

Article

Type Identification of Land Use in Metro Station Area Based on Spatial–Temporal Features Extraction of Human Activities

Dandan Xu ^{1,2}, Xiaodong Zhang ^{2,*}, Xinghua Zhang ¹  and Yongguang Yu ¹ ¹ School of Mathematics and Statistics, Beijing Jiaotong University, Beijing 100044, China² Beijing Municipal Institute of City Planning and Design, Beijing 100045, China

* Correspondence: zhangxiaodong@bmicpd.com.cn; Tel.: +86-10-88073732

Abstract: As a social carrier, a city is a place of human activities, and human activities also shape and influence the city through its dynamic demand. The in-depth understanding of urban functions is important for urban planning, and the full utilization of the spatial–temporal variation of human activities can be more effective in the identification of land functions. However, the complete extraction of time series features is one of the difficulties. To solve the above problems, the paper explores the identification of land use types based on human activity feature extraction, by taking Beijing as an example. Firstly, this paper constructs a time series that characterizing the change of passenger flow in the metro station area with AFC data, and realizes the feature extraction and type clustering of the time series. Secondly, this paper forms an index system for land use type identification by introducing POI-based indicators, which achieves a comprehensive representation of population activity data. Finally, this paper constructs a land use type identification model based on multi-source human activity data by using GBDT classifier. The results show that the model has high recognition accuracy. It is found that the fusion application of AFC and POI improves the land use recognition accuracy in the case of “Consistent change in time series but different types of demand”, and the different POI triggers the seemingly overall stable but random activity demand within the city. The research results promote the innovative application of open-source big data in the field of urban planning.

Keywords: metro station; time series clustering; type identification; land use



Citation: Xu, D.; Zhang, X.; Zhang, X.; Yu, Y. Type Identification of Land Use in Metro Station Area Based on Spatial–Temporal Features Extraction of Human Activities. *Sustainability* **2022**, *14*, 13122. <https://doi.org/10.3390/su142013122>

Academic Editors: Bin Meng, Dongsheng Zhan and Ran Liu

Received: 19 August 2022

Accepted: 8 October 2022

Published: 13 October 2022

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A city is a large human settlement. It can be defined as a permanent, densely populated place with administratively defined boundaries, whose members are primarily engaged in non-agricultural tasks. Urbanization is one of the most important human activities on earth, and the global urbanization process is showing an accelerated development [1]. Urbanization is the historical process by which a country or region gradually transforms its society from a traditional rural-type society dominated by agriculture to a modern urban-type society dominated by non-agricultural industries, such as industry and services, as its social productivity develops, science and technology progresses, and its industrial structure is adjusted. According to the United Nations, the proportion of urban population is predicted to reach 68% by 2050 [2]. Rapid urbanization has transformed the original natural landscape into an area for human activities. Functional monitoring and management of the area is particularly important for understanding the city, and is the key to rational urban planning and effective guidance of urban activities. Urban function identification has undergone a series of technological changes, showing a trend from passive to active perception at the data level.

Before the advent of big data era, urban functions were extracted mainly through field surveys and remote sensing [3]. Remote sensing is an important means to obtain urban information, and land use identification is one of the main tasks. Mao et al. [4]

used an object-oriented approach to multi-scale segmentation of historical images and fast recognition and classification of images by building convolutional neural networks. Kuang et al. [5] integrated remote sensing images and other information to achieve the identification of different urban land information and analyze the historical evolution of land functions. However, remote sensing image processing is mostly limited to single attributes of land or natural attributes due to the influence of technical methods, which is not suitable for urban spatial structure analysis with high requirements of spatial refinement [6]. Moreover, because of its difficulty in access, remote sensing data are difficult to use universally by researchers.

The emergence of social awareness presents an unprecedented opportunity to use open-source data for urban research. Among them, Point of Interest (POI) data are more widely used. Chi et al. [7] quantitatively identified single and mixed functions in cities by reclassifying POI data. Yu et al. [8] proposed a scheme to identify urban functions by integrating POI data and traffic cell boundaries, using traffic cells as the research unit. Yang et al. [9] fused the use of POI data and building morphology, and used AI methods and identification based on spatial calibration to achieve fine delineation of urban sites. Cao et al. [10] proposed an end-to-end deep learning fusion model of remote sensing and social perception data, and demonstrated its effectiveness in the functional identification of urban areas. These studies show that POI data have properties such as high timeliness and easy to capture; however, it cannot be used as a single data source for accurate quantification of the land use due to its low spatial area and volume description information.

As cities and human activities continue to influence each other, the advantages of cell phone signaling, social media, Automatic Fare Collection System (AFC), and other population activity data in identifying urban functions are becoming more and more obvious [11]. L Toole [12] and Noelia et al. [13] used cell phone signaling data to extract and analyze the human flow characteristics of different functional types of land, and explored the differences in human flow characteristics between different functional areas as a way to achieve urban neighborhood identification of functional land attributes. Based on the Integrated Circuit (IC) card data, Yong et al. [14] clustered the passenger flow of each station based on the K-Means++ algorithm from the perspective of the time series, and established the fitting equations of the passenger flow clustering results with the multidimensional parameters of land use features. Cao et al. [15] deduced the temporal distribution of passenger flow at metro stations based on the metro smart card data of Shenzhen, and performed a clustering analysis to achieve the identification of the types of occupational and residential land use around the metro stations. Han et al. [16] constructed an urban function identification model based on bus smart card data and POI data. Zhang et al. [17] proposed an algorithm based on convolutional neural network for the inference of land use characteristics of traffic neighborhoods based on bus and online taxi travel data sets. Gao et al. [18] used multi-scale geo-weighted regression to study the influence relationship between land use environment and morning peak outbound passenger flow, and then optimize the functional layout. In addition, Jiang et al. [19] estimated classified land use information based on social network data. In general, human activity data are highly usable in urban function identification; however, their presentation of land use functions is not well explored due to the challenges of complete feature extraction and representation. In addition, the type identification of land use based on human activity is mostly based on unsupervised classification, and its results are rarely compared with the actual land use types, which makes it impossible to discern the merits without benchmarks and difficult to discriminate the validity in terms of accuracy.

Therefore, this paper starts from the spatial and temporal characteristics of human activities, fuses and applies multi-source and open-source data to explore the intrinsic relationship between social behaviors and urban functions, and finally realizes the identification of land use types based on machine learning methods (Figure 1).

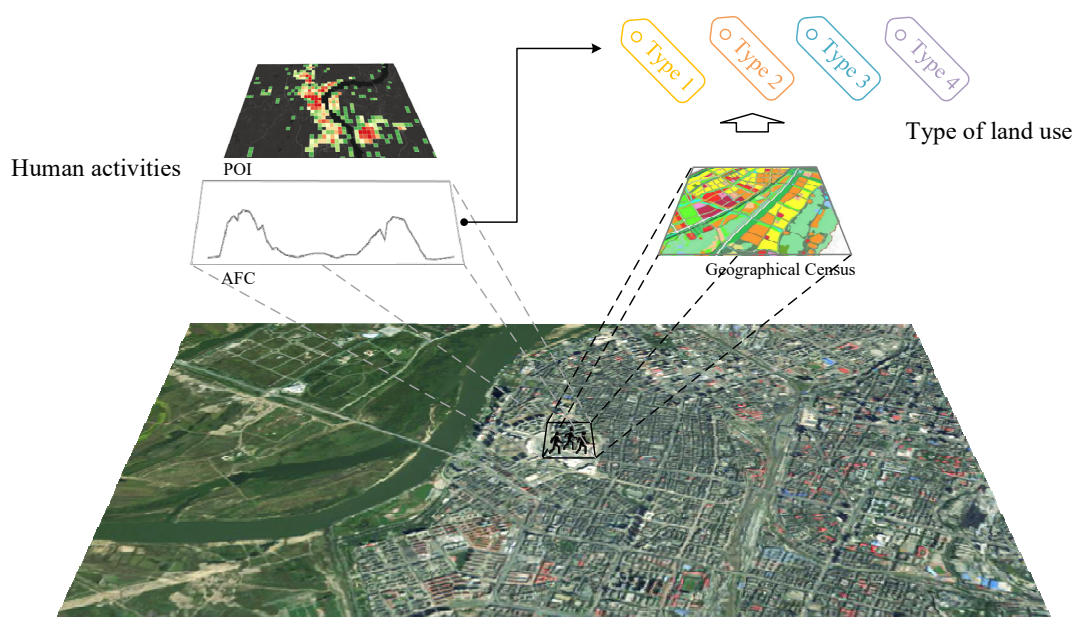


Figure 1. Research structure.

The paper is organized in the following structure: Firstly, a description of the data is given. Secondly, the research methodology is introduced. Thirdly, the spatial and temporal characteristics of the population activities are excavated. Fourthly, the land use function of the metro station area is analyzed. Finally, a land use type identification model is constructed, and the conclusions and reasons are discussed.

2. Data Collection

2.1. Auto Fare Collection Data

In recent years, with the rise of new data collection means and the strengthening of big data analysis capability, demographic activity data are widely used in urban research, and AFC data, as a typical demographic activity data, can reflect the spatial–temporal information of each metro user. The study shows that the rail passenger flow accounts for 29.1% of the total long-distance trip in Beijing in 2019, which represents the temporal pattern of residents' activities within the attraction of the metro station to a large extent. This paper takes AFC as the main data and integrates the application of POI to carry out the study.

The AFC data contain transaction information for a single passenger trip, which includes encrypted user ID, starting line and inbound station information, terminal line and outbound station information, inbound/outbound time, trip distance, and trip ID. The field information is shown in Table 1.

In this paper, the study is carried out on the example of Beijing metro stations. The AFC data of weekdays in July 2021 are selected, and the average daily passenger flow is 5.573 million. The data are aggregated by hour, and the morning peak hours for inbound and outbound passenger flows are 8:00 and 9:00, and the evening peak hours for inbound and outbound passenger flows are 18:00 and 19:00, respectively. In addition, the variance of inbound and outbound passenger flows during the morning and evening peak hours are compared separately (Figure 2). It is found that the dispersion of outbound passenger flows during the morning peak hours is high, as well as the inbound passenger flows during the evening peak hours. Differences in passenger flows due to land use types are prevalent, with commuter groups converging from different spaces of residence to go to a few of the same station areas for employment (e.g., Guomao Station, Xi'erqi Station), leading to short-term surges in passenger flows at these stations, which in turn produce significant

differences from other station flows. Therefore, the spatial–temporal data mining of human activities is important for the identification of land use types.

Table 1. AFC data information.

Field	Description	Example
USER_ID	Encrypted user ID	NVKR *****
START_LINE	Starting Line	Line 7
START_DIR	Starting direction	0
START_STATION	Inbound station ID	43
START_STATION_NAME	Inbound station name	Shuangjing
START_TIME	Entry time	06:34:41
START_LOG	Longitude coordinates of inbound station	116.46315
START_LAT	Start station latitude coordinates of inbound station	39.893464
END_LINE	Terminal line	Line 5
END_DIR	Terminal direction	0
END_STATION	Outbound station ID	53
END_STATION_NAME	Outbound station name	Dongdan
END_TIME	Outbound time	06:54:09
END_LOG	Longitude coordinate of outbound station	116.41848
END_LAT	Latitude coordinate of outbound station	39.908325
TRIP_DISTANSE	Trip distance	5156
ID	Trip ID	529360

*****: The encrypted information of the user ID.

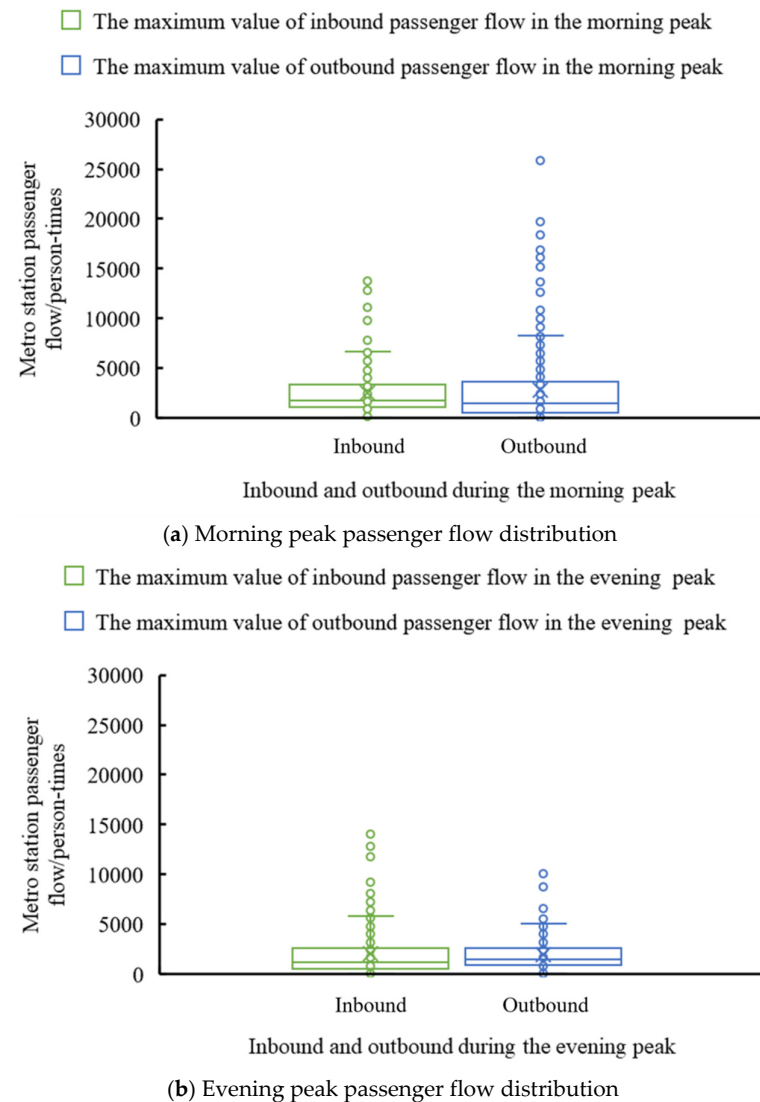


Figure 2. Comparison of passenger flow distribution during morning and evening peak hours.

2.2. Point of Interest Data

POI (Point of Interest) data refers to the point data in the Internet electronic map. With the popularity of Internet e-map services and LBS applications, POI has grown significantly in both conceptual scope and information depth. As a form of open-source data, POI is often used in the urban planning process to provide ideas for urban planning and design due to its relative ease of access. The sources of POI data in the planning field are generally from mapping companies, such as Baidu Maps, Gaode Maps, and Google Maps. The data for this study come from 2022 Gaode Map, which is obtained through open-source crawling. The data mainly include information on district, location name, address, type information (22 categories in total), spatial location, and crawling time. The details are shown in Table 2.

Table 2. POI data information.

Field	Description	Example
CITY	Administrative District	Beijing
NAME	Location Name	Huajiadi Park
ADDRESS	Location Address	150 m west of Huajiadi South Street
TYPE	POI Type	tourist attraction
LONGITUDE	Longitude	116.4641084026119
LATITUDE	Latitude	39.97475432622309
TIMESTAMP	Time	2022-05-18

3. Methodology

By ordering a set of random variables by time, the time series hide some important features, which is an important way to study social behavior. Feature analysis of a time series is generally carried out by two ideas, one is model-based clustering of the original data, such as the ARMA (Auto-Regressive and Moving Average Model) model [20] and HMM with Markov chain (Hidden Markov Models) [21]. The other idea is mainly for a high-dimensional time series, and the dimensionality reduction is achieved by feature extraction of time series, such as statistical feature-based clustering, shape-based clustering, and deep learning-based clustering. The difficulty lies in the effective expression and capture of time series features.

3.1. Tsfresh Method for Time Series Feature Extraction

Tsfresh algorithm is a feature engineering tool [4] dealing with relational databases of time series, which can automatically extract more than 1000 features from a time series. The general Tsfresh algorithm consists of three steps (Figure 3), based on feature extraction, p -value calculation, and multiple testing to complete the whole process of feature mining, feature dimensionality reduction, and type-supervised learning, respectively. Since this paper also needs to consider the POI spatial distribution features, only the feature extraction process is used in this algorithm.

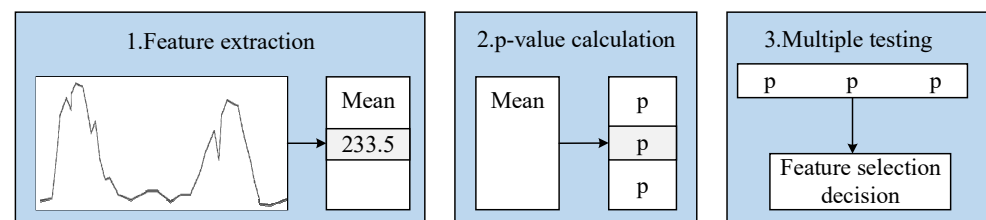


Figure 3. Description of Tsfresh algorithm.

3.2. Dynamic Time Warping Method for Time Series Clustering

Dynamic Time Warping (DTW) is one of the methods based on shape clustering (Figure 4). This method is a nonlinear regularization technique that combines time regularization and distance measurement. DTW not only eliminates the defects of Euclidean

Distance (ED) point-to-point matching, achieves one-to-many matching of data points to regularize time, and thus measures the time series of unequal length, but also has strong robustness to deviation and amplitude changes [22]. Therefore, this study is based on the DTW method for the clustering of a passenger flow time series.

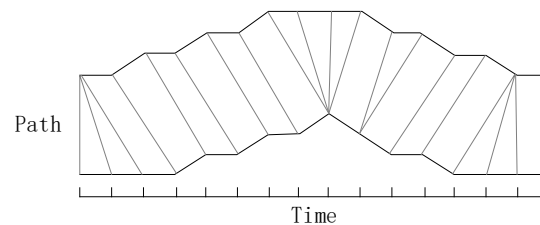


Figure 4. Curved paths between time series.

From the distance matrix D , the warping path used for distance calculation has multiple solutions. We only care about the path with the smallest distance and use this path as a similarity metric indicator. The dynamic time-warping distance between two time series Q and C is shown in Equation (1):

$$(Q, C) = \min\left(\frac{1}{K} \sum_{k=1}^k w_k\right), \quad (1)$$

where: $D(Q, C)$ is the dynamic distance between time series Q and C , K is the length of the time series, and w_k is the regularized path.

4. Spatial and Temporal Characteristics of Human Activities

Based on the AFC data, the hourly inbound passenger flow (TS_in), outbound passenger flow (TS_out), passenger flow difference (TS_(out-in)), and accumulated passenger flow difference (TS_ac) are constructed as the time series of human activities in the study. As shown in Figure 5, the overall human activity of Beijing metro stations shows a double-peak pattern, with the morning peak concentrated in the period of 6:00–10:00 and the evening peak concentrated in the period of 16:00–21:00. From the variation of TS_ac, the timing curve of some stations is concave and other stations are convex, and the population in the metro station area shows two special situations: attraction and dissipation. Due to the different functions of the land, the passenger flow in the same curve shape also shows differences, such as the start time of the increase and decrease of the regional passenger flow, the highest and lowest value of the passenger flow at a certain time, and the order and trend of the increase and decrease of the passenger flow in one day dimension. Therefore, detailed feature mining is needed to capture the impact of time series features on land use prediction.

For the feature analysis of the time series, one idea is to achieve dimensionality reduction by feature extraction of a time series, and another idea is to perform similarity carving of a time series. In this study, we first perform feature extraction of the AFC time series based on Tsfresh, and then perform similarity metric on passenger flow time series by DTW to obtain type labels, which are incorporated into the human activity index of each metro station area, respectively.

4.1. Feature Capture of AFC Time Series

For time series, Tsfresh can extract up to more than 1000 feature values. To avoid the influence of indicator redundancy and covariance, the MinimalFC-Parameters algorithm is used to extract the length, maximum, mean, median, minimum, standard_deviation, sum_values, and variance variables for each time series. Since the lengths of the sequences were all the same (04:00–24:00), the Length variables were excluded, and a total of 28 characteristic indicators of TS_in, TS_out, TS_(out-in), and TS_ac sequences were retained.

4.2. Classification of AFC Time Series

The eigenvalues of the time series describe to some extent the characteristics of the population situation over time, and may also ignore the pattern of change in the overall pattern of the time series. Therefore, in addition to extracting the features, the TS_ac time series is clustered and the similarity measure is implemented using DTW, and the results show that the best results are obtained when the number of clusters is seven. The seven clusters of stations after clustering are shown in Figure 5 in which the first and second clusters of stations have decreasing morning peak population and are relatively stable during the flat peak period and rebounding during the evening peak period, showing more obvious commuting characteristics of residence. The fourth and seventh clusters of stations have an increasing morning peak population and are relatively stable during the flat peak period and rebounding during the evening peak period, showing more obvious commuting characteristics of workplace; the third, fifth, and sixth clusters are relatively stable.

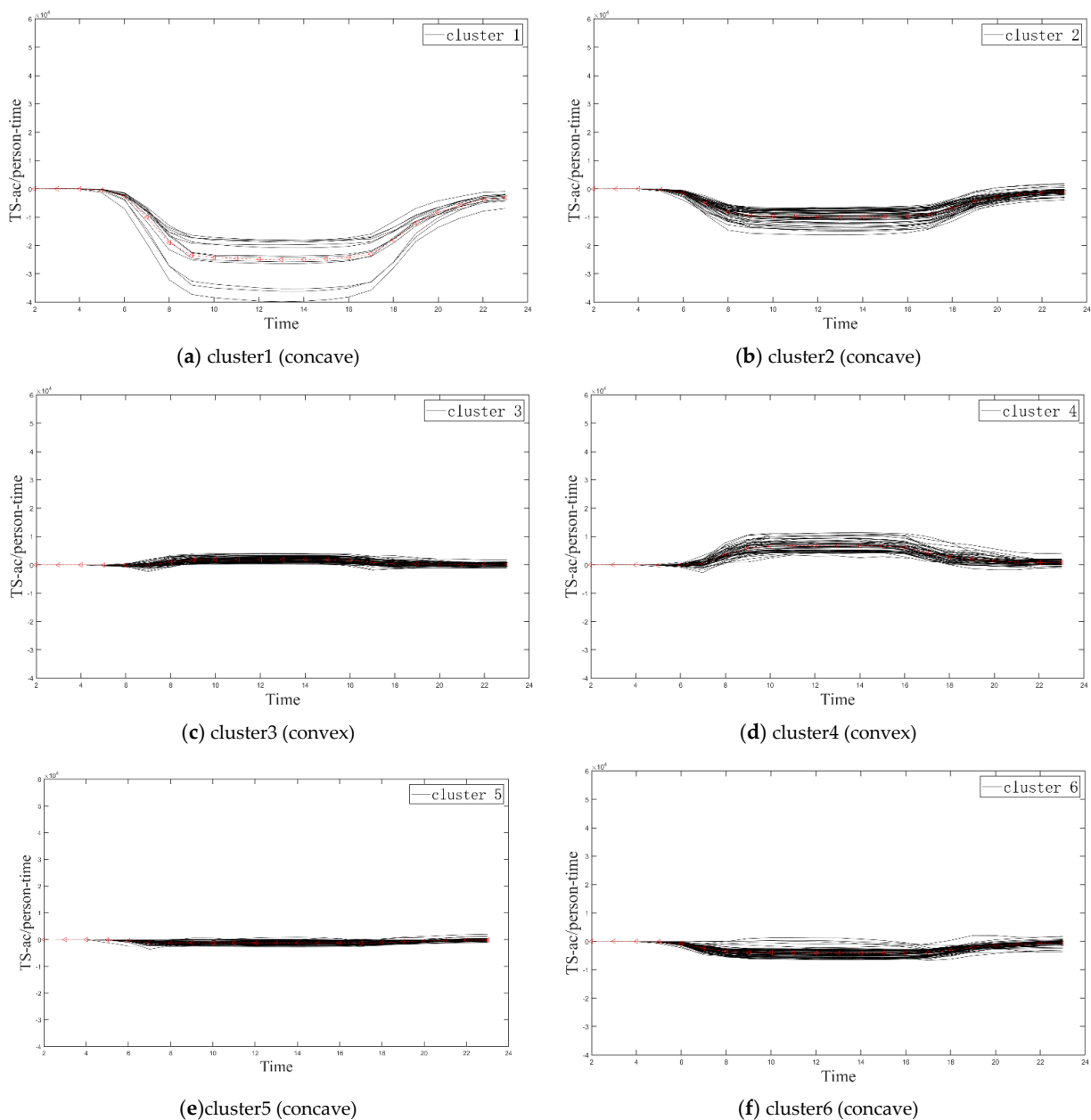
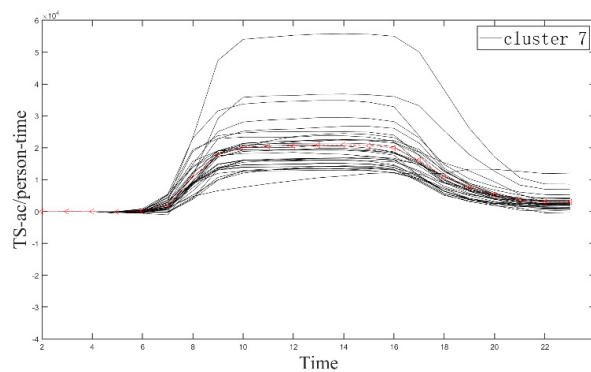


Figure 5. Cont.



(g) cluster7 (convex)

Figure 5. Time series of accumulated passenger flow difference (TS_ac) after clustering.

4.3. Spatial Distribution Feature Extraction of POI

The study is conducted using the Gaode Map POI data. Compared with previous years' data, the latest available POI data (2022) were cleaned and eliminated for invalid information. The “total number”, “proportion”, and “mixed entropy” of 22 types of POIs, including transportation facility services, accommodation services, sports and leisure services, and public facilities, are calculated for the metro station area to which they belong. Among them, the concept of “mixed entropy” is derived from information theory and is used to measure the degree of uncertainty of random events in a certain experiment. In this study, the concept of entropy is used to measure the variety of urban functions, as shown in Equation (2), a larger entropy value indicates that a comprehensive land use of the station area.

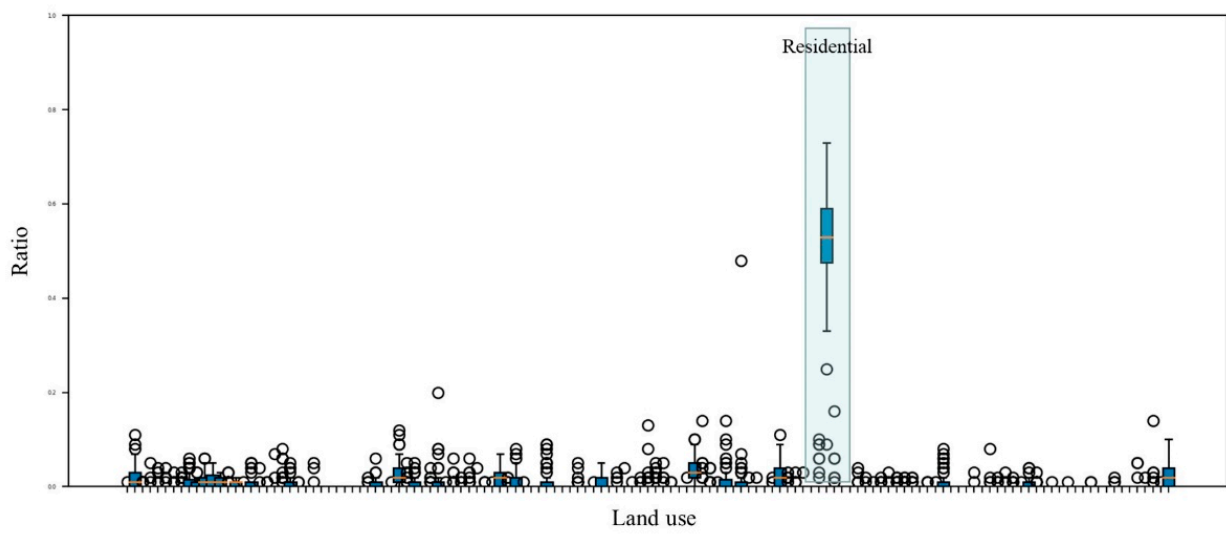
$$E = \sum_{i=1}^n p_i \log p_i \quad (2)$$

where: E is the land use mix entropy of the station area, n is the types of POI in a certain metro station area, and p_i is the proportion of a certain type of POI in the metro station area.

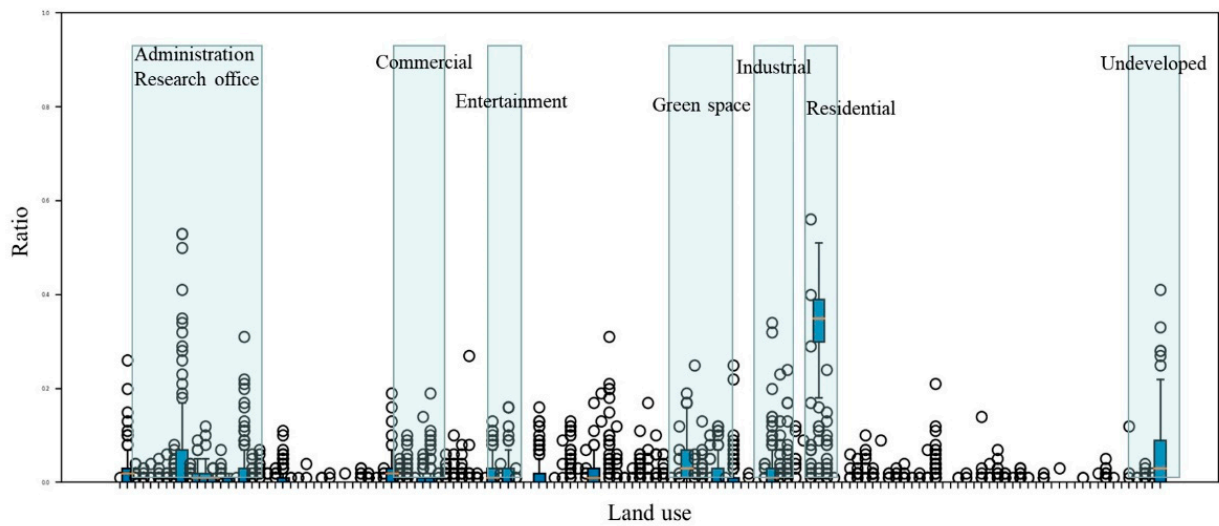
In addition to the characteristic values of the AFC time series, time series clusters, and POI distribution characteristic factors, this paper considers the urban development intensity and functional influence brought by urban space, and adds the loop location as one of the indicators. In summary, a total of 54 indicators (28 time series eigenvalues, 1 time series type, 24 POI indicators, and 1 location indicator) are established for the prediction of land use types in the metro station area.

5. Land Use Types in the Metro Station Area

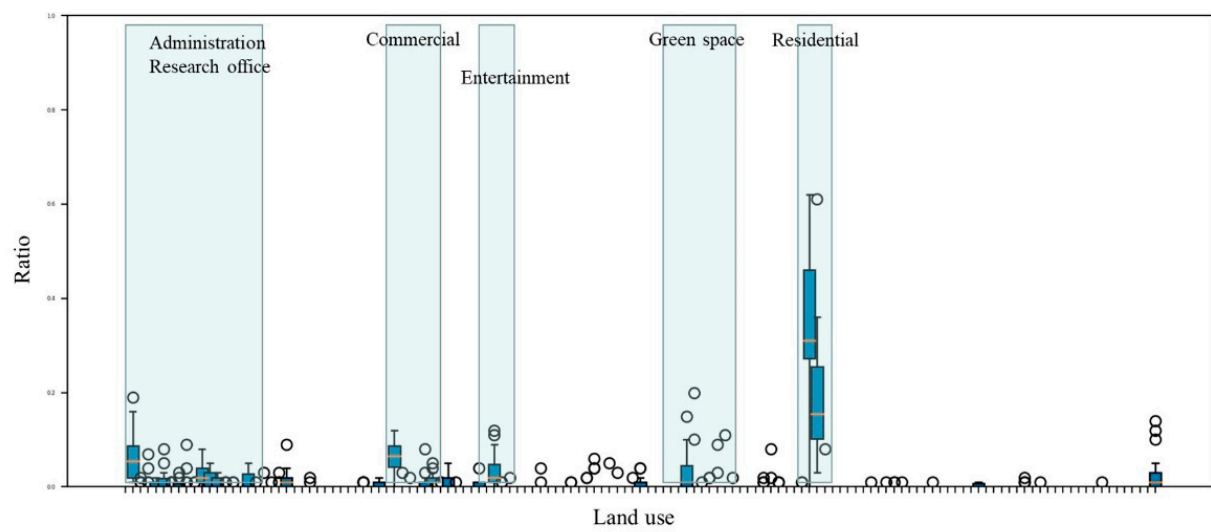
Using the “Beijing urban and rural planning land classification standard” as the benchmark, the land around Beijing’s metro stations (800-m buffer zone was selected) was functionally analyzed. After clustering by Kmeans algorithm, the functions around the Beijing rail stations were classified into the following four categories (Figure 6): (I) high-density residential-oriented single-type station area, (II) high-intensity mixed-type station area with a combination of jobs and residences, (III) low-density residential-oriented administrative center-type station area, and (IV) ecological-oriented development potential-type station area. The non-parametric test (Kruskal–Wallis test) was used to study the difference relationship between the classes, and the results showed that the 56 characteristics of the different categories of station areas were significantly different from each other, and the clustering results were valid.



(a) Land use type I

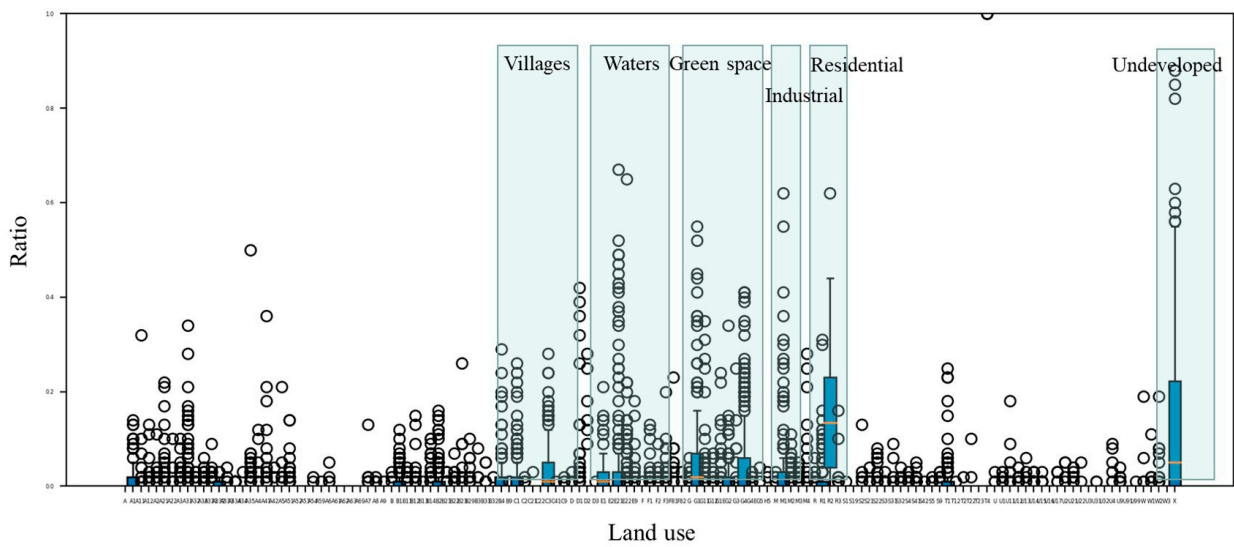


(b) Land use type II



(c) Land use type III

Figure 6. Cont.



(d) Land use type IV

Figure 6. Land use type of metro station.

Spatially (Figure 7), the high-density residential-oriented single-type stations (Type I) are mostly located in the Huitian area, Wangjing area, Shilihe area, and other large residential communities; the high-intensity mixed-type stations with a mix of jobs and residences (Type II) are mostly located in the key nodes of suburban lines such as northern Haidian, eastern Chaoyang, Fangshan, and Changping District; the low-density residential-oriented administrative center-type stations (Type III) are located in the urban core area, with low intensity of historical buildings such as “hutong”; the ecological residential-oriented. Most of the high-intensity stations (Type IV) are located between the Fifth and Sixth Ring Roads, with a high quality of living environment and a large number of land still to be determined.

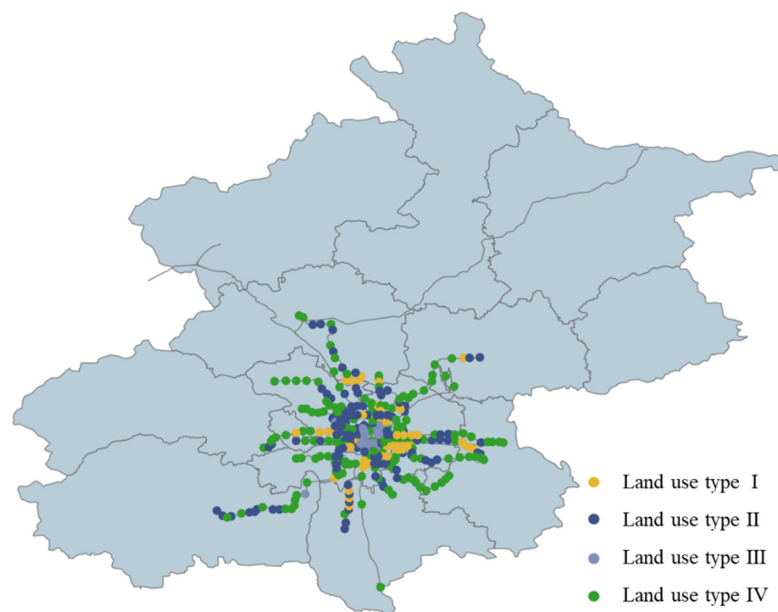


Figure 7. Spatial distribution of different types of stations.

6. Predict Land Use Types in the Metro Station Area

6.1. Gradient Boosting Decision Tree Based Land Use Type Identification Model

The Gradient Boosting Decision Tree (GBDT) is used to construct the land type recognition model, which is a type of boosting in integrated learning. The general idea is to use each weak learner to calculate the residuals between the current output and the true value, and then add up the residuals of each learner output to get close to the true value. Considering the limited number of samples, the model uses the K-fold Cross Validation (K-CV) method, which divides the original data into K groups and uses each subset as the validation set and the remaining K-1 subsets as the training set to obtain K models. The models are shown in Figure 8.

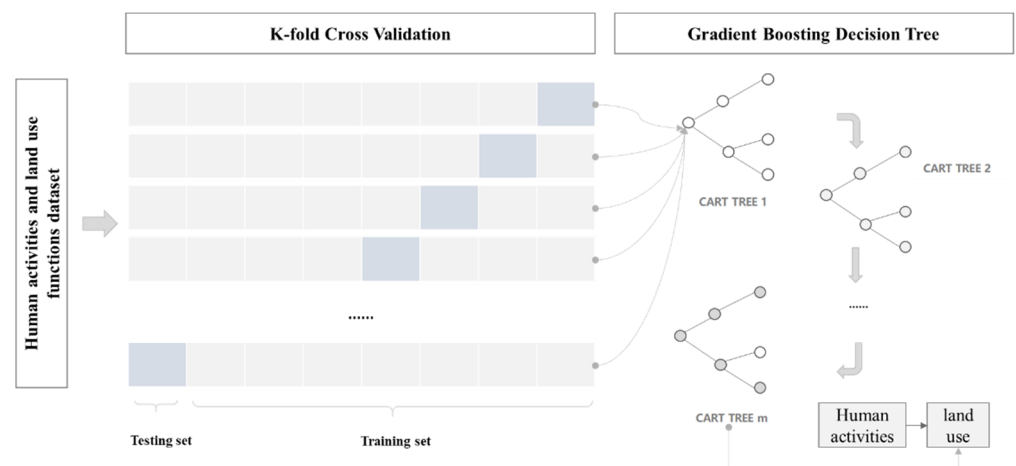


Figure 8. Model settings.

After tuning the model parameters, the final cross-validation K value was set to 30, the number of trees was 300, the maximum depth was 10, and the default log-likelihood loss function deviance was selected for the loss function. The final model accuracy was obtained with a model classification accuracy of 67.25% (accuracy: 66.92% \pm 21.35%), which is significantly better than previous models that used POI as a single data source [23]. The experimental results are shown in Table 3.

Table 3. Confusion matrix for GBDT.

	True 4	True 3	True 1	True 2	Class Precision
Pred. 4	117	1	4	34	75.00%
Pred. 3	3	19	1	4	70.37%
Pred. 1	5	0	22	10	59.46%
Pred. 2	25	2	23	72	59.02%
class recall	78.00%	86.36%	44.00%	60.00%	

6.2. Comparative Analysis of Different Classification Approaches

To explore the difference of classification accuracy, this paper compares the simulation results of other classifiers: Random Forest (RF), K-Nearest Neighbor (KNN), and Support Vector Machine (SVM), and the accuracy rates of the three identification methods are 65.79%. As is shown in Figure 9, in the prediction of the land function, the lower prediction accuracy is caused by the smaller number of samples in the third category and the poorer predictiveness of KNN and SVM for biased samples. In contrast, GBDT has better generalization ability and expressiveness on the dataset.

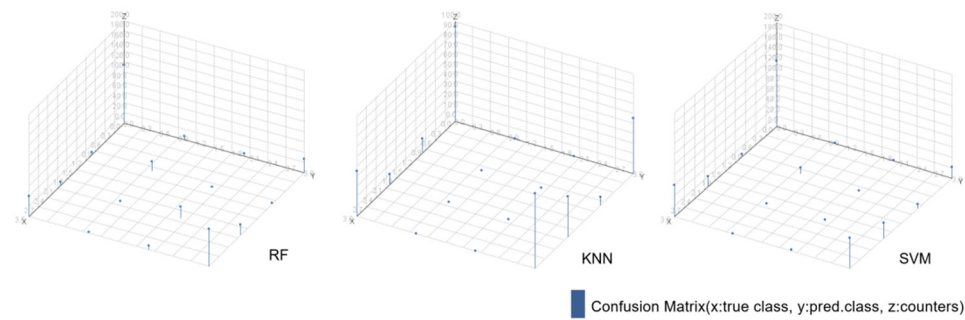


Figure 9. Confusion matrix for RF, KNN, and SVM.

6.3. Discussion and Analysis

The impact of human activities on urban functions can be more clearly portrayed by making full use of time series similarity and feature variables. The results of the weights of the attributes indicating predicted importance are shown in Figure 10. The AFC time series features show high contribution to the prediction of land use in which the median, sum_values, mean, maximum, variance, and standard_deviation of in-bound passenger flow (TS_in), outbound passenger flow (TS_out) are given higher weight. In addition, the time series type of TS_ac (Time-CLUSTER) is significant in the prediction model, which indicates that the difference in human activities is not only reflected in the total amount and dispersion level, but also in the time pattern of population activities. The conclusion further supports the initial conjecture of this paper that differences in human activities due to land types are prevalent, and that data mining of time series of activities is important for the identification of land use. The findings were similar to those of L Toole et al. [12–15]. However, the factors associated with the mechanisms of change in human activity were not discussed further in those studies.

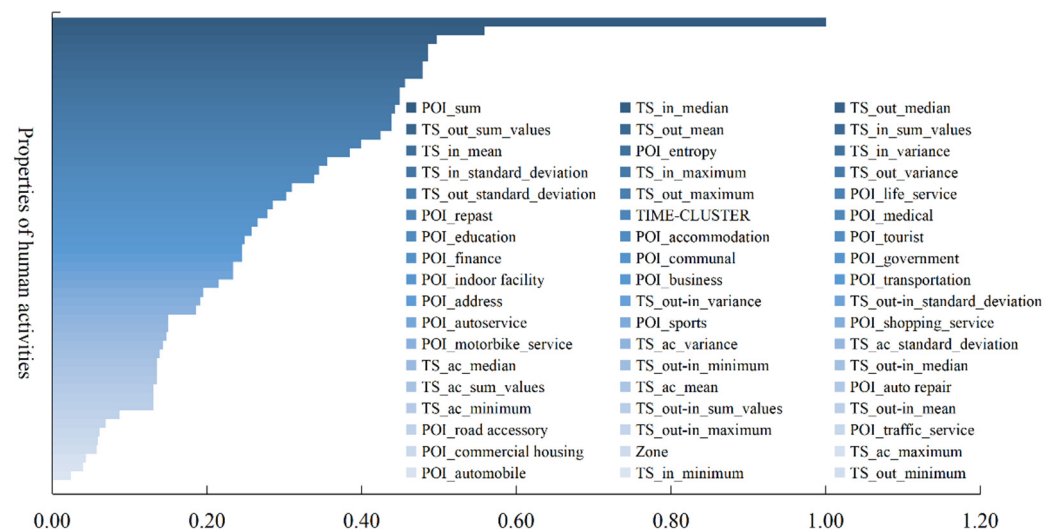


Figure 10. Weight of attributes.

The time series tends to reflect the differential population dynamic activity demand more, yet it is difficult to explain the demand triggering mechanism, so it is difficult to achieve high prediction accuracy by relying solely on time series attributes, the POI data introduced in the model compensates for this limitation.

POI has an improved effect on the model representation of time series data, which is caused by the fact that “different activity requirements may be time-series consistent in their occurrence”: POI can distinguish between cases in which “Same pattern of human activity time series, but different types of activity demands”. Further differentiation is possible. As the experimental results show (Figure 9), the total number of POIs, the mixed entropy of

POIs, and the POIs of living service facilities, catering facilities, medical facilities, scientific, educational and cultural facilities, accommodation facilities, scenic facilities, financial facilities, public facilities, and governmental facilities constitute important characteristics for predicting the type of land use in the area. This is due to the special attributes of POI in cities: POI is a key element in understanding cities, and its degree of mixing reflects the diversity of urban services in different regions; the higher the mixing entropy, the higher the chance of balancing supply and demand within the region; different urban points of interest trigger seemingly overall stable but internal random activity demand. The findings remedy the limitations of previous studies that have used POI or AFC data alone for poor land use prediction, and more importantly, the findings explain the mechanisms of interaction between human activity and land use types, contributing to a better understanding of urban dynamics [8,16,17].

The cases of station areas with different land use types were screened to compare the characteristics of time series and the differences of POI characteristics, as shown in Figure 11. Type I (high-density residential dominated single type station area, e.g., Huilongguan) shows a high intensity concentrated inbound and low intensity concentrated outbound in the morning peak, and its morning and evening differences are caused by a single residential function. Type II (high-intensity mixed station area, e.g., Qinghua East Road West Exit station area) with a certain proportion of residential land and office land has the same intensity in the morning and evening peaks. Type III (low-density residential-oriented administrative center station area, e.g., Nanluoguxiang Station) is mostly located in the historical and cultural hutong area in the core of the city, thus forming a lower intensity of residential population activities and a certain intensity of employment intensity, which is not significant compared to other types of double peaks. Type IV (ecologically oriented development potential station area, e.g., Forest Park South Gate Station) has a large area of green space or land to be developed. There are still major differences between the two at the level of entropy, i.e., diversity of land use.

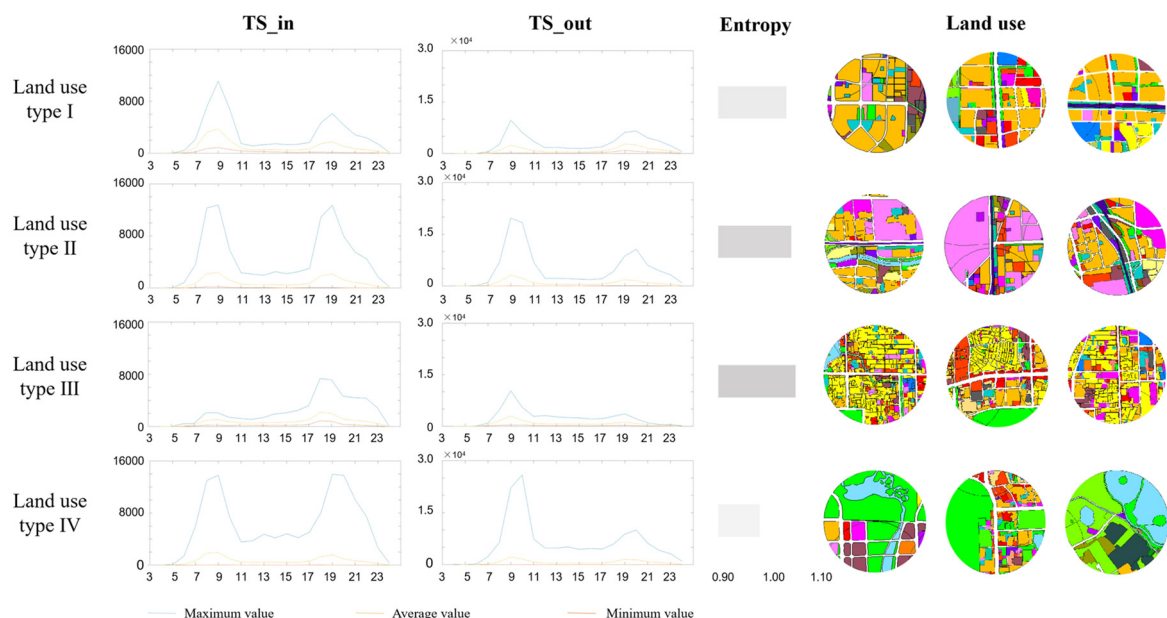


Figure 11. Time series and entropy in stations for different types of land.

The results of this study have positive implications for the planning and management of mega-cities like Beijing. For example, the findings of the study can directly support the monitoring of the status of land use types, which is necessary for the implementation of planning. Further, if we extend the study area to a larger spatial scale, we can use the model to discuss the relationship between occupants and residents. Another interesting

point is that if we take into account the temporal properties of the POI, the prediction of site types can be achieved more accurately.

7. Conclusions

As a physical carrier of social nature, the city is a place of human activity, and population activity is constantly shaping and influencing the city through its temporal dynamics and spatial differentiation characteristics. An in-depth understanding of urban functions is an important tool to assist urban planning. Therefore, this paper investigates the urban land use function from the spatial and temporal perspective of human activities.

AFC data and POI data have their own advantages in the expression level of land use functions, and the integration of the two has the potential to improve the accuracy of urban function identification, but previous studies have mostly focused on single spatial data and seldom dig deeper into this feature of temporal order. The expression of temporal features is still one of the difficulties. In this paper, we propose an urban land use function identification model that integrates AFC and POI data, and make prediction and specific analysis with the case of the Beijing metro station area. Firstly, the features of AFC time series are collected based on the Tsfresh algorithm, and the type of the AFC time series is also obtained by performing a DTW-based clustering. Secondly, the total number, proportion and mixed entropy of POIs are calculated and incorporated into the human activity index of each station area.

Through model discussions and case studies, some meaningful conclusions were obtained.

- (1) The demand for human activities brought about by different land use functions presents differences in the total amount, the degree of temporal dispersion, and the full day pattern of human activities, and such differences are widespread;
- (2) Based on the AFC time series data, the introduction of POI improves the accuracy of the model in identifying the cases where the temporal patterns of passenger flows are basically the same in terms of total volume, but different in types of demand (e.g., high-intensity mixed station area with a combination of employment and residence vs. ecologically-oriented station area with development potential). This indicates that the single time series data has limited expression of land use type, so the role of spatial distribution of points of interest needs to be considered comprehensively when modeling;
- (3) POI is a key element in forming and measuring urban vitality, and its mixing degree reflects the diversity of urban services. The higher the mixing entropy, the higher the chance of intra-regional supply and demand balance, and different POI trigger seemingly overall stable but internal random activity demand within the city.

The model proposed in this paper presents a high capability in land type identification, and the research results can promote the innovative application of open-source big data in the field of urban planning, as well as provide technical support for the investigation of land status in urban planning. Nonetheless, there are still many shortcomings in the study. In order to simplify the research scenario, this paper is launched with the metro station area as the research object, and the limited sample size affects the model accuracy to a certain extent, and we will continue to explore the possibility of model enhancement with the municipal land as the object. In addition, we will pay more attention to the interpretation of machine learning black box results in future works, and combine with practical scenarios such as urban physical examination for case application.

Author Contributions: Conceptualization, D.X.; data curation, X.Z. (Xinghua Zhang); formal analysis, D.X.; methodology, Y.Y.; project administration, X.Z. (Xiaodong Zhang); writing—original draft, D.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was sponsored by National Key Research and Development Program of China (2021YFA1000300, 2021YFA1000301) and Beijing Science and Technology Plan (Z211100004121014).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, M.X.; Zhang, H.; Liu, W.D.; Zhang, W. The global pattern of urbanization and economic growth: Evidence from the last three decades. *PLoS ONE* **2014**, *9*, e103799. [[CrossRef](#)] [[PubMed](#)]
2. United Nations. *World Urbanization Prospects: The 2018 Revision, Key Facts*; United Nations: New York, NY, USA, 2018.
3. Jiang, S.; Alves, A.; Rodrigues, F.; Ferreira, J., Jr.; Pereira, F.C. Mining point-of-interest data from social networks for urban land use classification and disaggregation. *Comput. Environ. Urban* **2015**, *53*, 36–46. [[CrossRef](#)]
4. Mao, L.D. Large-scale automatic identification of urban vacant land using semantic segmentation of high-resolution remote sensing images. *Landsc. Urban Plan.* **2022**, *222*, 104384. [[CrossRef](#)]
5. Kuang, W.H.; Zhang, S.W.; Liu, J.Y.; Shao, Q. Methodology for classifying and detecting intra-urban land use change: A case study of Changchun city during the last 100 years. *Can. J. Remote Sens.* **2010**, *14*, 345–355. (In Chineses)
6. Lu, B.Y.; Chen, Z.J.; Hou, C.H.; Wei, Q. Study on the spacial development of land urbanization in southern outskirts of Xi'an city based on GIS. *Hum. Geogr.* **2010**, *25*, 1003–2398. (In Chineses)
7. Chi, J.; Jiao, L.M.; Dong, T.; Gu, Y. Quantitative identification and visualization of urban functional area based on POI data. *J. Geomat.* **2016**, *41*, 68–73. (In Chineses)
8. Yu, B.B.; Wang, Z.B.; Chang, Y.B.; Han, J. Identify multi-level urban functional structures by using semantic data. *Sci. Surv. Mapp.* **2021**, *46*, 175–181. (In Chineses)
9. Yang, J.Y.; Shao, D.; Wang, Q.; Zhang, Y.H. Exploration on a method for precision identification of urban land use type using artificial intelligence: Based on big data of building forms and business poi data. *City Plan. Rev.* **2021**, *45*, 46–56. (In Chineses)
10. Cao, R.; Tu, W.; Yang, C.; Li, Q.; Liu, J.; Zhu, J.; Zhang, Q.; Li, Q.; Qiu, G. Deep learning-based remote and social sensing data fusion for urban region function recognition. *ISPRS J. Photogramm.* **2020**, *163*, 82–97. [[CrossRef](#)]
11. Kurowska, K.; Adamska-Kmie, D.; Kowalczyk, C.; Leń, P. Communication value of urban space in the urban planning process on the example of a Polish city. *Cities* **2021**, *116*, 103282. [[CrossRef](#)]
12. Toole, J.L.; Ulm, M.; Bauer, D. Inferring land use from mobile phone activity. In Proceedings of the 13th ACM SIGKDD International Workshop on Urban Computing, New York, NY, USA, 12 August 2012; p. 201.
13. Noelia, C.; Beníteza, F.G.; Romeroa, L.M. Land use inference from mobility mobile phone data and household travel surveys. *Transp. Res. Procedia* **2020**, *47*, 417–424.
14. Yong, J.; Zheng, L.; Mao, X.; Tang, X.; Gao, A.; Liu, W. Mining metro commuting mobility patterns using massive smart card data. *Phys. A Stat. Mech. Its Appl.* **2021**, *584*, 126351. [[CrossRef](#)]
15. Cao, R.; Tu, W.; Chao, B.C.; Luo, N.X.; Zhou, M. Identification and analysis of home and work regions in the vicinity of metro stations using smart card data. *J. Geomat.* **2016**, *41*, 74–78. (In Chineses)
16. Han, H.Y.; Yu, X.; Long, Y. Identifying urban functional zones using bus smart card data and points of interest in Beijing. *City Plan. Rev.* **2016**, *40*, 52–60. (In Chineses)
17. Zhang, Z.; Chen, Y.Y.; Liang, T.W. Inferring land use characteristics using travel patterns. *J. Transp. Syst. Eng. Inf. Technol.* **2020**, *20*, 31–35. (In Chineses)
18. Gao, D.H.; Xu, Q.; Chen, P.W.; Hu, J.; Zhu, Y. Spatial characteristics of urban rail transit passenger flows and fine-scale built environment. *J. Transp. Syst. Eng. Inf. Technol.* **2021**, *21*, 25–32. (In Chineses)
19. Christ, M.; Braun, N.; Neuffer, J.; Kempa-Liehr, A.W. Time series feature extraction on basis of scalable hypothesis tests. *Neurocomputing* **2018**, *307*, 72–77. [[CrossRef](#)]
20. Sang, A.; Li, S.Q. A predictability analysis of network traffic. *Comput. Netw.* **2002**, *39*, 329–345. [[CrossRef](#)]
21. Eddy, S.R. Hidden Markov models. *Curr. Opin. Struct. Biol.* **1996**, *6*, 361–365. [[CrossRef](#)]
22. Xu, D.; Bain; Rong, J.; Wang, J.; Yin, B. Study on clustering of free-floating bike-sharing parking time series in Beijing subway stations. *Sustainability* **2019**, *11*, 5439. [[CrossRef](#)]
23. Koo, Y.Y.; Wu, X.Y.; Huang, Z.Q.; Feng, Y.C.; Fang, C.G. Rapid determination and validation of the status quo of urban parcel land types in Luzhou city aided by POI data. *Chin. J. Agric. Resour. Reg. Plan.* **2019**, *40*, 72–79.