion of genuine positives in the data set divided by the total number of positive cases. The area under the receiver-operating curve (AUC) is called the area under the ROC curve. F1 is the precision and recall weighted harmonic mean.

$$F1= \frac{2\times\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$$

This research uses Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) to evaluate the results. MSE is a statistic that computes the average squares of mistakes or deviations. It is the distinction between the estimator and the estimated value. MAE is used to determine the accuracy of forecasts or predictions concerning actual results. The root means square error (RMSE) is the square root of the arithmetic mean of the squares of a collection of values. It is a metric for the inaccuracy of the estimator's fit to the data.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

where N is the number of testing samples, and y and $\hat{y}$ are the actual and prediction values.

4.3 Parameters settings and candidate explanatory variables

The data is processed using Python 3.8, ArcGIS 10.2, and MATLAB 2020b. All experiments were conducted on a Windows server (Intel(R) Core i5-7200U CPU @2.50GHz 2.71 GHz). The weekly averages of transit vehicle data and traffic events are selected as an original data set. The sample set includes 3924 pieces of ridership data and land-use labels in the covered region in the Bay Area, of which 70% (2746) are for training samples, 15% (598) for validation, and 15% (589) for testing. Table 1 shows the parameters of baseline models.

Table 1. Parameters of baseline models

Model	Parameter 1	Parameter 2	Parameter 3	Parameter 4
AdaBoost	Base estimator	Number of estimators	Classification algorithm	Regression loss function
	Tree	50	SAMME.R	Linear
KNN	Number of neighbors	Metric	Weight	
	5	Euclidean	Uniform	
Neural Network	Activation	Neurons in hidden layers	Solver	Regularization
	ReLu	64	Adam	0.0001
RF	Number of trees	Growth control		
	10	5		
SVM	Cost	Regression loss epsilon	Kernel	Numerical tolerance
	1	0.1	Sigmoid	0.001

The descriptive statistics of explanatory variables across different TAZ are summarized in Table 2.

Table 2. Descriptive statistics of variables

Variable	Description	Min	Max	Mean	S.D.
Traffic events-related variables					
Total events	Traffic events count in each TAZ	0	1493	78.73	168.46
Transit vehicle data-related variables					
Total vehicle on weekends	Taxi count on weekends in each TAZ	0	131925	836.05	6487.39
Total vehicle on weekdays	Taxi count on weekdays in each TAZ	0	182262	1141.17	7712.47
Vehicle speed on weekends	Taxi speed on weekends in each TAZ (mph)	0	36.17	7.61	6.08
Vehicle speed on weekdays	Taxi speed on weekdays in each TAZ (mph)	0	42.73	9.92	5.97
Traffic road network-related variables					
Total roads	Road segments count in each TAZ	0	100	24.21	14.61
Total length	Road segments length in each TAZ (ft)	0	80696.44	14106.48	9821.74
Street percentage	Street segments /total road segments in each TAZ (%)	0	1	0.57	0.33
Avenue percentage	Avenue segments /total road segments in each TAZ (%)	0	1	0.23	0.24
Highway length	Total highway length in each TAZ (ft)	0	3399	85.27	314.80
Intersection	The number of intersections in each TAZ	0	100	18.83	13.47
Stop	Total traffic stop/intersection in each TAZ (%)	0	1	0.64	0.62
Signal	Total traffic signal/intersection in each TAZ (%)	0	1	0.14	0.23
Land-use-related variables					
CIE index	The type of cultural, institutional, and educational in each TAZ	0	434782	48641.32	
MED index	The type of medicine in each TAZ	0	838225	20259.38	
MIPS index	The type of management, information, and professional services in each TAZ	0	1979905	141318.83	
PDR index	The type of production, distribution, and repair in each TAZ	0	515686	52926.50	
RES index	The type of residence in each TAZ	0	2371	371.59	322.71
RETAIL index	The type of retail and entertainment in each TAZ	0	656973	48562.81	
VISITOR index	The type of hotels and visitor services in each TAZ	0	182184	4445.26	

4.4 Targets labelling

There are no widely accepted criteria to clarify which indicators should be used to measure the characteristics of human activity in a region. Two sets of open-source land-use-related data in San Francisco are used to label the targets.

One crucial factor is that dataset 1 is categorical data that only shows the land-use type of a given piece of the region, while dataset 2 shows the intensity of each land-use type. Two experiments are conducted to show the performance of these ML methods with these two targets.

Experiment 1: The land-use type of zones in San Francisco provided by the San Francisco zoning map is used to identify the type of land-use, as shown in figure 3. Five land-use categories are included, Office, Commercial, Industry, Resident, and Transit. Considering that not all machine learning classification methods, such as DNN, can directly use discrete variables as targets, it is necessary to convert the land-use type variables into integer or one-hot codes (Harris & Harris, 2013) before training. This categorical indicator is transferred into a one-hot code to facilitate the usage of ML-based classification.

Experiment 2: Eight indicators from dataset 2 and the road network are used to identify the intensity of land-use types. It includes CIE (cultural, institutional, and educational), MED (medical), MIPS (office (management, information, and professional services)), PDR (industrial (production, distribution, and repair)), RES (residential), RETAIL (retail, and entertainment), VISITOR (hotels and visitor services) and Road. The labels are calculated based on the labelling process shown in section 3.2. It should be noted that these indicators are normalized, and the top and lower bounds of land-use characteristics are set to 100 and 0, respectively, for comparison, as shown in Figure 5.

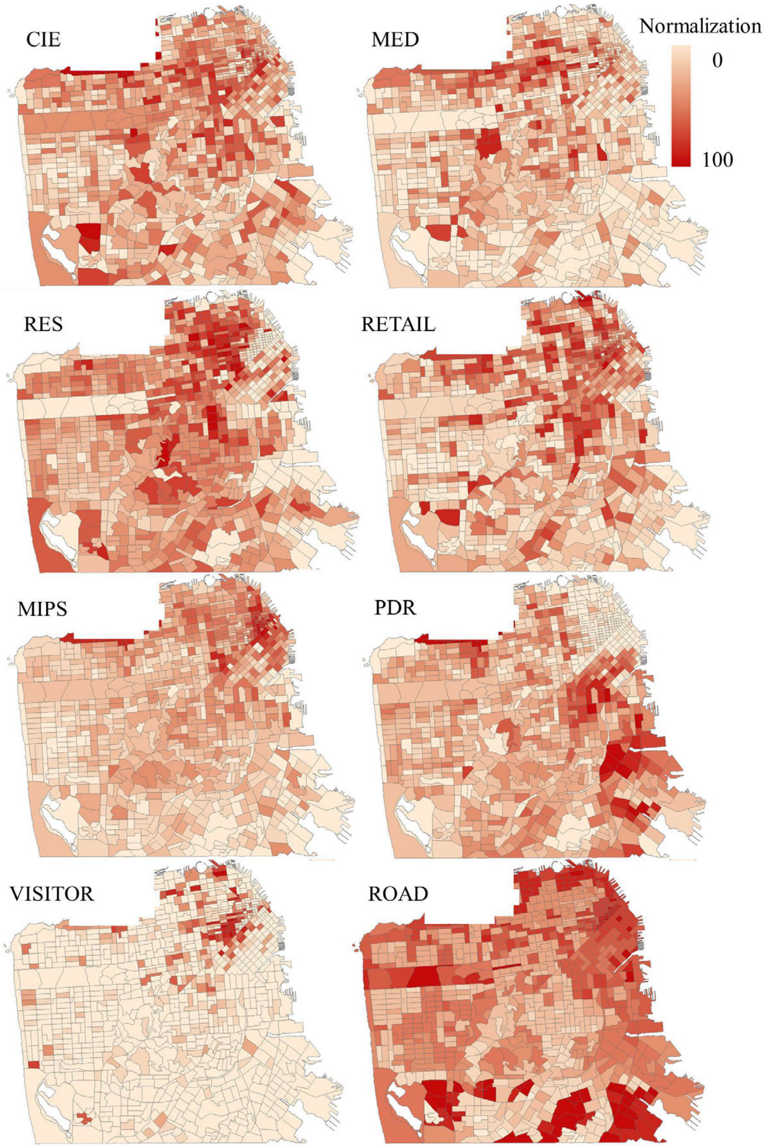


Figure 5. Spatial distributions of normalized land-use characteristics

4.5 Results analysis of land-use type in the classification task

In Table 3 and Figure 6, the results land-use type classification tasks are analyzed. Six classification methods are employed to demonstrate the usefulness of the proposed Ensemble learning approaches to the land-use identification problem: AdaBoost, KNN, Nave Bayes, DNN, RF, and SVM. Numerous performance metrics, including AUC, CA, F1, Precision, and Recall, are used to assess the performance of two suggested Ensemble learning algorithms, as shown in Table 3. In the classification tasks, the Ensemble learning methods of Random Forest outperform other methods with precision indices of 0.8143 (Office), 0.8112 (Commercial), 0.9069 (Industry), 0.9266 (Resident), and 0.9619 (Transit).

Table 3. Results of different classification models

Index: Office	AUC	CA	F1	Precision	Recall
AdaBoost	0.7435	0.7667	0.7631	0.7618	0.7667
KNN	0.7343	0.7245	0.7106	0.7256	0.7245
Naive Bayes	0.6482	0.6327	0.6348	0.6385	0.6327
Neural Network	0.7780	0.7517	0.7464	0.7493	0.7517
Random Forest	0.8576	0.8143	0.7993	0.8143	0.8143
SVM	0.4787	0.4150	0.3912	0.4672	0.4150
Index: Commercial	AUC	CA	F1	Precision	Recall
AdaBoost	0.7021	0.7020	0.7021	0.7022	0.7020
KNN	0.7506	0.7007	0.6975	0.7115	0.7007
Naive Bayes	0.6522	0.6190	0.6173	0.6226	0.6190
Neural Network	0.7732	0.7075	0.7075	0.7077	0.7075
Random Forest	0.8611	0.8109	0.8106	0.8112	0.8109
SVM	0.6099	0.5986	0.5760	0.6304	0.5986
Index: Industry	AUC	CA	F1	Precision	Recall
AdaBoost	0.7054	0.8946	0.9084	0.9246	0.8946
KNN	0.8111	0.9456	0.9309	0.9216	0.9456
Naive Bayes	0.7587	0.5850	0.6967	0.9510	0.5850
Neural Network	0.8230	0.9388	0.9223	0.9064	0.9388
Random Forest	**0.8772**	**0.9490**	**0.9274**	**0.9069**	**0.9490**
SVM	0.7434	0.9524	0.9292	0.9070	0.9524
Index: Resident	AUC	CA	F1	Precision	Recall
AdaBoost	0.6644	0.9184	0.9267	0.9357	0.9184
KNN	0.5782	0.9592	0.9425	0.9264	0.9592
Naive Bayes	0.5615	0.4728	0.6109	0.9291	0.4728
Neural Network	0.4799	0.9490	0.9374	0.9261	0.9490
Random Forest	0.8163	0.9626	0.9442	0.9266	0.9626
SVM	0.5169	0.9524	0.9391	0.9262	0.9524
Index: Transit	AUC	CA	F1	Precision	Recall
AdaBoost	0.6272	0.9524	0.9575	0.9633	0.9524
KNN	0.7364	0.9762	0.9697	0.9680	0.9762
Naive Bayes	0.7394	0.6769	0.7861	0.9726	0.6769
Neural Network	0.8865	0.9694	0.9653	0.9619	0.9694
Random Forest	0.7483	0.9762	0.9644	0.9529	0.9762
SVM	0.4410	0.9762	0.9644	0.9529	0.9762

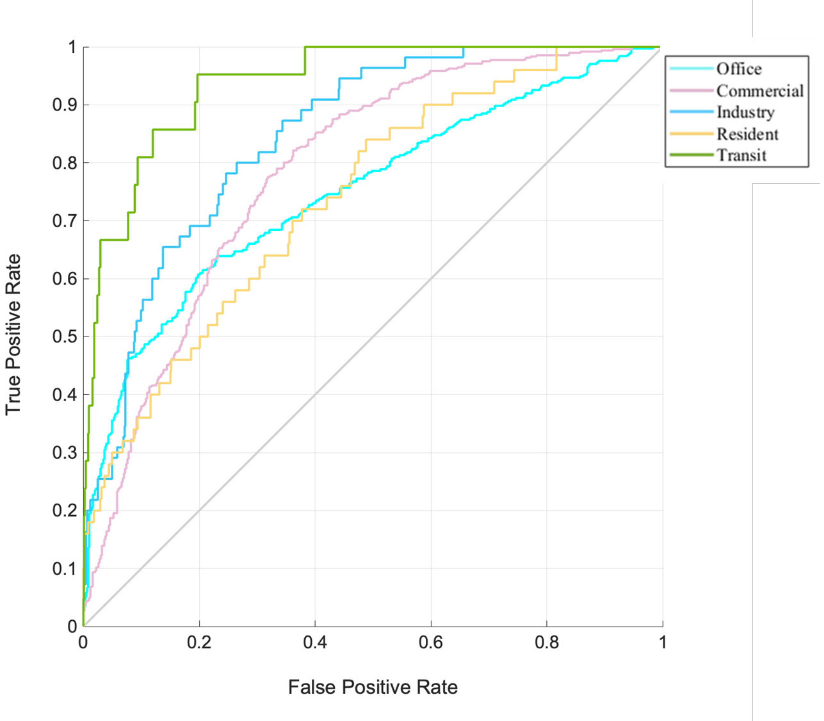


Figure 6. The ROC curve of classification

The following conclusions are made:

(a) The proposed method effectively analyses the non-linear relationship between traffic data and land-use characteristics. Among all six methods and five indicators, the average AUC is 0.703. When considering only the model with the best results (RF), it has a mean AUC of 0.832, which shows that the proposed framework and classification models effectively identify the land-use type.

(b) The Ensemble learning class (RF, Adaboost) performs well among the six approaches, particularly for RF. The average improvement rate of RF compared with other methods is 23.26%, 16.44%, 13.45%, 6.94%, and 16.44% (AUC, CA, F1, Precision, and Recall). The average difference rate of AdaBoost compared with other methods is 1.99%, 9.24%, 8.65%, 3.93%, and 9.24% (AUC, CA, F1, Precision, and Recall).

(c) Identification accuracy varies according to land-use type. Transit has the highest AUC, followed by commercial, industry, office, and residents, as seen in Table 3. The primary explanation for this anomaly is that the paper's traffic flow data is approximated using transit vehicle data. Since taxis primarily serve activities that occur in places with developed public transportation, characteristics of the traffic, production, and services are more easily recognizable. On the other hand, residential travel activities (commuting, etc.) are typically served by more cost-effective transit modes (e.g., public transport, private car, etc.). As a result, the characteristics of residential land-use are not easily defined.

The preceding two points demonstrate that the identification results for the land-use type classification problem are highly accurate and that the comparison of AUC performance among different types is interpretable.

4.6 Results analysis of land-use characteristics in a regression task

To further illustrate the effects of Ensemble learning on land-use characteristics identification, a regression task with eight types of land-use characteristics is used. Note that the classification task identifies whether the output is or not, while regression fits the value. The regression task is more challenging to conduct compared with classification tasks. The ensemble learning approaches of Random Forest and AdaBoost outperform other machine learning techniques in regression tasks. With MAE values of 9.6791 and 5.3061, the Random Forest achieves higher performance in the CIE, MIPS, and ROAD tests. With MAE values of 1.9396, 7.7664, 4.3976, 7.1906, 1.9784, and 15.7901, the AdaBoost performs better in the MED, RES, RETAIL, PDR, and VISITOR categories.

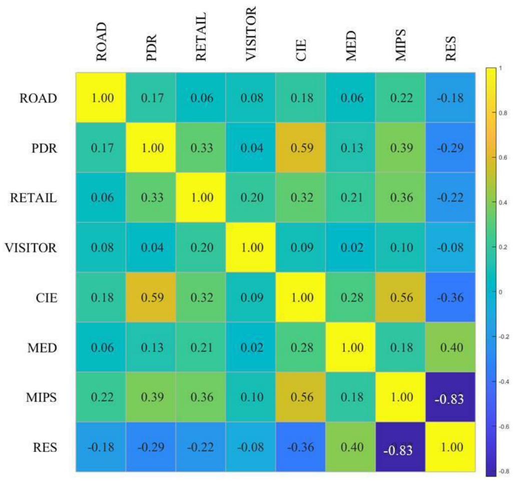
Table 4. Errors of different regression models for land-use characteristics

Index: CIE				Index: MED			
Model	MSE	RMSE	MAE	Model	MSE	RMSE	MAE
AdaBoost	292.5284	17.1035	8.3743	**AdaBoost**	32.1146	5.6670	1.9396
KNN	350.6007	18.7243	10.9243	KNN	37.8619	6.1532	2.8791
DNN	317.4010	17.8158	10.5387	DNN	34.8511	5.9035	3.0731
RF	281.4229	16.7757	9.6791	RF	35.9118	5.9926	2.8076
SVM	1156.9176	34.0135	32.7328	SVM	39.4065	6.2775	2.7915
Index: RES				Index: RETAIL			
Model	MSE	RMSE	MAE	Model	MSE	RMSE	MAE
AdaBoost	129.7341	11.3901	7.7664	**AdaBoost**	56.6257	7.5250	4.3976
KNN	191.3608	13.8333	10.4488	KNN	79.5044	8.9165	6.1269
DNN	157.4161	12.5466	9.4015	DNN	66.1084	8.1307	5.4035
RF	145.1692	12.0486	8.8998	RF	62.8751	7.9294	5.1386
SVM	223.6157	14.9538	11.9381	SVM	86.0764	9.2777	6.3094
Index: MIPS				Index: PDR			
Model	MSE	RMSE	MAE	Model	MSE	RMSE	MAE
AdaBoost	83.7651	9.1523	4.7833	**AdaBoost**	213.1752	14.6005	7.1906
KNN	107.0992	10.3489	6.2064	KNN	283.2154	16.8290	9.7930
DNN	90.4003	9.5079	5.8238	DNN	228.1301	15.1040	8.7667
RF	79.8896	8.9381	5.3061	RF	222.5345	14.9176	8.4509
SVM	169.2044	13.0079	10.8010	SVM	609.9646	24.6975	23.1050
Index: VISITOR				Index: ROAD			
Model	MSE	RMSE	MAE	Model	MSE	RMSE	MAE
AdaBoost	35.9167	5.9931	1.9784	**AdaBoost**	421.7490	20.5365	15.7901
KNN	54.6992	7.3959	3.7552	KNN	516.4865	22.7263	18.7403
DNN	39.5807	6.2913	3.6719	DNN	437.0028	20.9046	17.3060
RF	43.2343	6.5753	3.4134	RF	413.0251	20.3230	16.4886
SVM	68.7001	8.2886	6.6185	SVM	651.6228	25.5269	21.1781

In this case, some conclusions are drawn:

(a) The results show that the proposed framework can effectively identify the intensity of land-use characteristics. Previously published research fitted four land-use characteristics (residence, work, consumption, and transportation) (Zhao et al., 2020) using data from San Francisco's bike-share system, with a mean square error of 324.64 using a ten-layer DNN. This paper shows a significant performance improvement, with an average MSE of 171.36, when traffic flow and traffic events are used as inputs.

(b) When compared to other methods, Adaboost or RF always produces the best results, with an improvement of 57.95%, 2.21%, 58.52% (MSE, RMSE, MAE), and 55.68%, 20.20%, 37.54% (MSE, RMSE, MAE) compared with the average performance of other models. This conclusion shows that Ensemble learning models are more effective at solving complex recognition tasks (regression) compared with easier tasks (classification).



Note: ROAD: road; PDR: industrial; RETAIL: retail and entertainment; VISITOR: hotels, and visitor services; CIE: cultural, institutional, and educational; MED: medical; MIPS: office ;RES: residential

Figure 7. Correlation analysis of land-use characteristics identification error

This correlation matrix is calculated using the RF error sequences for each land-use characteristic indicator. Interestingly, it is closely related to our common sense, demonstrating the interpretability of the Ensemble learning model presented in this paper.

(a) Among these characteristics, PDR, RETAIL, CIE, and MIPS are all related to secondary (industry) or tertiary (services) activities, which mean that the relevant outgoing travellers will arrive during the morning peak and depart during the evening peak on weekdays, with a stop on non-working days. Correspondingly, as shown in the figure, any pair has a correlation coefficient greater than 0.3.

(b) Intuitively, the production-oriented travel characteristics differ significantly from the life-oriented indicators. The proposed ensemble model also captures this common sense. The figure shows a negative correlation coefficient between the abovementioned groups and RES. RES and MIPS are significantly negatively correlated, corresponding to the activity characteristics of large cities with high commuters.

(c) MED is a feature between production and life, as some travellers - patients - may appear at any time, while others - most staff - adhere to a regular work schedule. In this correlation analysis, the MED has a high positive correlation with the RES and a positive correlation with the production-related characteristics.

(d) One can also observe a slight positive correlation between VISITOR and RETAIL and a small positive correlation between MED and CIE. These functional overlaps also correspond to common sense.

4.7 Analysis of extension along the time dimension

This section discusses the durability of the model in different time periods. The performance of the ensemble learning model is compared based on average traffic data from January 2020 to December 2020. Since the ensemble learning model labels are based on the land-use map of April 2020 while the time range of the test set is from January 2020 to December 2020, it is unreasonable to use MSE to characterize the error. However, the reasonable result should be positively correlated with the land-use labels. The stronger the correlation, the more convincing the results are. Therefore, the performance of the ensemble learning model from January to December is measured through the correlation index R-square. The final extension results of four seasons are provided in Table 5 in the following ways to indicate how well the proposed model performs for various land-use features.

Table 5. The final extension results of four seasons

AdaBoost	CIE	MED	RES	RETAIL	MIPS	PDR	VISITOR	ROAD
Dec.-Feb.	0.3157	0.4615	0.6315	0.7246	0.4861	0.4751	0.4761	0.5704
Mar.-May.	0.3126	0.4765	0.6234	0.7562	0.4816	0.4532	0.4631	0.5674
Jun.-Aug,	0.3256	0.4613	0.6213	0.7312	0.4892	0.4265	0.4563	0.5632
Sept.-Nov.	0.3029	0.4536	0.6245	0.7245	0.4765	0.4658	0.4665	0.5786
RF	CIE	MED	RES	RETAIL	MIPS	PDR	VISITOR	ROAD
Dec.-Feb.	0.3216	0.4423	0.6312	0.7012	0.4832	0.3954	0.4365	0.6036
Mar.-May.	0.3546	0.4516	0.6214	0.6988	0.4923	0.4025	0.4826	0.5926
Jun.-Aug,	0.3248	0.4426	0.6348	0.7132	0.4928	0.3974	0.4623	0.6126
Sept.-Nov.	0.3045	0.4361	0.6313	0.7213	0.4887	0.3954	0.4589	0.5914

Several conclusions can be drawn from the extension experiment:

(a) Adaboost model performs better in the following land-use types: MED, RETAIL, MIPS, PDR, and VISITOR. The activities related to medical treatment, industry and hotel services are more random in daily human life. AdaBoost model is more sensitive to the temporospatial demand of fluctuations. On the other hand, the Random Forest model performs better in labels with high regularity. Education, residence, and work activities have regular daily activities and a clear weekday timetable. This group has contributed to rush hour on weekdays.

(b) In terms of the time dimension, the model performs relatively well in summer due to changes in demand scale. Meanwhile, the recognition of education, medical, retail and office attributes is not stable, especially for zones with balanced characteristics. The possible reason is that the human activity level is in dynamic fluctuations in highly developed urban regions such as the Bay Area.

(c) Overall, the resident and road index has the most stable performance among different months compared with other land-use types. It is easy to derive the level of residence of the zones from the promotion figures. This is due to the fact that the label of the home is based on the population density survey conducted by the planning department, which is more accurate. In addition, traffic facilities, such as road length and road type, in most areas are relatively stable and are not significantly affected by time.

5 Conclusions

Land-use identification requires data-driven approaches with high performance and input factors with dynamic and low-cost data sources. Developing Ensemble learning methods and multivariate traffic data provided opportunities for identifying land-use features and intensity. In this paper, we build a novel framework with the Ensemble learning method to quantitatively analyze and identify land-use characteristics. The results show that the proposed methods perform well in both land-use type classification and density regression experiments which average improve 12.63%, 12.84%, 11.05%, 5.44%, 12.84% (AUC, CA, F1, Precision, Recall) in classification tasks and 56.81%, 21.20%, 47.29% (MSE, RMSE, MAE) in regression tasks than other models. The Random Forest model performs better in labels with high regularity, such as education, residence, and work activities. The model performs relatively well in summer due to changes in demand scale. The outcome accurately aligns with people's common sense of urban land function, proving the suggested framework's interpretability.

In future work, we will test the proposed framework with other state-of-the-art Machine learning models on a more significant number of datasets to ensure that it is reliable. Additionally, it is necessary to apply the model for many months to determine the temporal causal link between traffic volume, traffic incidents, and the features of human activities.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (51925801, 51878165, and 71871057), the Fundamental Research Funds for the Central Universities (2242019R40060, 2242020K40063), the National Key R&D Program of China (2018YFE0120100), and Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX22_0274).

References

Bao, J., Liu, P., Yu, H., & Xu, C. (2017). Incorporating Twitter-based human activity information in spatial analysis of crashes in urban areas. *Accident Analysis & Prevention, 106*, 358–69. https://doi.org/10.1016/j.aap.2017.06.012

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Duan, Y., Lei, K., Tong, H., Li, B., Wang, W., & Hou, Q. (2021). Land-use characteristics of Xi'an residential blocks based on pedestrian traffic system. *Alexandria Engineering Journal, 60*(1), 15–24.

Fekih, M., Bonnetain, L., Furno, A., Bonnel, P., Smoreda, Z., Galland, S., & Bellemans, T. (2022). Potential of cellular signaling data for time-of-day estimation and spatial classification of travel demand: A large-scale comparative study with travel survey and land-use data. *Transportation Letters, 14*(7), 787–805. htttps://doi.org/10.1080/19427867.2021.1945854

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences, 55*(1), 119–39. https://doi.org/10.1006/jcss.1997.1504

Harris, D. M., & Harris, S. L. (2013). *Digital design and computer architecture* (2nd ed.). Amsterdam: Elsevier.

Jedwab, R., Loungani, P., & Yezer, A. (2021). Comparing cities in developed and developing countries: Population, land area, building height and crowding. *Regional Science and Urban Economics, 86,*103609. https://doi.org/10.1016/j.regsciurbeco.2020.103609

Jia, R., Khadka, A., & Kim, I. (2018). Traffic crash analysis with point-of-interest spatial clustering. *Accident Analysis & Prevention, 121,* 223–230. https://doi.org/10.1016/j.aap.2018.09.018

Kasanko, M., Barredo, J. I., Lavalle, C., McCormick, N., Demicheli, L., Sagris, V., & Brezger, A. (2006). Are European cities becoming dispersed? *Landscape and Urban Planning, 77*(1–2), 111–30. hppts://doi.org/10.1016/j.landurbplan.2005.02.003

Krause, C. M., & Zhang, L. (2019). Short-term travel behavior prediction with GPS, land use, and point-of-interest data.*Transportation Research Part B: Methodological, 123*, 349–361. https://doi:10.1016/j.trb.2018.06.012

Li, Z., Luan, W., Zhang, Z., & Su, M. (2020). Relationship between urban construction land expansion and population/economic growth in Liaoning Province, China. *Land Use Policy, 99,* 105022. https://doi.org/10.1016/j.landusepol.2020.105022

Liu, Y., Wang, F., Xiao, Y., & Gao, S. (2012). Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning, 106*(1), 73–87. https://doi.org/10.1016/j.landurbplan.2012.02.012

Mendonça, R., Roebeling, P., Martins, F., Fidélis, T., Teotónio, C., Alves, H., & Rocha, J. (2020). Assessing economic instruments to steer urban residential sprawl, using a hedonic pricing simulation modelling approach. *Land Use Policy, 92,* 104458. https://doi.org/10.1016/j.landusepol.2019.104458

Moeckel, R., Heilig, M., Hilgert, T., & Kagerbauer, M. (2020). Benefits of integrating microscopic land use and travel demand models: Location choice, time use & stability of travel behavior. *Transportation Research Procedia, 48*, 1956–1967. https://doi.org/10.1016/j.trpro.2020.08.226

Pavlyuk, D. (2020). Towards ensemble learning of traffic flows' spatiotemporal structure. *Transportation Research Procedia, 47*, 361–368. https://doi.org/10.1016/j.trpro.2020.03.110

Phillips, T., & Abdulla, W. (2021). Developing a new ensemble approach with multi-class SVMs for manuka honey quality classification. *Applied Soft Computing, 111,* 107710. https://doi.org/10.1016/j.asoc.2021.107710

Wang, C., Xu, C., & Fan, P. (2020). Effects of traffic enforcement cameras on macro-level traffic safety: A spatial modeling analysis considering interactions with roadway and land use characteristics. Accident *Analysis & Prevention, 144,* 105659. https://doi.org/10.1016/j.aap.2020.105659

Wang, M., & Debbage, N. (2021). Urban morphology and traffic congestion: Longitudinal evidence from US cities. *Computers, Environment and Urban Systems, 89,* 101676. https://doi.org/10.1016/j.compenvurbsys.2021.101676

Wu, Y., Shan, J., & Choguill, C. L. (2021). Combining behavioral interventions with market forces in the implementation of land-use planning in China: A theoretical framework embedded with nudge. *Land Use Policy, 108,* 105569. https://doi.org/10.1016/j.landusepol.2021.105569

Xiao, J. (2019). SVM and KNN ensemble learning for traffic incident detection. *Physica A: Statistical Mechanics and its Applications, 517,* 29–35. https://doi.org/10.1016/j.physa.2018.10.060

Xu, W., & Yang, L. (2019). Evaluating the urban land-use plan with transit accessibility. *Sustainable Cities and Society, 45,* 474–85. https://doi.org/10.1016/j.scs.2018.11.042

Zhang, J., Chen, M., & Hong, X. (2021). Nonlinear process monitoring using a mixture of probabilistic PCA with clusterings. *Neurocomputing, 458,* 319–326. https://doi.org/10.1016/j.neucom.2021.06.039

Zhao, D., Hu, X., Xiong, S., Tian, J., Xiang, J., Zhou, J., & Li, H. (2021). K-means clustering and KNN classification based on negative databases. *Applied Soft Computing, 110,* 107732. https://doi.org/10.1016/j.asoc.2021.107732

Zhao, J., Fan, W., & Zhai, X. (2020). Identification of land-use characteristics using bicycle sharing data: A deep learning approach. *Journal of Transport Geography, 82,* 102562. https://doi.org/10.1016/j.jtrangeo.2019.102562

Zheng, G., Chai, W. K., Katos, V., & Walton, M. (2021). A joint temporal-spatial ensemble model for short-term traffic prediction. *Neurocomputing, 457,* 26–39. https://doi.org/10.1016/j.neucom.2021.06.028

Zhu, Z., Cui, X., Zhang, K., Ai, B., Shi, B., & Yang, F. (2021). DNN-based seabed classification using differently weighted MBES multifeatures. *Marine Geology, 438,* 106519. https://doi.org/10.1016/j.margeo.2021.106519