

Article

Site Selection of Digital Signage in Beijing: A Combination of Machine Learning and an Empirical Approach

Su Li ^{1,†}, Yuxue Wang ^{1,†}, Mengmeng Hao ^{2,*}, Dong Jiang ², Xun Zhang ^{1,2,*} and Rui Zhou ³

¹ Beijing Key Laboratory of Big Data Technology for Food Safety, School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China; 1830401028@st.btbu.edu.cn (Y.W.); lis@reis.ac.cn (S.L.)

² Key Laboratory of Resources Utilization and Environmental Remediation, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; jiangd@igsnr.ac.cn

³ Institute of Earthquake Forecasting, China Earthquake Administration, Beijing 100036, China; cnuarui@sina.com

[†] These authors contributed equally to this work.

^{*} Correspondence: zhangxun@btbu.edu.cn (X.Z.); haomm.16b@igsnr.ac.cn (M.H.); Tel.: +86-010-68985704 (X.Z.); +86-010-6488-9221 (M.H.)

Received: 11 March 2020; Accepted: 3 April 2020; Published: 4 April 2020

Abstract: With the extensive use of digital signage, precise site selection is an urgent issue for digital signage enterprises and management agencies. This research aims to provide an accurate digital signage site-selection model that integrates the spatial characteristics of geographical location and multisource factor data and combines empirical location models with machine learning methods to recommend locations for digital signage. The outdoor commercial digital signage within the Sixth Ring Road area in Beijing was selected as an example and was combined with population census, average house prices, social network check-in data, the centrality of traffic networks, and point of interest (POI) facilities data as research data. The data were divided into 100–1,000 m grids for digital signage site-selection modelling. The empirical approach of the improved Huff model was used to calculate the spatial accessibility of digital signage, and machine learning approaches such as back propagation neural network (BP neural networks) were used to calculate the potential location of digital signage. The site of digital signage to be deployed was obtained by overlay analysis. The result shows that the proposed method has a higher true positive rate and a lower false positive rate than the other three site selection models, which indicates that this method has higher accuracy for site selection. The site results show that areas suitable for digital signage are mainly distributed in Sanlitun, Wangfujing, Financial Street, Beijing West Railway Station, and along the main road network within the Sixth Ring Road. The research provides a reference for integrating geographical features and content data into the site-selection algorithm. It can effectively improve the accuracy and scientific nature of digital signage layouts and the efficiency of digital signage to a certain extent.

Keywords: digital signage; site selection; multisource information; multiscale analysis

1. Introduction

Digital signage is a multimedia audio–visual system that releases business, financial, and entertainment information through terminal display devices in public places [1,2]. Compared with traditional TV and newspaper advertisements, digital signage can be used for personalized and

customized advertising based on different audiences [3,4]. Currently, digital signage advertising has gradually become the main trend in the development of outdoor advertising, supporting the overall growth of outdoor advertising [5]. Digital signage has been developed over more than 20 years, and its application has spread to all areas of life [6,7]; in addition, its wide application has brought certain application value to society. Currently, studies on digital signage are mainly divided into two categories. The first is the study of consumers' behaviour by using digital signage. Marion et al. [8] analysed the consumer's response to digital signage advertisements in the shopping environment and found that using digital signage in malls would result in positive emotions for customers and could increase impulse purchases and store loyalty. The presence of a digital signage system reduced the perceived waiting time while also creating a favourable waiting experience [9]. In short, digital signage can promote a retail atmosphere and stimulate consumption behaviour [10–14]. The second focus is on the system development for digital signage management and content distribution. Khue et al. [15] designed a scalable real-time digital signage system that can effectively handle real-time tasks, such as emergency/instant messaging and system status monitoring. Sheppard et al. [16] used voice input and natural language processing methods to implement digital signage in the form of voice forwarding, enabling users to find information faster than with standard touch-screen-only methods. In addition, emerging technologies represented by big data technology have promoted the development of intelligent platforms for digital signage terminal management, automatic scheduling, and content distribution [17–20]. In summary, the research on digital signage mainly focuses on consumption behaviour and information system construction and lacks the site selection of digital signage. However, empirical digital signage site selection and advertisement placement are all performed manually, and there is a lack of data and method standards, which make it difficult to meet the needs of advertisers and media dealers. Therefore, precise location models should be introduced to digital signage for standardized management, which is an urgent problem for digital signage enterprises.

Site-selection models are mainly divided into empirical location models and machine learning methods. (1) Empirical location models include the gravity model [21], Huff model [22], multiplicative competitive interaction (MCI) model, multi-criteria decision-making (MCDM) model [23–25], and analytical hierarchy process (AHP) model [26–28]. Suarez-Vega et al. [22] combined the Huff model with a geographically weighted regression model to present an application in which parameters showed spatial heterogeneity and analysed the location of a new store. Amparo et al. [29] conducted a principal component analysis (PCA) of 16 types of supermarket influential factors, such as house prices, transportation centrality, and supermarket sales types, to exclude relevant influential factors and then used the MCI model to obtain supermarket location. Nevertheless, these models have a limited site-selection range and are not suitable for large-scale site selection. Considering that various factors have different impacts on the location results, Velasquez et al. [25] used a combination of the MCDM model and AHP model to select the locations of retail stores, thereby improving the effectiveness and accuracy of the location results. The Reilly model and Huff model consider the spatial distance between the audience and the facilities to be deployed for location selection; the MCI model, MCDM, and other models consider the economic and transportation information of the facilities to be located for location selection. The empirical location model can solve the location problem well in a small region or with a small amount of data. However, in the face of complex social and geographical environments, the large amount of data and calculations have increased dramatically, the computational complexity has increased, and multiple source site selection has become an NP-hard (nondeterministic polynomial) problem [30]. The empirical site-selection model is hard to solve, and there are some problems with the range and number of sites. Therefore, machine learning is introduced to adapt to large-scale calculations and to improve the intelligence and accuracy of the site-selection model to a certain extent. (2) Machine learning uses predicting methods such as decision tree and neural network to solve location problems [31–38]. Yang et al. [35] developed a HoLSAT (hotel location selection and analysis toolset) application combining WebGIS and machine learning algorithms to obtain hotel locations. Lu et al. [36] combined neural network regression prediction and the MCDM model to predict hotel deployment locations based on taxi GPS

data. Liu et al. [37] built an urban billboard site-selection system named SmartAdP based on taxi trajectory data; Wang et al. [38] constructed a hybrid BP neural network (backpropagation neural network) algorithm based on urban point-of-interest data to perform shop location selection and visualized the location results. The empirical site-selection model has the advantages of comprehensive consideration of the distance and commercial environment of the study area, but its site-selection scope and quantity are limited and difficult to map and quantify. Machine learning methods can adapt to large-scale and complex calculations. Therefore, machine learning has been introduced to improve the efficiency of big data calculations and expand the scope of site selection and the results of site selection and has gradually become a research focus.

Moreover, Gahegan M. [39] indicated that in making use of machine learning, which generally leads to models that cannot be understood by humans or related to domain theory, we can combine empirical models with machine learning methods to improve computing efficiency and the interpretability of the model. The Huff model is one of the mature theoretical site-selection models, and we innovatively improve the model to adapt it to digital signage site selection and combine it with machine learning methods to adapt to large-scale complex calculations and to improve calculation efficiency and the interpretability of the model. Therefore, this paper aims to propose a digital signage site-selection model combination of machine learning and an empirical approach based on multisource factors to improve the accuracy and interpretability of digital signage site selection models. We introduce factors such as social, economic, and demographic ones into the site selection process and effectively solve the issue of digital signage scientific layout and precision marketing.

The region within the Sixth Ring Road in Beijing was taken as the research area, and outdoor commercial digital signage was taken as the research object. This study comprehensively considers the influential factors of 19 types of digital signage, such as population, housing prices, and social network check-in, and explores how the different scales impact the site-selection process through multiscale grid analysis. The site-selection process includes three aspects: (1) A modified Huff model is used to analyse and evaluate the accessibility of digital signage, that is, the preliminary site-selection process. (2) The BP neural network model is used to predict the layout potential of digital signage. (3) The overlay analysis of digital signage accessibility, audiences, and layout potential is used to determine the site for deploying digital signage. Furthermore, the results of site selection are compared with the empirical methods to verify the validity and correctness of the site-selection results obtained in this paper. On the one hand, we provide some grid cells which are suitable for digital signage layout within the Six Ring Road of Beijing. In addition, we provide a research idea that combines empirical methods with machine learning methods and provide a reference for the integration of geographical features and their data elements in the site selection algorithm, improving the accuracy and interpretability of existing site selection models. At the same time, for other commercial facility location issues, this method can be used as a reference. Thus, it can effectively increase the accuracy of advertisement placement and the scientificity of digital signage deployment.

2. Study Area and Data Source

2.1. Study Area

Beijing is the political centre, cultural centre, technological innovation centre, and international communication centre of China. It is a modern international city. As a comprehensive megacity, the rapid development of Beijing's urban economy has also brought new opportunities to the development of the digital signage industry. Beijing is an important city where the global digital signage industry gathers. In central Beijing and its suburbs, digital signage accounts for 85% of the total signage in Beijing [20]. The 2014 Beijing Population Sample Survey Report showed that 79.5% of Beijing's permanent population is concentrated within the Sixth Ring Road [40]. The digital signage data are obtained from previous projects and describe the operational costs, screen types, and addresses of the 5,823 digital signs in the study area (Figure 1). The area within the Sixth Ring Road

in Beijing was selected as the study area (as shown in Figure 1), including Dongcheng District, Xicheng District, Haidian District, Chaoyang District, most of Fengtai District, Shijingshan District, Shunyi District, Changping District, Tongzhou District, Daxing District, Fangshan District, and parts of Mentougou District.

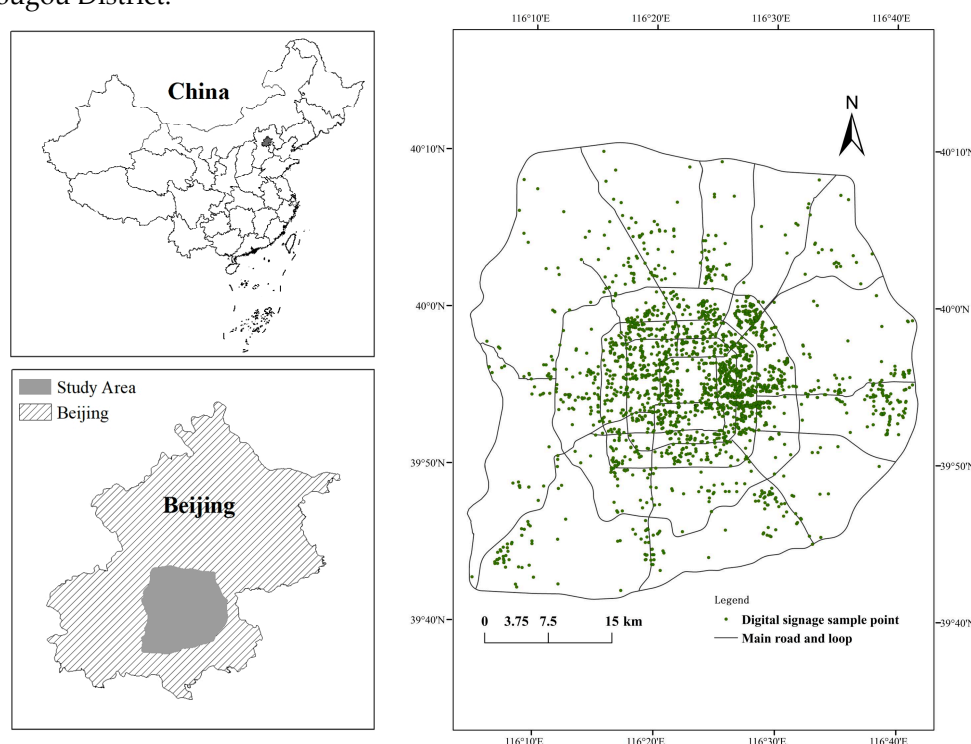


Figure 1. Study area.

2.2. Data Source

As a nonlinear complex system, the distribution of digital signage needs to comprehensively consider the interaction between various factors. Demographics, transportation, competition, and facilities are the main factors affecting the layout and location of digital signage [20,37,38]. Therefore, we select 19 indicators as shown in Table 1, including the spatial information and broadcast price of digital signage, the number of POI facilities, Sina Weibo check-in data, census data, and transportation network centrality index data (calculated based on the street network) [41,42].

Table 1. Description of Data.

Data Type	Indicators	Source
Population	Residents	The third census and the third economic census in Beijing
Traffic Count	Transportation network centrality index	Statistic calibre
Competitors	Basic data of digital signage	Project group accumulation
Social Media	Social network check-in data	Sina Weibo (https://weibo.com/)
Housing Price	Housing price	LianJia (https://bj.lianjia.com)
Point of interest (POI) Facilities	14 kinds of POI facilities data (restaurant/shopping malls/communal facilities/culture education departments/transportation services facilities/life service facilities/etc.)	Google (https://developers.google.cn/places/web-service/intro)

The digital signage assumed in this study included outdoor commercial digital signage with broadcasting video, animation, pictures or text, as well as all of the outdoor commercial large digital signage distributed within the Sixth Ring Road in Beijing based on our database established by field

investigation. The information displayed on the digital signage should have at least 100 m visibility, that is, it can be seen clearly 100 meters away from the audience.

3. Methods

The site selection of digital signage is the process of identifying new places to deploy digital signage by combining digital signage data and influential factors. The improved Huff model is an empirical approach that can calculate the distance between the audience and the digital signage and then select grids with low spatial accessibility. The machine learning methods analyse the number of commercial outlets, POI facilities, and other influential factors in the grid to obtain the layout potential of digital signage and then select grids with a high deployment potential as the place to be deployed. Compared with previous studies, by combining the two methods, the spatial characteristics and the commercial potential of the location are integrated into the content-based site-selection method. This not only improves computing efficiency but also improves the interpretability of the model. This process mainly includes three parts: data spatialization, site selection, and model verification (Figure 2).

First, the data are processed and spatialized at multiple scales by ESRI® ArcGIS™ 10.3. Then, the improved Huff model is used to calculate the spatial accessibility of digital signage, and machine learning methods such as the BP neural network are used to calculate the digital signage layout potential within the Sixth Ring Road in Beijing. Finally, the MCI model, BP neural network model, and Huff model are used to contrast with our model. The validity and accuracy of the digital signage site-selection model proposed in this paper are verified by cross validation and ROC curves.

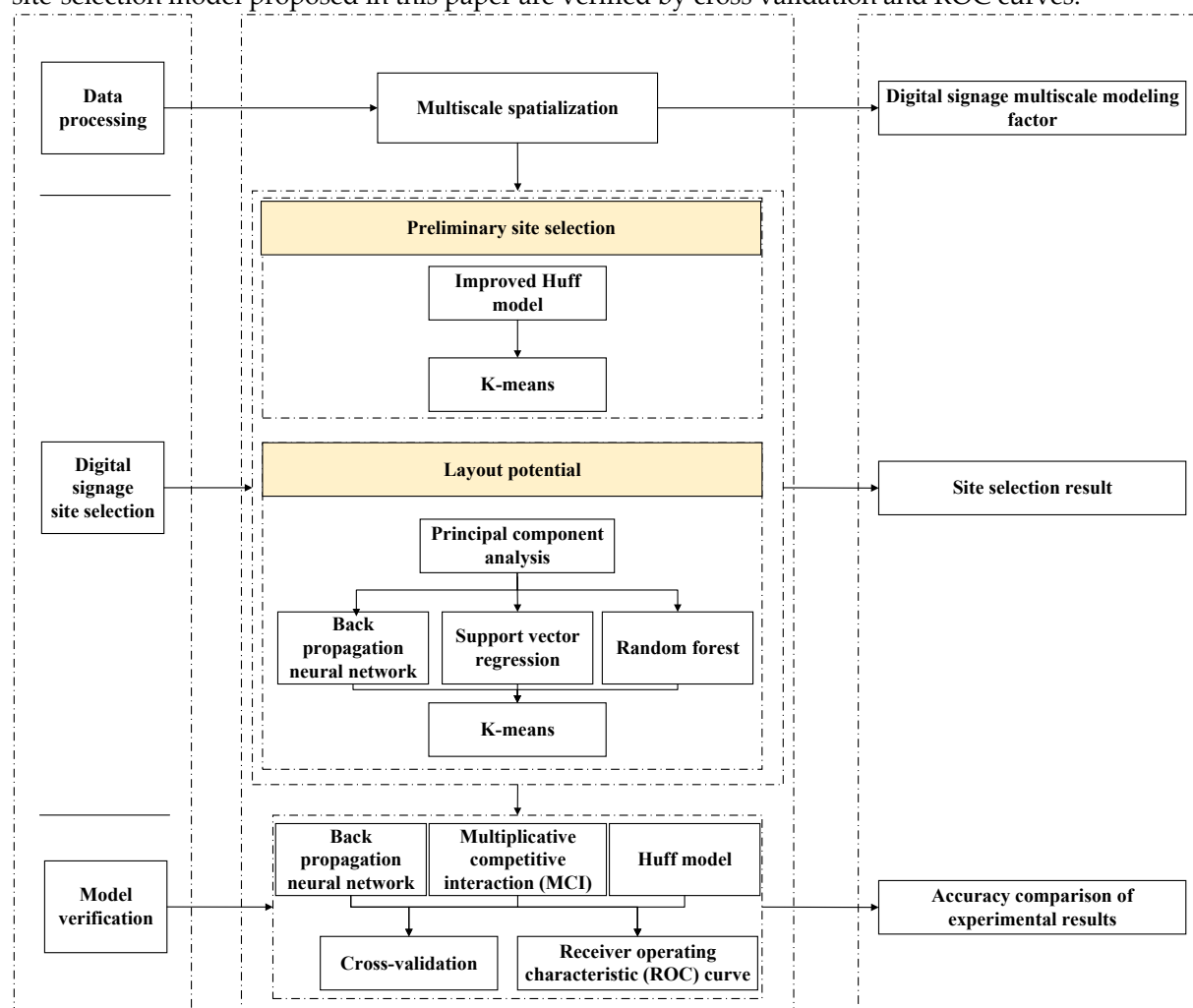


Figure 2. Technical flowchart of the location algorithm.

3.1. Data Processing

Due to the multisource heterogeneity of digital signage influential factors, to obtain the spatial characteristics of digital signage and influential factors, it is necessary to build a standard grid and suitable scales to obtain digital signage modelling factors. Grid-based digital signage modelling factors can not only reflect spatial feature information more realistically and intuitively but can also provide a unified spatial reference for data fusion and subsequent analysis.

In the process of spatialization, it is considered that different experimental results will occur due to different spatial scales (i.e., there is a modifiable areal unit problem [43–49] (MAUP)). According to Wang et al. [38], dividing the study area into a 100 m to 1,000 m grid scale will meet the needs of site selection experiments. Therefore, to accurately evaluate the potential of digital signage within the Sixth Ring Road in Beijing and to find a more suitable spatial scale, this paper selects 100 m to 1,000 m for spatial scale division experiments.

The specific experimental process is shown in Figure 3. First, data processing is performed on digital signage influential factor data, and then multiscale areal interpolation is performed, that is, 100 m to 1,000 m standard grids are used for the features using ESRI® ArcGIS™ 10.3. The number of standard grids at 10 different scales is shown in Table 2, so that 10 scale location modelling factors are obtained.

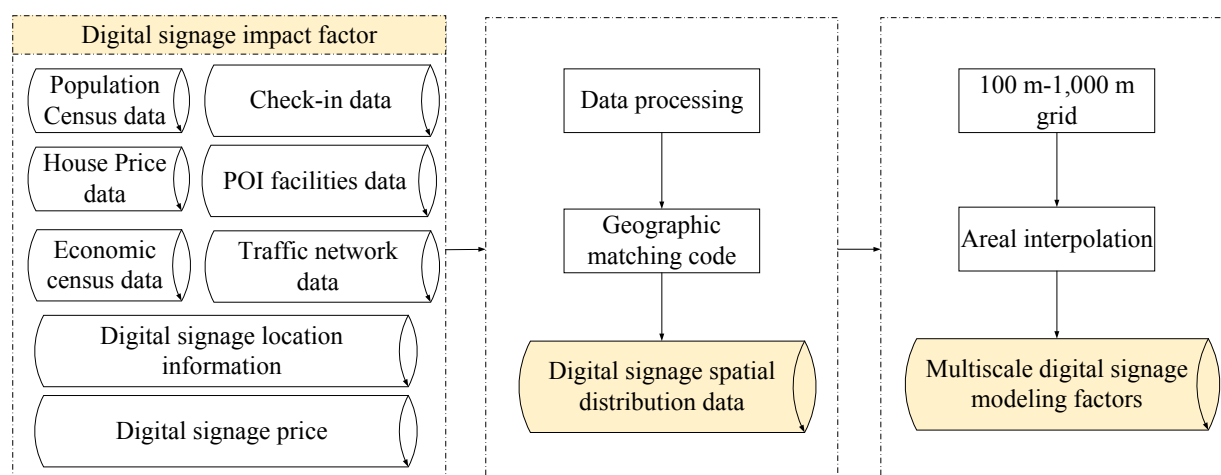


Figure 3. Processing flow of digital signage modelling factors.

Table 2. Number of standard grids at 10 different scales.

Scale (m)	Number of Grids
100×100	227,330
200×200	56,837
300×300	25,262
400×400	14,200
500×500	9,095
600×600	6,367
700×700	4,640
800×800	3,561
900×900	2,812
1,000×1,000	2,275

3.2. Location Model

The site-selection process is mainly divided into two steps. The first step is to use the modified Huff model to calculate the spatial accessibility of digital signage and use the K-means clustering algorithm to classify the calculation results into 3 levels, high, medium and low, to obtain the spatial accessibility of digital signage at ten scales (from 100 m to 1,000 m). The second step is to calculate the layout potential of the digital signage in the area by the BP neural network, random forest, and

support vector machine regression algorithms and use the clustering algorithm to classify the calculation results into high, medium, and low. Finally, overlay analysis is used to obtain the site-selection results.

3.2.1. Preliminary Site Selection

The preliminary site selection refers to the calculation of the spatial accessibility of digital signage in the area, and the preliminary selection results are based on the broadcast price, number of check-ins (data type indicating the audience), and the distance between the digital signage and the audience. The Huff model, as an empirical commercial site-selection model, takes the distance between the retail store and the customer and the area of the retail store itself into account. Therefore, a modified Huff model centred on the Weibo check-in data is used to calculate the spatial accessibility of digital signage. Then, the K-means clustering algorithm is used to classify the spatial accessibility calculation results, and clusters of different levels of digital signage spatial accessibility are obtained.

1. Huff model

The Huff model is an empirical business location model [50]. The model uses distance and the area of the sales area to calculate the attractiveness of a retail store to consumers, that is, the probability that a customer chooses a store. Variable P_{ij} represents the probability that a consumer at location i will spend at store j ; the formula can be written as

$$P_{ij} = \frac{S_j / T_{ij}^\beta}{\sum_{j=1}^n S_j / T_{ij}^\beta} \quad (1)$$

where T_{ij} represents the time to reach the store (this refers to the calculation of the distance to the store); S_j represents the area of the sales area of the store; and β is a parameter estimated from experience that indicates the impact of the time required for consumers to form different behaviours, generally set as 2.

2. Modified Huff model

Based on the calculation framework of the Huff model, combined with the experimental needs of digital signage site selection, the model was improved to concentrate on the Weibo check-in point in the standard grid to calculate the attractiveness of digital signage to the audiences in the unit area (Figure 4) and to characterize the spatial accessibility of digital signage, as shown in formula (2):

$$a_i = \sum_{j=1}^m \frac{num_i * p_j}{d_{ij}^2} \quad (2)$$

$$A_G = \frac{\sum_{i=1}^n a_i}{n}$$

where i denotes Weibo check-in point data; j denotes digital signage point data; num_i denotes the number of check-in times of the check-in point; p_j denotes the broadcast price of the digital signage; d_{ij} denotes the distance between the check-in point and the digital signage; and a_i represents the sum of one of the check-in points (indicates the audience) in a grid and all the digital signage data points ($num_i * p_j / d_{ij}^2$), indicating the ability of the check-in point to reach the digital signage. A_G characterizes the average ability of a grid to reach a digital signage, and n is the number of check-in points in a grid.

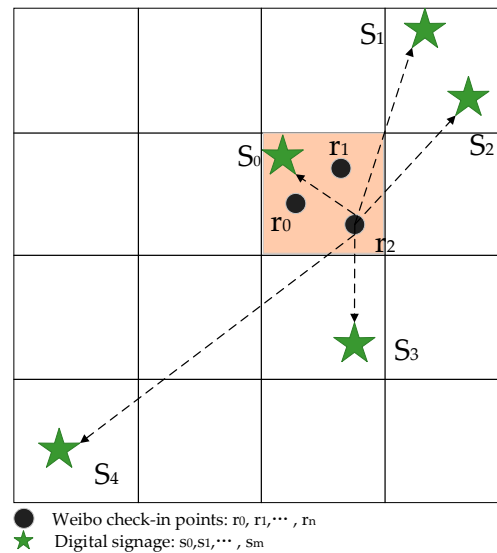


Figure 4. Principle of the spatial accessibility model. The yellow region denotes the target region, and the blank regions are the eight nearby regions.

3. K-means clustering

The K-means [51] model is a clustering method based on distance division. The clustering process is as follows: First, it is necessary to determine K centroids, that is, the number of clusters that are desired to be aggregated; second, the Euclidean distance of each index is calculated, and the distances are integrated between the calculated objects; finally, the program is run continuously, and each group randomly selects the centroid until the sum of the squared errors is the smallest [52].

It is necessary to distinguish regions with different spatial accessibility and layout potentials. Since the data are numerical data, different levels of regions need to be divided by Euclidean distance. Therefore, the K-means algorithm is needed to obtain clusters of different levels.

4. Calinski–Harabasz index (CH index)

The CH index is a statistical measurement index to detect the distribution effect. The specific formula is shown in formula (3), where $Tr(B_k)$ and $Tr(W_k)$ are the distances between categories and within categories, respectively. The CH index indicates that the difference between measurement categories is greater than the difference within categories; that is, when the clustering result is optimal, CH has the maximum value [53].

$$CH(k) = \frac{Tr(B_k)}{Tr(W_k)} \frac{n-k}{k-1} \quad (3)$$

where B_k is the intraclass divergence matrix, W_k is the interclass divergence matrix, and the calculation of W_k and B_k is as follows:

$$B_k = \sum_q n_q (c_q - c)(c_q - c)^T \quad (4)$$

$$W_k = \sum_{q=1}^k \sum_{x \in c_q} (x - c_q)(x - c_q)^T \quad (5)$$

where n is the number of points in the data, c_q is the set of points in cluster q , c is the centre of the sample point, and n_q is the number of points in class q .

The *CH* index is used to evaluate the clustering effect of different parameters in the above algorithm. The higher the index value, the better the clustering effect, and then the parameters in the algorithm are determined.

3.2.2. Calculation of Digital Signage Layout Potential

The layout potential of digital signage means selecting multiscale modelling factors as independent variables and digital signage advertising prices as dependent variables to predict the potential within a unit area. Typical prediction algorithms include the BP neural network, support vector machine regression, and random forest. Due to the difference between the dataset and the data characteristics, there is no universal prediction method applicable to all datasets; therefore, we conduct comparative experiments, taking the root mean square error (*RMSE*) as the evaluation index and selecting a suitable algorithm and the appropriate scale to predict the layout potential of digital signage in the region.

1. BP neural network

The BP neural networks [54] consist of input, hidden, and output layers. The training process is as follows: first, the synaptic weight and threshold matrix of the network are initialized, and the training samples are presented; second, the forward propagation and the error back propagation are calculated, then the weight is updated; and last, iteration is performed, using the new samples to perform forward propagation calculations and error back propagation calculations until the stopping criterion is met.

2. Support vector machine regression (SVR)

SVR regression, in simple terms, finds a regression plane that makes the distance of all the data of a set to the plane the closest [55]. Different from the empirical regression model, support vector regression assumes that as long as $f(x)$ and y do not deviate too much, the prediction can be considered correct, and the loss is not calculated. Specifically, the threshold α is set, and only $|f(x)-y|>\alpha$ is calculated as the loss value of the data point. The commonly used kernel functions are RBF (radial basis function), linear, and poly.

3. Random forest (RF)

Random forest is a structured supervised learning method. The calculation process using the regression prediction is as follows: first, extract N sample units from the original data randomly and generate a regression tree; second, randomly extract m at each node; finally, integrate the result of each regression tree and generate predicted values [56].

In the above prediction models, the grids without digital signage are used as the training set of the model, and the grids with digital signage are used as the test set of the model.

4. Root mean square error (RMSE)

The root mean square error is used to measure the deviation between the predicted value and the true value. The smaller the *RMSE* value, the smaller the error of the algorithm. The formula is

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x - \bar{x})^2}{n}} \quad (6)$$

To exclude the possible correlation between multidimensional features, the principal component analysis method was introduced to reduce the dimensions of multidimensional feature data, and the results of dimensionality reduction were used as model inputs.

5. Principal component analysis (PCA)

PCA is a data-compression algorithm [57]. When dealing with multidimensional data, PCA is used to filter out data attributes with higher similarity, thereby achieving the purpose of dimensionality reduction and accelerating the speed of data processing. The specific calculation process is removing the average value, calculating the covariance matrix, calculating the eigenvalues and eigenvectors of the covariance matrix, sorting the eigenvalues, retaining the eigenvectors

corresponding to the first N largest eigenvalues, and transforming the data into a new space constructed by the N feature vectors obtained above.

3.3. Model Verification

In the model verification, the digital signage site-selection method combined with the modified Huff model and BP neural network model is compared with the empirical Huff model, BP neural network model, and MCI location model. The receiver operating characteristic curve (ROC curve) is used as the evaluation standard to verify the effectiveness and accuracy of the digital signage site-selection method proposed in this paper.

The ROC curve can reflect the sensitivity and accuracy of the model when selecting different thresholds. When the distribution of positive and negative samples changes, the shape of curve can be kept basically unchanged, so this evaluation index can reduce the interference caused by different test sets and more objectively measure the performance of the model itself [58].

The horizontal axis of the curve is the false positive rate (FPR). The false positive rate indicates how many negative samples in the sample are predicted to be positive samples. There are two possibilities for this case. One is to change the original negative value—the class is predicted as a positive class (false positive, FP)—and the other is to predict the original negative class as a negative class (true negative, TN), that is,

$$FPR = \frac{FP}{TN + FP} \quad (7)$$

The vertical axis of the curve is the true positive rate (TPR). The true positive rate represents the proportion of samples that are predicted to be positive relative to the total number of positive samples. There are two possibilities for predicting positive results. One is the positive class (true positive, TP), and the other is to predict the negative class as the positive class (false positive, FP), that is,

$$TPR = \frac{TP}{TP + FN} \quad (8)$$

In this paper, the false positive rate (FPR) indicates the proportion of samples without digital signage selected as samples to be deployed, and the true positive rate (TPR) means the proportion of samples with digital signage selected as samples to be deployed. The closer the ROC curve is to the upper left, the higher the true class rate, the lower the negative class rate, and the better the model effect.

In the Huff model, multiscale digital signage modelling factors are used as input data to obtain the attractiveness of the computing centre to the surrounding grids, which characterizes the possibility of digital signage in this grid. After calculating the accuracy of the model through the ROC curve, the probability is divided into multiple thresholds of 0–1 with a step size of 0.1, and finally, the accuracy of the Huff model is obtained.

In the BP neural network, multiscale digital signage modelling factors are used as input data, and the normalized digital signage broadcast price is used as output to predict the digital signage layout potential in the sixth ring of Beijing. The grid where the potential is high will be selected as the unit to be deployed. After calculating the accuracy of the model through the ROC curve, the layout potential is divided into multiple thresholds of 0–1 with a step size of 0.1, and finally, the accuracy of the BP neural network model is obtained.

In the MCI site-selection model, multiscale digital signage modelling factors are used as input data, the grid digital signage layout probability is output, and the probability value is set as the threshold of the ROC curve with a step size of 0.1. Finally, a threshold with a high accuracy rate is selected. Grids with a probability of layout greater than the selected threshold are used as units to be deployed.

1. Cross validation

The cross-validation method is used to improve the accuracy of the three regression prediction models, such as the BP neural network, through multiple sampling and training processes. In the

model, the training and validation datasets are randomly selected, so the quality of the training and validation data may be uneven. To reduce the accuracy loss during data segmentation, this experiment uses ten-fold cross validation and randomly divides the dataset into four. In this process, data division was repeated 10 times. Therefore, the entire training-verification process was performed 40 times. The final error is the average error of the 40 iterations.

4. Results and Discussion

4.1. Preliminary Site-Selection Results

The method described in Section 3.2.1 is used to calculate the spatial accessibility of digital signage in 100–1,000 m standard grid units. Moreover, the K-means clustering algorithm is used to cluster the spatial accessibility results and adjust the parameters (K value) of the algorithm for multiple experiments to obtain accurate clustering results. The CH index is a statistical measurement index to detect clustering effects, usually used to evaluate the effect of the clustering algorithm. Therefore, we use the CH index described in Section 3.2.1 as an evaluation index. The clustering quality of the algorithm with different parameters is shown in Figure 5. As K increases, the CH index continues to fall; when the K value is 3, the CH index is relatively high. Therefore, the K-means algorithm is used to divide accessibility into three categories: low, medium, and high (Figure 5).

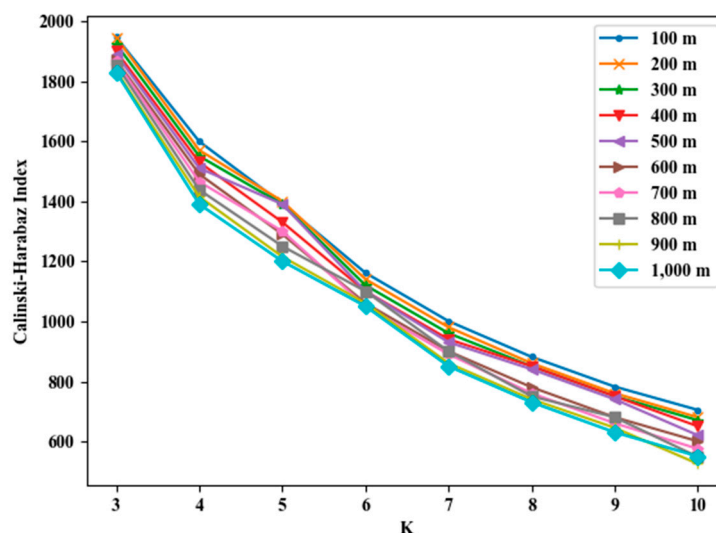


Figure 5. Calinski–Harabasz index of each K .

The experiments are performed on 10 scales from 100 m to 1,000 m, so the spatial accessibility results are also obtained at each scale. At the same time, we use ESRI® ArcGIS™ 10.3 to conduct a spatial query to obtain the number of Weibo check-ins in the standard grid within the Sixth Ring Road of Beijing. The number of check-ins indicates the visibility of digital signage. The highly visible grid unit is equipped with digital signage, which is more likely to be seen by the audience.

Taking a grid unit of 100 m as an example, Figure 6 (a) shows the result of spatial accessibility. The yellow grid in the figure indicates that the accessibility is low, and digital signage is urgently needed. The green grid indicates that the accessibility is medium, and digital signage can still be deployed; the blue grid indicates that the accessibility is high, and there is no need to deploy digital signage. Figure 6 (b) shows the distribution of the number of Weibo check-ins in the grid. The yellow grid in Figure 6 (b) indicates that the area has a low number of check-ins, which indicates that the area has a relatively small number of people. If digital signage is placed here, the visibility is relatively low, and the blue grid indicates that the number of check-ins is high, meaning a wider audience here. If digital signage is installed here, the visibility is high, and it has a very high deployment value.

Overlaying the two layers of Figure 6 (a) and 6 (b), the grid units with medium and low spatial accessibility and a high number of Weibo check-ins are the preliminary results of the positions where the digital signage will be deployed (Figure 6c).

The spatial accessibility of digital signage is mainly affected by the distance between the signage and the audience, the number of check-ins, and the price and quantity of the digital signage. Therefore, the primary results were mainly distributed near Weigong Village, Wangfujing, Xidan, and the main transportation lines. In general, various commercial districts and tourist attractions have greater appeal to an audience. Because of the higher number of check-ins, more visibility and relatively lower spatial accessibility, digital signage is required. As shown in Figure 6 (c), the areas around Weigongcun, Sihui Building Materials Market, Wangfujing, Xidan, and Huilongguan Wholesale Market basically cover the culture, franchise, shopping centre, and comprehensive business district. The attractiveness and influential scope of these areas are extensive and suitable for layout digital signage. With the government and universities relocating, audiences in some areas, such as Fangshan, Daxing, and Tongzhou, have gradually increased, and digital signage here will also have greater benefits.

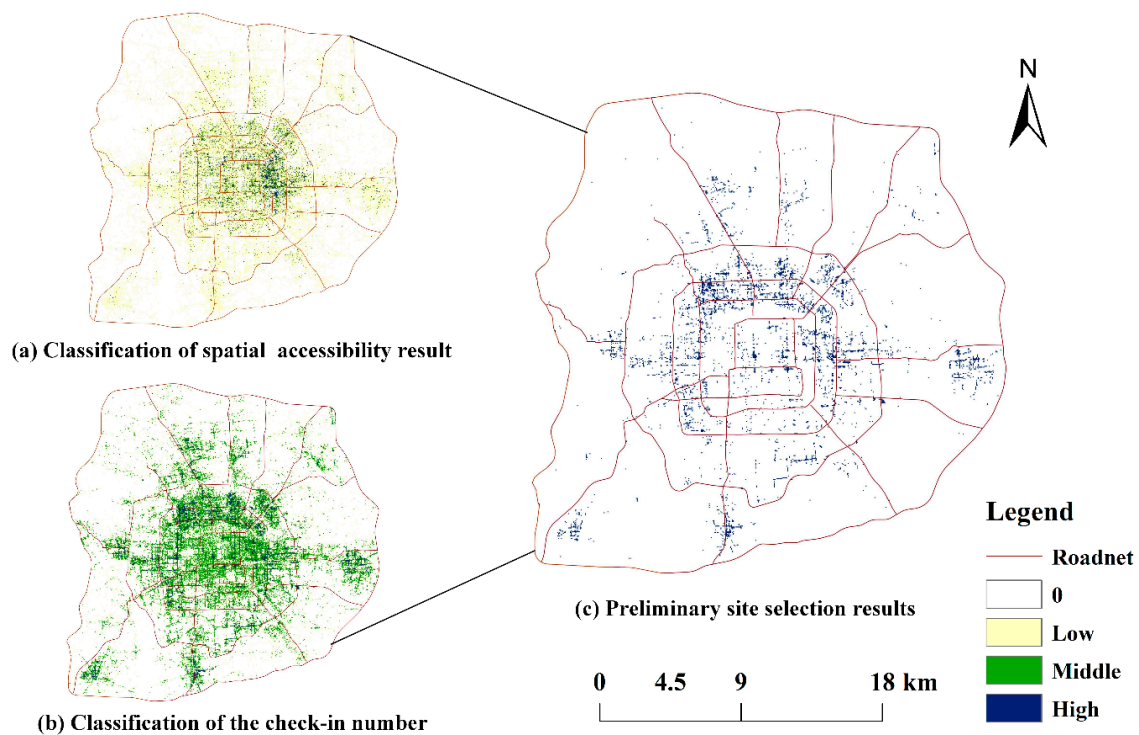


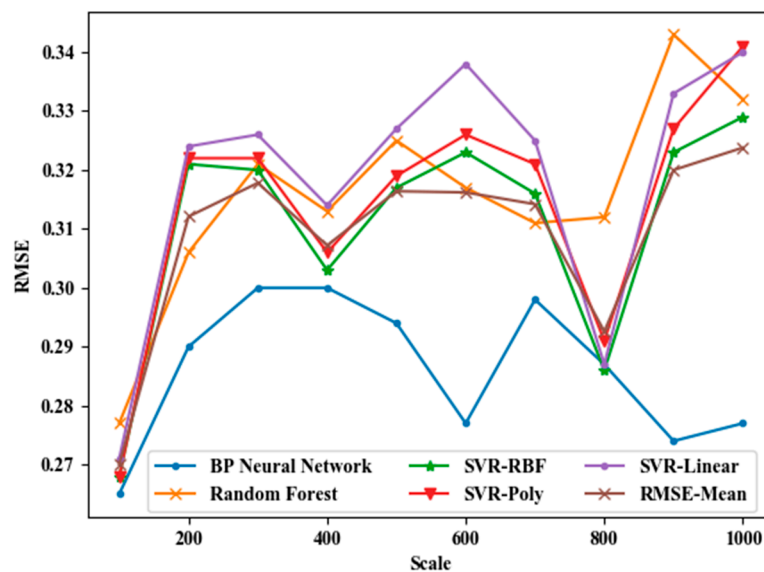
Figure 6. Preliminary site-selection results ((a) Classification of spatial accessibility result; (b) Classification of the check-in number; (c) Preliminary site selection results).

4.2. Layout Potential Results

The layout potential of digital signage is calculated with the method described in Section 3.2.2. The RMSE described in Section 3.2.2 is generally used as an evaluation index for prediction algorithms. This paper uses it to evaluate the prediction effects of five regression prediction algorithms. The results are shown in Table 3 and Figure 7. With the increase of the grid scale, the RMSE values of five algorithms generally increased gradually. When the scales are 100 m, the errors are relatively small, and the RMSE values of the five algorithms are 0.277, 0.265, 0.268, 0.271, and 0.268. The relatively small RMSE value is 0.265, and the algorithm with a relatively small error is the BP neural network (Figure 7). Therefore, the digital signage modelling factors at 100 m grid scale and the BP neural network algorithm are selected to calculate the digital signage layout potential within the Sixth Ring Road in Beijing.

Table 3. RMSE value of the comparison of multiscale modelling factor regression prediction.

Scale	Random Forest	BP Neural Network	RBF	Linear	Poly	Mean
100 m	0.277	0.265	0.268	0.271	0.268	0.270
200 m	0.306	0.290	0.321	0.324	0.322	0.312
300 m	0.321	0.300	0.320	0.326	0.322	0.318
400 m	0.313	0.300	0.303	0.314	0.306	0.307
500 m	0.325	0.294	0.317	0.327	0.319	0.316
600 m	0.317	0.277	0.323	0.338	0.326	0.316
700 m	0.311	0.298	0.316	0.325	0.321	0.314
800 m	0.312	0.287	0.286	0.287	0.291	0.293
900 m	0.343	0.274	0.323	0.333	0.327	0.320
1,000 m	0.332	0.277	0.329	0.340	0.341	0.324

**Figure 7.** Comparison of multiscale modelling factor regression prediction results.

To exclude the correlation between 19 types of digital signage modelling factors from the experiment, PCA is introduced to perform analysis of the above features to achieve the purpose of feature dimension reduction, thereby reducing the *RMSE* value and improving the accuracy of the regression analysis of digital signage potential. It can be seen from the results after using PCA (Table 4, Figure 8) that the minimum value of *RMSE* at this time is 0.264, which is relatively smaller than that before dimension reduction. After extracting the principal components of the modelling factor sequence, the calculation error of the digital signage potential is relatively reduced.

Table 4. Regression analysis results after feature dimensionality reduction

Method	RMSE
BP Neural Network	0.264
Random Forest (RF)	0.271
SVR (kernel: RBF)	0.271
SVR (kernel: Linear)	0.271
SVR (kernel: Poly)	0.270

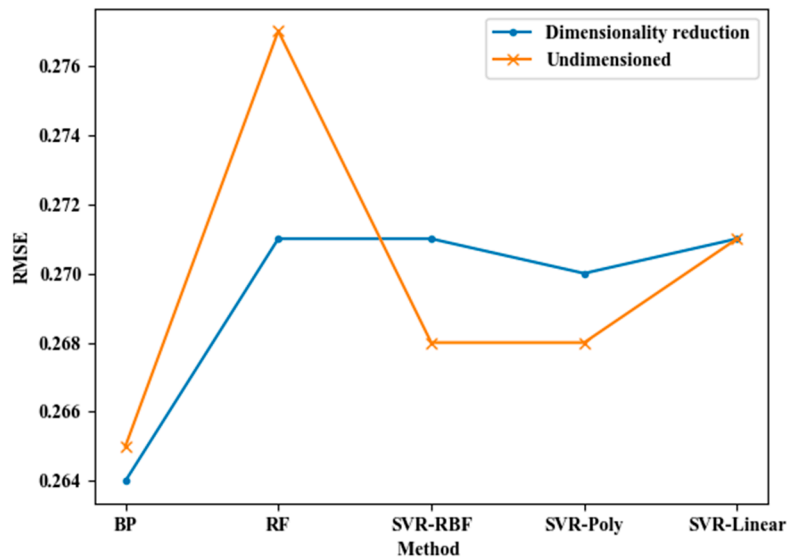


Figure 8. Comparison of regression analysis results before and after feature dimension reduction.

We take the digital signage modelling factor at a 100 m grid scale after dimensionality reduction as input and the normalized digital signage broadcast price as the output. The model is trained for the digital signage layout potential. The K-means clustering algorithm divides the result into three levels, low, medium, and high, as shown in Figure 9. Among the grids, the blue area is the area with high potential for digital signage deployment, mainly distributed in the CBD, Olympic Park, and other commercial areas and tourism areas. The result is consistent with the digital signage that should be deployed in a large number of check-in locations, likely to be seen by the audience with a higher probability. The green area shown in the figure is a medium-level area of layout potential, mainly distributed in Fangshan Changyang, Tongzhou Baliqiao, Daxing Zaoyuan, and other places. These developing business districts are characterized by large populations and fewer commercial facilities, so more digital signage needs to be deployed. The yellow areas shown in the figure are the areas with a low potential for deployment. There are basically no large-scale shopping malls or business districts, and the areas are less attractive to nearby residents. The possibility of digital signage being noticed by the audience is low, which leads to the low layout potential of digital signage in this place, and deployment is not recommended.

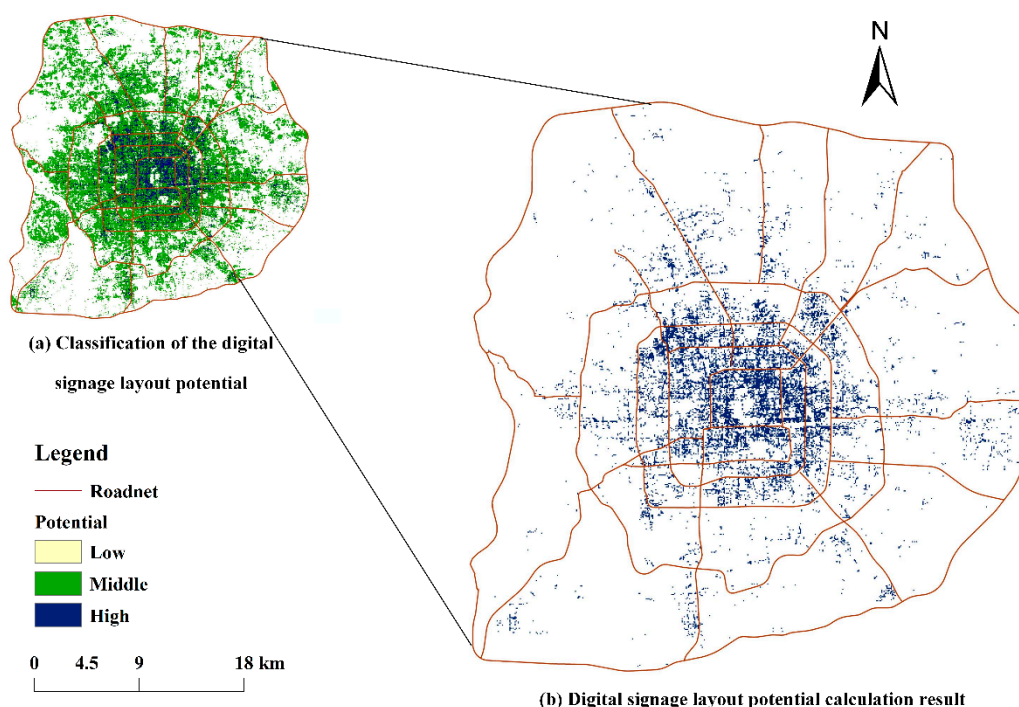


Figure 9. Digital signage layout potential calculation result ((a) Classification of the digital signage layout potential; (b) Digital signage layout potential calculation result).

4.3. Results and Analysis of Digital Signage Location

To verify the accuracy of the abovementioned digital signage site-selection method, we compare our method with the BP neural network model in the machine learning method and the Huff and MCI empirical model of commercial geography. The ROC curve (described in Section 3.3) characterizes the accuracy of the site-selection results. A lower false positive rate and higher true positive rate, that is, the closer the ROC curve is to the upper left, indicate the model has higher accuracy and better site selection effect.

The location results of each model are shown in Figure 10. The result of the BP neural network model is shown in Figure 10a. Using 100–1,000 m modelling factors as input of the BP neural network model, as the modelling scale increases, the ROC curve gradually approaches the upper left of the coordinates; at 700 m, the experimental effect is the best and gradually decreases after 800 m. Therefore, using the BP neural network model to predict the location of digital signage, the prediction effect is the best at the 700 m scale.

The Huff model is one of the main methods of commercial site selection, and its variables are the store area and the distance of the store from the neighbourhood. The Huff model was modified to adapt to the digital signage site-selection experiment. When the modelling scale is 100 m, the model effect is better. As the modelling scale increases, the model effect gradually decreases and reaches a certain convergence (Figure 10b).

The MCI model characterizes the probability of the audience choosing this region for consumption under the influence of the population, economy, and other geographical environments [29]. The mentioned multiscale factors are brought into the model, and the locations with higher probability are recommended as the new locations of digital signage. At each modelling scale, as the threshold increases, the model effect gradually converges. When the modelling scale is 100 m, the model location results are better. As the modelling scale increases, the MCI model location effect gradually decreases (Figure 10 (c)).

The ROC curve of the Huff–BP model proposed in this paper is the blue curve shown in Figure 10 (d). Based on the selection of the ROC curves of the above site selection models (BP neural network, Huff model, and MCI model) with better scales, it can be seen that the true positive rate of Huff–BP

is relatively high and its false positive rate is relatively low—the curve is closest to the top left, which shows the location effect of this experiment is relatively good, followed by the BP neural network model. Both the Huff model and the MCI model are probabilistic models, and the location effects of the two are similar, thereby verifying the accuracy and reliability of this experiment. In summary, the current methods and parameters and the results of the site selection are relatively good, and the results may be more satisfactory if we change the methodology.

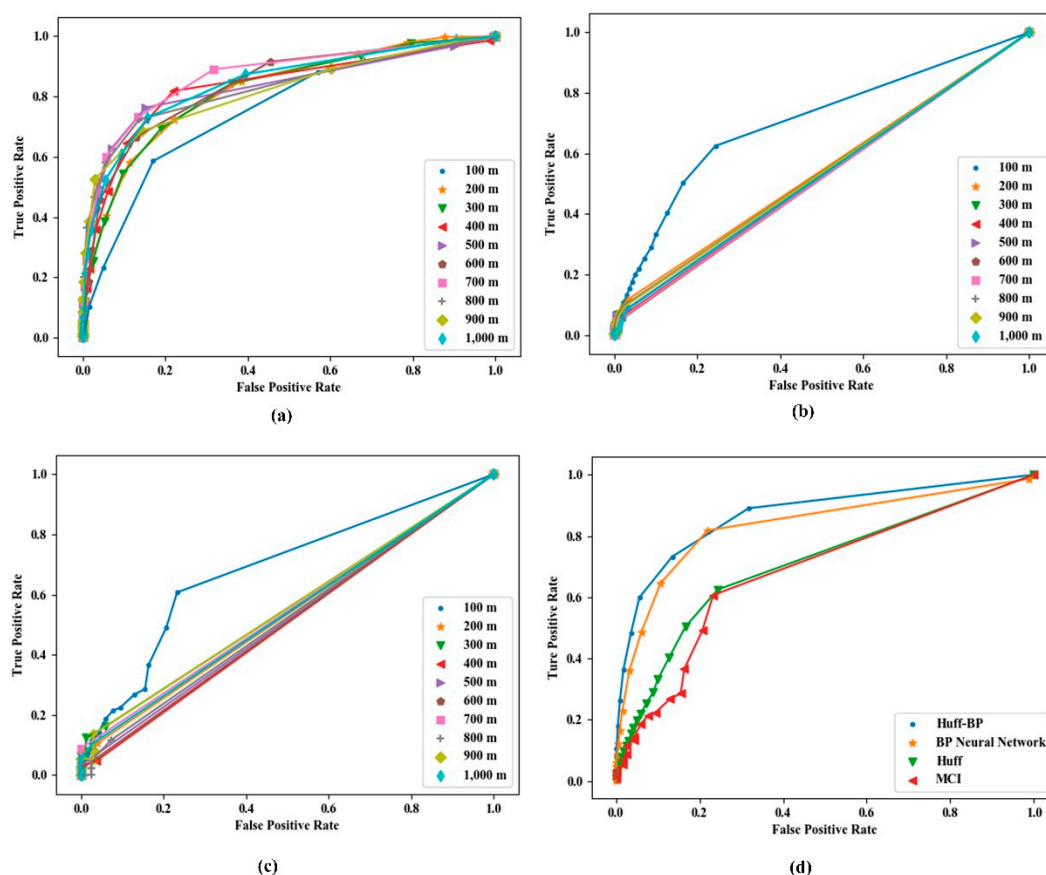


Figure 10. ROC curve of four algorithms under different scales ((a) BP Neural Network; (b) Huff model; (c) MCI model; and (d) ROC curve of Huff–BP model and the other three models with a better scale).

The abovementioned preliminary signage site-selection results and digital signage layout potential results are overlaid to analyse areas that simultaneously satisfy low spatial accessibility, high check-in numbers, and high layout potential (areas for digital signage to be selected), as shown in Figure 11.

As shown in Figure 11, the sites are mainly located in Sanlitun, 798 Art District, Beijing West Railway Station, and other places. These locations are divided into three categories: (1) Olympic Park, Sanlitun, 798 Art District, and other places in the figure that are convenient for transportation, have complete catering and entertainment facilities, and are part of the well-known cultural and entertainment industry zone in Beijing. This area has a large number of check-ins and high potential for deployment. Deploying more digital signage is recommended. (2) Fangshan Liangxiang, Daxing Huangcun, Tongzhou Beiyuan, etc., have gradually increased their commercial development due to the construction of Beijing’s subcentres and the relocation of some universities. This area has high commercial value and layout potential, and digital signage is recommended. (3) Beijing Railway Station, Beijing West Railway Station, Beijing South Railway Station, Siyuanqiao, etc., are passenger logistics gathering places. These locations have high population mobility throughout the year, and there are many commercial outlets nearby, so digital signage is also recommended.

When we trained the model, we used the digital signage point data and digital signage modelling factors within the Six Ring Road of Beijing and adjusted the parameters based on the training results, so the model is more suitable for the digital signage site selection experiment within the Six Ring Road of Beijing, that is, the results are adaptable to this specific study area.

In summary, we selected some grids from a relatively macro scale that are suitable for digital signage layout within the Six Ring road of Beijing and provide a digital signage site selection method. Moreover, we provided a research idea that combines empirical site selection methods with machine learning methods to improve the interpretability, computing efficiency. Our study area is in Beijing, training model with data of Beijing, so it is likely more suitable for cosmopolitan cities with a large number of check-ins and high prosperity. For other commercial facility location issues, the site selection method proposed in this paper can also be used as a reference.

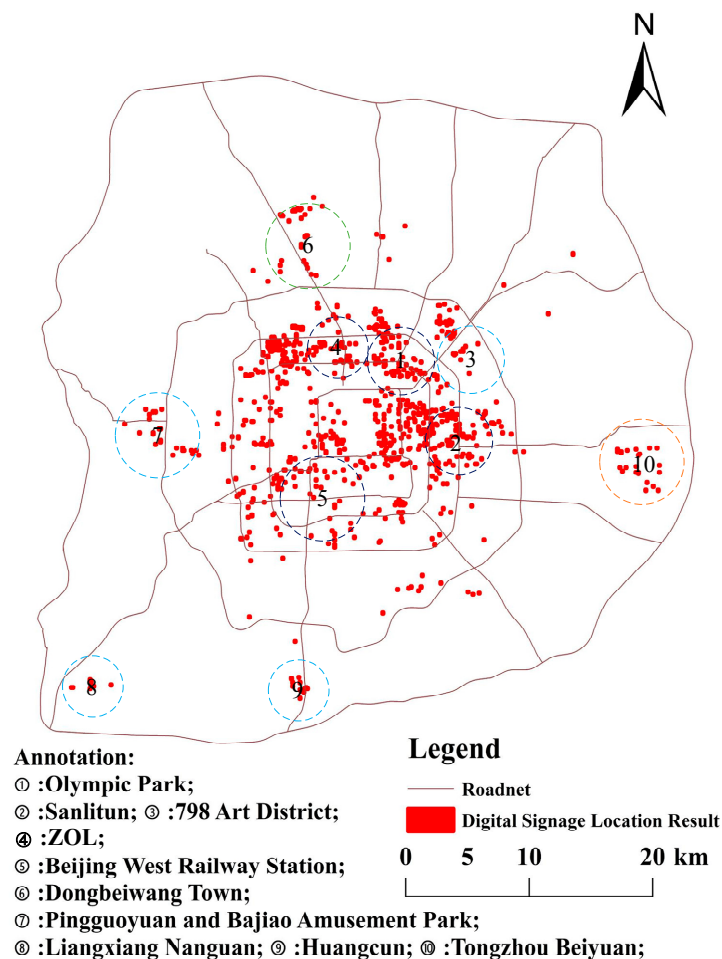


Figure 11. Digital signage location results.

5. Conclusions

With the wide use of digital signage, the precise location of digital signage is an urgent issue for digital signage companies and management agencies. We proposed a hybrid Huff and BP neural network model that comprehensively considers the population, economy, transportation, and other factors that affect the location of digital signage and performs standard grid processing to form multiscale modelling factors. We modified the Huff model to calculate the spatial accessibility of digital signage and comprehensively calculate the layout potential of digital signage using multiple machine learning methods such as BP neural networks. The Huff–BP, MCI, Huff, and BP neural network models were compared. The conclusions are as follows:

(1) A set of multifactor and multiscale digital signage modelling factors was constructed. Based on the influential factors of digital signage, such as broadcasting price, POI facilities, census,

transportation network centrality index, and social network check-ins, we unified the spatial scale with multiscale grid processing of 100 m to 1,000 m.

(2) A digital signage site-selection method that combines a machine learning algorithm with a modified Huff model was proposed. Based on the location and price of digital signage, the Huff model was improved to calculate the spatial accessibility of digital signage, and preliminary selection was made for the unit. By using different prediction algorithms to calculate the digital signage deployment potential, the results show that the 100 m scale and BP neural network algorithm had better performance than RF and SVR algorithm under the other scales. The results of site selection were obtained by overlaying the preliminary selection results and the layout potential results.

(3) The site-selection results indicated that the areas suitable for digital signage are mainly distributed in Sanlitun, Wangfujing, Financial Street, Beijing West Railway Station, and areas along the main road network. These areas are mostly well-known cultural and entertainment industrial districts and passenger logistics gathering places in Beijing. They have relatively better commercial outlets, a higher population, and convenient transportation, which can maximize the benefits of digital signage.

Our research provides a reference for integrating geographical features and content data into the site-selection algorithm for digital signage. Furthermore, we provide some grids from a relatively macro scale that are suitable for digital signage layout within the Six Ring road of Beijing and provide a digital signage site selection method. The method can effectively improve the accuracy and scientific nature of digital signage deployment, maximize the deployment benefits, and optimize the allocation of digital signage resources. In addition, this paper provides a research idea that combines empirical methods with machine learning methods to improve the model's interpretability and computing efficiency. For other commercial facility location issues, the location method proposed in this paper can also be used as a reference. However, one of the most important factors for the site location of digital signage is visibility from streets. Therefore, our future research will aim at the location of digital signage along the street network. Based on the micro-location theory, introducing factors, such as the position and orientation of digital signage into more precise location experiments, and incorporating constraints into the model, such as user needs and deployment costs, will achieve an optimal layout of digital signage.

Author Contributions: Conceptualization, Xun Zhang; Data curation, Rui Zhou; Funding acquisition, Xun Zhang and Dong Jiang; Investigation, Xun Zhang; Methodology, Su Li; Project administration, Xun Zhang; Software, Mengmeng Hao; Supervision, Dong Jiang; Validation, Yuxue Wang; Visualization, Yuxue Wang; Writing—original draft, Yuxue Wang; Writing—review & editing, Xun Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Strategic Priority Research Program of the Chinese Academy of Science (grant number XDA23000000); the China Postdoctoral Science Foundation (grant number 2017M620885); the Support Project of High-level Teachers in Beijing Municipal Universities in the Period of 13th Five-year Plan (grant number CIT&TCD201904037); the R&D Program of Beijing Municipal Education Commission (grant number KM202010011011); and the National Natural Science Foundation of China (grant number 61802010).

Acknowledgements: We would like to acknowledge the Beijing Key Laboratory of Big Data Technology for Food Safety and Key Laboratory of Resources Utilization and Environmental Remediation for providing a research grant to conduct this work. We express gratitude to the editors for the editing assistance. Last, we would like to thank the reviewers for their valuable comments and suggestions on our paper.

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

References

1. Hossain, M.A.; Islam, A.; Le, N.T.; Lee, Y.T.; Lee, H.W.; Jang, Y.M. Performance analysis of smart digital signage system based on software-defined IoT and invisible image sensor communication. *Int. J. Distrib. Sens. Netw.* **2016**, *12*, 1–14.
2. Ravník, R.; Solina, F. Audience Measurement of Digital Signage: Quantitative Study in Real-World Environment Using Computer Vision. *Interact. Comput.* **2013**, *25*, 218–228.
3. Want, R.; Schilit, B.N. Interactive Digital Signage. *Comput.* **2012**, *45*, 21–24.

4. Davies, N.; Langheinrich, M.; Jose, R.; Schmidt, A. Open Display Networks: A Communications Medium for the 21st Century. *Comput.* **2012**, *45*, 58–64.
5. Lin, L.; Zhigang, Z. Analysis of Digital Signage Advertising Operation. *Youth Journal*. **2017**, *21*, 96–97 (in Chinese).
6. Battiato, G.E.; Cavallaro, A.; Distant, C.; Battiato, S. Special issue on “Video analytics for audience measurement in retail and digital signage.” *Pattern Recognit. Lett.* **2016**, *81*, 1–2.
7. Ullah, F.; Sarwar, G.; Lee, H., et al. Control framework and services scenarios of provisioning N-Screen services in interactive digital signage. *Teh. Vjesn.* **2014**, *21*, 1321–1328.
8. Garaus, M.; Wagner, U. Let me entertain you – Increasing overall store satisfaction through digital signage in retail waiting areas. *J. Retail. Consum. Serv.* **2019**, *47*, 331–338.
9. Garaus, M.; Wagner, U.; Manzinger, S. Happy grocery shopper: The creation of positive emotions through affective digital signage content. *Technol. Forecast. Soc. Chang.* **2017**, *124*, 295–305.
10. Alfian, G.; Ijaz, M.F.; Syafrudin, M.; Syaekhoni, M.A.; Fitriyani, N.L.; Rhee, J. Customer behavior analysis using real-time data processing. *Asia Pac. J. Mark. Logist.* **2019**, *31*, 265–290.
11. Kim, J.S. A Study of Contact Frequency and Consumer Preference for Digital Signage Advertisement. In Proceedings of the e-Business and Telecommunication Networks; Springer Science and Business Media LLC, 2012; Vol. 338, pp. 181–187.
12. Yoon, S.; Kim, H. Research into the Personalized Digital Signage Display Contents Information through a Short Distance Indoor Positioning. *Int. J. Smart Home*. **2015**, *9*, 171–178.
13. Ijaz, M.F.; Tao, W.; Rhee, J.; Kang, Y.-S.; Alfian, G. Efficient Digital Signage-Based Online Store Layout: An Experimental Study. *Sustain.* **2016**, *8*, 511.
14. Umor, E.F. THE ROLE OF DIGITAL SIGNAGE ADVERTISING IN ENHANCING PATRONAGE AMONG ADVERTISERS AND POTENTIAL CONSUMERS; THE UYO CITY OUTLETS IN PERSPECTIVE. *Commun. Rev.* **2017**, *1*, 174.
15. Khue, T.D.; Binh, N.T.; Chang, W.; Kim, C.; Chung, S.-T. Design and implementation of MEAN stack-based scalable real-time Digital Signage System. In Proceedings of the 2017 8th International Conference of Information and Communication Technology for Embedded Systems (IC-ICTES); Institute of Electrical and Electronics Engineers (IEEE), 2017; pp. 1–6.
16. Sheppard, D.; Felker, N.; Schmalzel, J. Development of Voice Commands in Digital Signage for Improved Indoor Navigation Using Google Assistant SDK. *2019 Ieee Sens.* **2019**, 1–5.
17. Chen, Y. *Optimization Design and Implementation of Multimedia Information Publishing System*; Fudan University: Shanghai, China, 2013. (In Chinese)
18. Inoue, H.; Suzuki, K.; Sakata, K.; Maeda, K. Development of a Digital Signage System for Automatic Collection and Distribution of Its Content from the Existing Digital Contents and Its Field Trials. *2011 Ieee/Ipsj Int. Symp. Appl. Internet.* **2011**, 463–468.
19. Ma, T.; Guo, L.; Tang, M.; Tian, Y.; Al-Rodhaan, M.; Al-Dhelaan, A. A Collaborative Filtering Recommendation Algorithm Based on Hierarchical Structure and Time Awareness. *Ieice Trans. Inf. Syst.* **2016**, *E99*, 1512–1520.
20. Xie, X.; Zhang, X.; Fu, J.; Jiang, D.; Yu, C.; Jin, M. Location Recommendation of Digital Signage Based on Multi-Source Information Fusion. *Sustain.* **2018**, *10*, 2357.
21. O’Roarty, B.; Patterson, D.; McGreal, S.; Adair, A. A case-based reasoning approach to the selection of comparable evidence for retail rent determination. *Expert Syst. Appl.* **1997**, *12*, 417–428.
22. Suárez-Vega, R.; Gutiérrez-Acuña, J.L.; Rodríguez-Díaz, M. Locating a supermarket using a locally calibrated Huff model. *Int. J. Geogr. Inf. Sci.* **2014**, *29*, 217–233.
23. De Figueiredo, C.J.J.; Mota, C.M.D.M.; De, J. A Classification Model to Evaluate the Security Level in a City Based on GIS-MCDA. *Math. Probl. Eng.* **2016**, 2016, 1–10.
24. A Badri, M. Combining the analytic hierarchy process and goal programming for global facility location-allocation problem. *Int. J. Prod. Econ.* **1999**, *62*, 237–248.
25. Velasquez, M.; Hester, P.T. An analysis of multi-criteria decision making methods. *Int. J. Oper. Res.* **2013**, *10*, 56–66.
26. Allahi, S.; Mobin, M.; Vafadarnikjoo, A.; Salmon, C. An Integrated AHP-GIS-MCLP Method to Locate Bank Branches. In proceedings of Industrial and Systems Engineering Research Conference (ISERC), 30 May–2 June 2015, Nashville, Tenn., USA.

27. Szeremeta-Spak, M.D.; Colmenero, J.C. A two-stage decision support model for a retail distribution center location. *Rev. Fac. De Ing. Univ. De Antioq.* **2015**, *74*, 177–187.
28. Chang, C.-W.; Wu, C.-R.; Chen, H.-C. Using expert technology to select unstable slicing machine to control wafer slicing quality via fuzzy AHP. *Expert Syst. Appl.* **2008**, *34*, 2210–2220.
29. Baviera-Puig, A.; Buitrago-Vera, J.; Escriba-Perez, C. GEOMARKETING MODELS IN SUPERMARKET LOCATION STRATEGIES. *J. Bus. Econ. Manag.* **2016**, *17*, 1205–1221.
30. Xia, L.; Xiaoping, L.; Shaoying, L. *Intelligent GIS and spatial optimization*. Science Press: Beijing China, 2010.
31. Cortes, C.; Gonzalvo, X.; Kuznetsov, V.; Mohri, M.; Yang, S.; AdaNet: Adaptive Structural Learning of Artificial Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), 6–11 August 2017, Sydney, Australia.
32. Pradhan, B. A comparative study on the predictive ability of the decision tree, support vector machine and neuro-fuzzy models in landslide susceptibility mapping using GIS. *Comput. Geosci.* **2013**, *51*, 350–365.
33. Al-Sharif, A.A.; Pradhan, B. A novel approach for predicting the spatial patterns of urban expansion by combining the chi-squared automatic integration detection decision tree, Markov chain and cellular automata models in GIS. *Geocarto Int.* **2015**, *30*, 1–24.
34. Zhou, G.; Wang, L. Co-location decision tree for enhancing decision-making of pavement maintenance and rehabilitation. *Transp. Res. Part C: Emerg. Technol.* **2012**, *21*, 287–305.
35. Yanga, Y.; Tang, J.; Luo, H.; Law, R. Hotel location evaluation: A combination of machine learning tools and web GIS. *Int. J. Hosp. Manag.* **2015**, *47*, 14–24.
36. Lu, Y.; Zhu, S.; Zhang, L. A Machine Learning Approach to Trip Purpose Imputation in GPS-Based Travel Surveys. Available online: <http://onlinepubs.trb.org/onlinepubs/conferences/2012/4thITM/Papers-R/0117-000075.pdf> (accessed on 28 February 2019)
37. Liu, D.; Weng, D.; Li, Y.; Bao, J.; Zheng, Y.; Qu, H.; Wu, Y. SmartAdP: Visual Analytics of Large-scale Taxi Trajectories for Selecting Billboard Locations. *Ieee Trans. Vis. Comput. Graph.* **2017**, *23*, 1–10.
38. Luyao, W.; Hong, F.; Yankun, W. Site Selection of Retail Shops Based on Spatial Accessibility and Hybrid BP Neural Network. *Isprs Int. J. Geo-Inf.* **2018**, *7*, 202.
39. Gahegan, M. Fourth paradigm GIScience? Prospects for automated discovery and explanation from data. *Int. J. Geogr. Inf. Sci.* **2019**, *34*, 1–21.
40. Zhang, X.; Ma, G.; Jiang, L.; Zhang, X.; Liu, Y.; Wang, Y.; Zhao, C. Analysis of Spatial Characteristics of Digital Signage in Beijing with Multi-Source Data. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 207..
41. Okabe, A. Spatial Analysis Along Networks. In *Encyclopedia of GIS*; Springer Science and Business Media LLC, 2017; pp. 1938–1948.
42. Sevtsuk, A. Path and Place: A Study of Urban Geometry and Retail Activity in Cambridge and Somerville, MA. PhD Thesis, Massachusetts Institute of Technology, Cambridge USA, 11 August 2010.
43. Kwan, M.-P. The Uncertain Geographic Context Problem. *Ann. Assoc. Am. Geogr.* **2012**, *102*, 958–968.
44. Haynes, R.; Daras, K.; Reading, R.; Jones, A. Modifiable neighbourhood units, zone design and residents' perceptions. *Heal. Place.* **2007**, *13*, 812–825.
45. Houston, D. Implications of the modifiable areal unit problem for assessing built environment correlates of moderate and vigorous physical activity. *Appl. Geogr.* **2014**, *50*, 40–47.
46. Nakaya, T. An Information Statistical Approach to the Modifiable Areal Unit Problem in Incidence Rate Maps. *Environ. Plan. A: Econ. Space.* **2000**, *32*, 91–109.
47. Dark, S.J.; Bram, D. The modifiable areal unit problem (MAUP) in physical geography. *Prog. Phys. Geogr. Earth Environ.* **2007**, *31*, 471–479.
48. Viegas, J.M.; Martinez, L.; Silva, E. Effects of the Modifiable Areal Unit Problem on the Delineation of Traffic Analysis Zones. *Environ. Plan. B: Plan. Des.* **2009**, *36*, 625–643.
49. Clark, A.F.; Scott, D. Understanding the Impact of the Modifiable Areal Unit Problem on the Relationship between Active Travel and the Built Environment. *Urban Stud.* **2013**, *51*, 284–299.
50. Barr, P.S.; Stimpert, J.L.; Huff, A.S. Cognitive change, strategic action, and organizational renewal. *Strat. Manag. J.* **1992**, *13*, 15–36.
51. Hartigan, J.A.; Wong, M.A. Algorithm AS 136: A K-Means Clustering Algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1979**, *28*, 100.
52. Huang, J.Z.; Ng, M.K.; Rong, H.; Li, Z. Automated variable weighting in k-means type clustering. *Ieee Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 657–668.

53. De Amorim, R.C.; Hennig, C. Recovering the number of clusters in data sets with noise features using feature rescaling factors. *Inf. Sci.* **2015**, *324*, 126–145.
54. Sadeghi, B. A BP-neural network predictor model for plastic injection molding process. *J. Mater. Process. Technol.* **2000**, *103*, 411–416.
55. Balabin, R.; Lomakina, E.I. Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. *Anal.* **2011**, *136*, 1703.
56. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
57. D'Aspremont, A.; El Ghaoui, L.; Jordan, M.I.; Lanckriet, G.R.G. A Direct Formulation for Sparse PCA Using Semidefinite Programming. *Siam Rev.* **2007**, *49*, 434–448.
58. Pepe, M.S.; Janes, H.; Gu, J.W. Letter by Pepe et al regarding article, "Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation.* **2007**, *116*, e132.



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