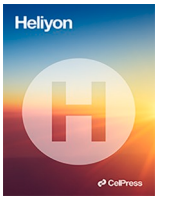
Heliyon 9 (2023) e21213

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Applying modified-data mining techniques to assess public transportation vulnerable urban and suburban city areas

Donghee Oh [a], Sangjae Lee [b], Juneyoung Park [a,b,*], Jaehong Park [c], Chang-Gyun Roh [c]

[a] Department of Smart City Engineering, Hanyang University, Ansan, 15588, Republic of Korea
[b] Department of Transportation and Logistics Engineering, Hanyang University, Ansan, 15588, Republic of Korea
[c] Department of Highway and Transportation Research, Korea Institute of Civil Engineering and Building Technology, Goyang, 10223, Republic of Korea

ARTICLE INFO

Keywords:
Data mining
Bus demand forecasting
Geographic information system
Demand responsive transit
Vulnerable areas

ABSTRACT

To guarantee the right to move for residents in areas where public transportation is insufficient, research is needed to identify vulnerable areas and prepare measures. This paper defines the vulnerable regions of public transportation within various city types in Korea. In order to identify appropriate areas to apply the Demand Responsive Transit (DRT), the regions with vulnerability were compared with a specific city (Yangsan-*si*) which already the DRT system was successfully adopted. To collect monthly bus data, web-data crawling method was performed and processed with coordinating program by matching GPS coordinate. The public transportation demand was predicted for each grid cell size (100 m, 250 m, and 500 m) by different methodologies. Various data mining models based on regression were analyzed to predict bus demand of vulnerable areas. Among models, a modified model was suggested to combine Automated machine learning models for high prediction performance. The modified model outperformed other methods as 0.685 and prediction performance was appropriate at 100 m rectangle grid. Regional characters of DRT bus allocation areas were extracted by K-means clustering method and differentiate urban and sub-urban types. The findings of this study provide valuable insights into conditions that DRT bus stop can be installed. The urban bus stop areas located in metropolitan cities and the suburban bus stop allocation areas located in countryside. The study results can be used as policy data for the successful introduction to prevent social exclusion and improve resident welfare in the future.

1. Introduction

1.1. Overview

Public transportation, such as buses and subways, is the only means of transportation for citizens who do not have cars, and it is an essential service for living [1]. The administrative division of the Republic of Korea consists of 8 provinces and 1 special city. The province includes sub-administrative districts such as '*si*' and '*gun*'. '*si*' is an area with a population of 100,000 or more, and '*gun*' is an

* Corresponding author. Department of Smart City Engineering, Hanyang University, Ansan, 15588, Republic of Korea.
E-mail addresses: ehdgml1247@naver.com (D. Oh), asd4950@hanyang.ac.kr (S. Lee), juneyoung@hanyang.ac.kr (J. Park), jhpark@kict.re.kr (J. Park), rohcg@kict.re.kr (C.-G. Roh).

https://doi.org/10.1016/j.heliyon.2023.e21213
Received 23 June 2023; Received in revised form 13 October 2023; Accepted 18 October 2023
Available online 24 October 2023

area with a population of less than 100,000. Population-induced demand significantly influences the availability of public transportation. According to the 'Resident Registration Population Statistics' of Republic of Korea's Ministry of Public Administration and Security, the population in the capital area (50.1 %) outnumbered that in the non-capital areas as of 2019. Thus, the capital areas are witnessing an increase in population density and the rural areas are showing decreased population density and outflow of people. As population density decreases, public transportation and service supply demands decrease, reducing the efficiency of traditional fixed-route bus-based public transportation systems. For example, fixed-route public transportation discontinued after a short operation period in rural transportation services owing to a fare cost recovery rate, limiting the utilization of important services [2]. Hence, public transportation in vulnerable areas is more crucial to people who have difficulty take transportation. Moreover, the freedom to migrate is limited for residents in public transportation vulnerable areas, contributing to social exclusion. Social exclusion refers to social problems related to the fragmentation of conventional social institutions, less involvement in regular societal processes, and increased deprivation among specific social groups [3,4]. A tackling social exclusion concerns involvement in activities which provide social interaction with others in the community [5]. It is becoming an essential component of social policy debate, but it is becoming to limit areas of economic poverty and income disadvantage [6,7]. To prevent social exclusion, an accessibility analysis such as medical facilities and public transportation is being performed in several studies [8,9]. Local governments are focusing on implementing a demand-responsive transit (DRT) system that combines the benefits of buses and taxis to reduce residents' social exclusion in public transportation vulnerable areas and ensure mobility. The DRT system is a transportation service that adapts the route or schedule of vehicles to meet passenger needs [10]. This study considered introducing the DRT system to reduce social exclusion and expand mobility rights for residents in public transportation vulnerable areas. Yangsan-*si*, which adopted the DRT system, was used as a standard region to examine public transportation vulnerable areas around Gyeongsangnam-*do*. GIS analysis was used to create data, classify grids, and examine predictive performance by grid size using various datamining methods such as classification and regression tree (CART), random forest (RF), support vector machine (SVM), Multivariate adaptive regression splines (MARS), Automated machine learning (AUTOML)_default, and Automated machine learning (AUTOML)_modified model. And then conducting K-means clustering model to classify type of DRT bus stop allocation. The study chose a grid technique appropriate for the domestic environment to derive characteristics that support demand-responsive bus stop construction and identify public transportation vulnerable areas and DRT system priority locations.

1.2. Literature review

In this study, several studies were reviewed to analyze for selecting and visualizing areas vulnerable to public transportation. Geographic information systems (GISs) and data mining models include various analysis methods and have been used in studies that is similar to this research topic. GIS analysis was considered for accessibility and visualization, and data mining models were considered for selecting public transportation vulnerable areas.

1.2.1. Geographic information systems (GISs)

GIS analysis was considered for accessibility and visualization, and data mining methodology was considered for selecting vulnerable areas for public transportation. GIS can be used to select a facility's location based on geographical distribution and spatial structure. Moreover, it can analyze information on spatially distributed phenomena [11]. It is necessary to create data for analysis by combining geographical data by grid through the GIS program. At this time, since each data affects a range of total data characteristic, it is necessary to establish standards of data scope through the literature in which GIS analysis has been performed. Ma, Q. et al. used GIS analysis to develop a shared electric scooter and subway connection plan. Land use data and location data were used, and restaurants within an OD(Origin-Destination) of 100 m were counted [12]. Sadeek used a GIS map to create rectangular grids ranging from 50 to 400 m and analyzed to examine transportation accessibility, crime trends, and land use. Through using SVM and GIS analysis, a 300 m*300 m grid was proposed as the best outcome [13]. Yao determined the predictive demand for public transportation; the SOM algorithm, a data mining tool. It was used to assign weights with the results classified and shown in the GIS using the natural break method [14]. Xue et al. investigated bus and subway accessibility to forecast house values based on transportation accessibility [*15*]. Yi and Kim examined medical facility accessibility and determined the optimal radius at 2100 m [16]. Kim et al. used a GIS buffer analysis to confirm the spatial distribution of vulnerable classes of public transportation services with the range of influence of the bus stop set to 300 m, while isolated areas where is more than 300 m from bus stop were assessed as public transportation vulnerable areas [1]. Lee et al. used a buffer analysis to assess the impact on student obesity. For comparison, the impact ranges for schools (elementary, middle, and high school) were set to 1000 m or 1500 m, respectively. The parameter stated in "Rules for determination, structure, and installation criteria of *si/gun* planning facilities" was elementary school walking distance [17]. Rattan et al. used GIS analysis to assess walking distance. The maximum walking distance in the analysis was 400 m from each public transportation method and 1500 m to an elementary school [18]. The buffer range for hospitals, elementary schools, middle schools, and high schools was set at 2100 m, 1000 m, 1000 m, and 1500 m, respectively, and the bus stop buffer range at 300 m and 500 m.

1.2.2. Data mining techniques

Data mining has been used as an analytical technique in various scientific domains for years [19]. The prediction accuracy was improved by introducing and comparing data mining techniques to overcome the constraints of the statistical empirical model. To select public transportation vulnerable areas, we reviewed similar data mining topic in the regression analysis series that can compare predictive performance with the same evaluation indicators. Xue et al. revealed that the RF model predicts regression well. RF models have allowed for complex non-linear interactions in modeling numerous variables through the substantial use of big data [15].

Tehrany et al. assessed flood-vulnerable areas using SVM and GIS. The study proved the accuracy of the SVM model and demonstrated their efficiency for generating maps in the GIS environment [20]. Pradhan compared decision trees(DT), SVM, and neural fuzzy inference system models for analyzing landslide vulnerability. The DT model generated the most accurate predictions, although all three methods produced efficient results [21]. Rahmati et al. used RF and MARS models to find and map potential groundwater areas. Both models performed well and It was an example that demonstrated the performance of the tree-based model [22]. Chen et al. compared the spatial prediction methodologies of RF, LMT, and CART for landslide vulnerability assessment and determined that the RF model generated the best results [23]. Lee et al. assessed the best personal mobility service area using an ensemble-based data mining technique. Though RF and gradient boost methods(GBM) offered excellent performance, the gradient boost analysis was more accurate than RF. They evaluated model performances using the root mean square error (RMSE) and coefficient of determination (R^2) metrics [24]. Cong et al. developed a traffic prediction model using the least square SVM model and evaluated it using the mean square error (MSE) and mean absolute percentage error (MAPE) [25]. Stadler et al. forecasted bus demand for rural area using XGBoost, RF, and Naïve bayes model. The best prediction performance is 87 % and the model was RF. They used a different time of day intervals and weather to predict dynamic result [26]. Khan et al. performed bus optimization, such as predicting the number of passengers and allocating scheduling. Based on the MARS algorithm, they tried to predict the optimal demand, and predicted the number of passengers on all routes, including date, day of the week, time, and seats [27]. Imhof & Blättler modeled demand of DRT service by using rural area data in Swiss. they analyzed service area by 300*300 m raster and use the random forests algorithm to predict demand within and across areas [28]. Caicedo et al. developed the LSTM forecasting model to predict short-term public transportation demand using smart cards and time series data. It was mentioned that the LSTM model is suitable for predicting demand in dynamic situations due to COVID-19 [29]. Ma et al. performed a time series analysis to predict short-term bicycle sharing demand and introduce a system. They made predictions based on stations and proposed the STGA-LSTM (Spatial-Temporal Graph Attentional Long Short-Term Memory) framework, which focuses on both temporal and spatial dimensions [30]. Yang et al. used a multivariate linear regression model and SVR (Support Vector Regression) to determine the influence of environmental factors on the spatial distribution of bicycles, and analyzed it on a 500 m*500 m grid. The performance of SVR based on the Gaussian Radial Basis function was better, and it was confirmed that regional factors such as financial institutions, residential areas, and commercial areas had a significant influence on the installation area of shared bicycles [31]. Cui et al. (2018) focused on the recommendation of the number of shared bicycles near the subway station and take the volume of the station's outbound passenger flow as the potential demand. They developed a novel passenger flow forecast model on advanced Xgboost method and the idea of sliding window and recommended a suitable number of shared bicycles for a subway station [32].

Based on previous studies, GIS analysis was conducted to combining data and visualizing results by applying buffer analysis, data combination, and Jenks' natural break method. This combined data was processed to be analyzed and used for various data mining models. CART, SVM, RF, MARS, and AutoML were used among the data mining models, and the prediction results were compared and evaluated through RMSE, MAE, and R^2.

In this study, it is the main novelty that used successful cases of DRT operation as a reference and predicted bus demand in the supra regions of successful cases by considering the results according to regional characteristics. demand-responsive transit is a transportation which effects are depending on the installed location, so selecting a site is important According to the scope of bus stop influence, the study area was divided into three grid sizes (100 m, 250 m, 500 m). Since there are differences in results depending on bus demand, regional location and characteristics, data which may affect bus demand were collected and combined. Existing literature used time series data and performed time series analysis. However, in this study, bus demand data, which is the dependent variable, can be used to collect monthly demand data for each stop through a web-data crawling method, resulting in various results depending on the scope of influence of the bus stop. Following is a summary of the specific research objectives and its essential intellectual merits.

1. We proposed a combination among the AutoML model packages not only using best model: Deep Neural Network(DNN), Gradient Boosting Model(GBM)
2. To suggest best performance model and grid size, we used comparison of evaluation indicators such as RMSE, MAE, R^2. We can identify the lower values of analysis results.
3. A novel definition and standards of public transportation vulnerable area with prediction results were proposed; simply an area where there is a bus demand and far from bus stop, or an area where the demand is numerically above a certain value and far from bus stop over 300 m.
4. It was identified that vulnerable areas can be divided two types of DRT allocation area; urban DRT bus stops and suburban DRT bus stops and each type of area has common characteristics.

2. Materials and methods

2.1. Methods

Methods for measuring access to public transportation are divided into infrastructure-based accessibility measurements, location-based accessibility measurements, human-based measurements, and utility-based measurements [33–35]. In this study, accessibility is measured based on location. We combined each data using buffer analysis with 100, 250, and 500 m grids and assessed the predictive performances of specific vulnerable grid of total data. Many researchers have used various methodologies such as statistical methods, time series analysis, machine learning, deep learning to develop bus demand forecasting or prediction models. We used tree-based CART and RF, SVM, MARS, and AutoML models to predict public transportation bus demand. After evaluating outcomes, the grid

size and model with the highest prediction were chosen. Based on the results of regression models, we established definition of public transportation vulnerable areas. Through using K-means clustering model, regional characteristics of areas where DRT bus stops might be introduced were obtained by matching with grid where DRT bus stops are installed. Fig. 1 depicts a simplified framework of the analysis process.

2.1.1. Classification and regression tree(CART)

CART is a non-parametric model that specifies the functional form and does not rely on predictor additivity assumptions. If the output is a categorical variable, the Gini index is employed as an impurity measure; if the output variable is continuous, binary separation is conducted using a variance. The branching process begins with variables of high importance. Its goal is to reduce the "impurities" of nodes in a classification tree [36].

2.1.2. Random forest regression(RF)

RF is an ensemble technique for tree-based prediction models, such as decision trees, developed to solve the problem of decision tree analysis [37]. It reaches consensus by gathering classification data from several trees built through training. Data is sampled randomly for each tree, the parameters are adjusted to diversify the tree's properties, and the classification results of each tree are collected to form a forest [15].

2.1.3. Support vector machine regression(SVR)

SVM is a supervised learning model that learns from training data containing labels for classification or regression analysis. The hyperplane classifies data, and after training on the given data, it learns the class of the new data [38]. The model is developed as **Equations (1)–(4)**:

$$y = \left(w^{T}.\Phi(x) + b\right) + \text{noise}\right) \tag{1}$$

$$\frac{1}{2}w^{T}.w + C\sum_{i=1}^{N}\varepsilon_{i} + C\sum_{i=1}^{N}\varepsilon \bullet I \tag{2}$$

$$w^{T}.\Phi(\varepsilon) + b - yi \leq \epsilon + \varepsilon \bullet i \tag{3}$$

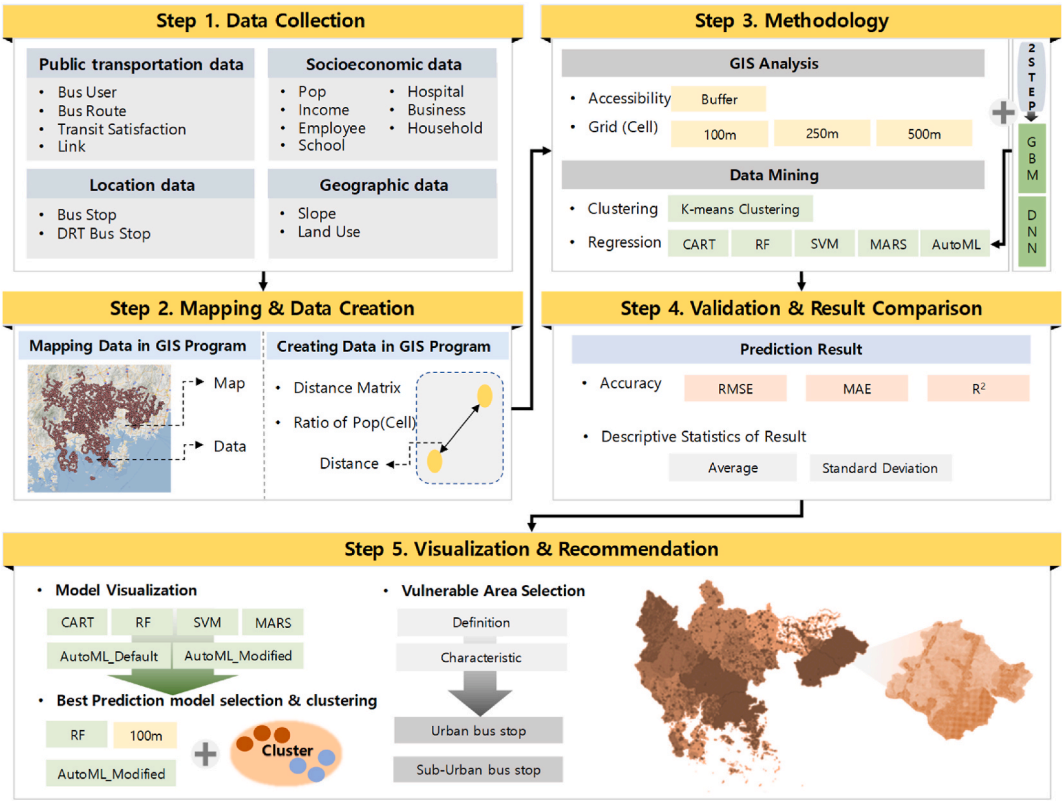


Fig. 1. An overview of the framework.

$$\varepsilon \bullet i, \varepsilon_i \geq 0, i = 1 \tag{4}$$

The noise parameter (ε) in the regression model is expressed as error tolerance. w is the coefficient vector, b is a constant, and Φ represents the kernel function. C is a positive constant that controls the degree of loss when an error occurs, N is the sample size, and ϵ X are slack variables that specify the upper and lower calibration errors of ε [39].

2.1.4. Multivariate adaptive regression splines(MARS)

The MARS method is appropriate for high-dimensional regression problems with several input variables. It is a generalization of stepwise linear regression or an enhancement of decision trees. MARS employs a piecewise linear regression basis function as $(x-t)_+(t-x)_+$ at knot points defined at value t, where $(x)_+ = x(f(x > 0))$ reflects only the positive component of the value x in parenthesis as **Equation (5)**:

$$(x-t)_+ = \begin{cases} x - t, & x\rangle t \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

$$(t-x)_+ = \begin{cases} t - x, & x\langle t \\ 0 & \text{otherwise} \end{cases}$$

The class of the basis function is expressed as splines with observed values x_{1j} as knot points for each input variable X_j. The MARS model is expressed as **Equation (6)**:

$$B = (X_j - t)_+, (t - X_j)_+ : t \in x_{1j}, \cdots, x_{nj} \text{ for } j = 1, \cdots, p \tag{6}$$

B_m (X) is a basis function in B or the product of two or more basis functions that belong to B. Rather than using the original input values, forward-step linear regression is used for modeling. When B_m (X) is given, we estimate the B_m coefficients that minimize the sum of squared errors, and the β_0 is intercept as **Equation (7)**:

$$\mu(X) = \beta_0 + \sum_{m=1}^{M} \beta_m B_m(X) \tag{7}$$

If only one basis function in B is used, the model is additive and solely employs the primary effect. The basis function in MARS is chosen using forward selection. Thus, $\beta_0(X) = 1$ is fed to the model, variables and knot points that minimize the sum of squared errors at each step are determined, and the corresponding basis function pair is added to the model. **Equation (8)** is the criterion for selecting a basis function:

$$GCV(m) = \frac{\sum_{i=1}^{n} (y_i - \widehat{\mu_m}(x_i))^2}{(1 - C(m)/n)^2} \tag{8}$$

where $\widehat{\mu_m}$ is the fitted value of μ(x) based on m terms, n is observations, and C(m) is the complexity function defined by the number of parameters. Finally, a model with m* terms where m* = argminGCV(m) is selected [40].

2.1.5. Automated machine learning (AutoML)

AutoML is a systematic model that automates the algorithm selection and hyper-parameter tuning processes. AutoML consists of the following three key components: a search space, a search strategy, and a performance evaluation strategy. The search space refers to a set of hyper-parameters and the range of each hyper-parameter. The search strategy refers to the strategy of selecting the optimal hyper-parameters from the search space. The performance evaluation strategy refers to the method used to evaluate the performance of the trained models. In the study, the H2O AutoML platform was adopted for the assessment of Bus demand. Generalized linear model with regularization (GLM) is an extended form of a linear model. Given the input variable x, the conditional probability of the output class falling within the class c of observations is defined as **Equation (9)** Where β_C is the vector of coefficients for class c:

$$\widehat{y_c} = \Pr(y = c|x) = \frac{ex^T\beta_C + \beta_{C0}}{\sum_{k=1}^{K}(ex^T\beta_k + \beta_{k0})} \tag{9}$$

The distributed random forest (DRF) is an ensemble learning approach based on decision trees. In the DRF training process, multiple decision trees are built. To reduce the variance, the final prediction was obtained by aggregating the outputs from all decision trees. Like the DRF, extremely randomized trees (XRT) is based on multiple decision trees, but randomization is strongly emphasized to reduce the variance with little influence on the bias. The following main innovations are involved in the XRT process: random division of split nodes using cut points and full adoption of the entire training dataset instead of a bootstrap sample for the growth of trees. The DNN in H2O AutoML is based on a multilayer feedforward artificial neural network with multiple hidden layers. There are many hyper parameters involved in DNN training, which makes it notoriously difficult to manually tune. Cartesian and random grid searches are available in H2O AutoML for DNN hyper-parameter optimization. GBM is an ensemble learning method. The basic idea of GBM is to combine weak base learners for the generation of strong learners. The objective is to minimize the error in the objective function

through an iterative process using gradient descent. In addition, stacked ensembles can be built using either the best-performing models or all the trained models [41]. AutoML allows you to select the optimal model without tuning hyper-parameters, but models that combine individual models presented in AutoML's format can have excellent predictive performance. If the optimal model selected from the AutoML model is stacked ensemble, the result of combining various models is presented, so you can proceed with the analysis by combining individual methods. Therefore, in this study, the models were combined by two STEPs and AutoML_Modified was presented.

STEP1. Consider the characteristic of data. As the analysis data in this study was collected through individual sources, there are many variables that have the same results depending on the administrative district. In the case of a 100 m grid, there is a variable having a value of '0' because it is generated with multiple grids. Since data of the dependent variable is predicted from a range of 1–6,000, the range of errors may be increased. The distribution of data can consider many dimensions due to the enormous number of explanatory variables, and it is necessary to apply a methodology that can clearly visualize the data.

STEP2. Choose models to combine. Among the results excluding AutoML, the analysis result of the RF model was the highest. This is a model through a bagging algorithm and is one of the ensemble models. Since the bagging method analyzed in parallel, it has the characteristic of temporarily extracting and analyzing data. In contrast, the gradient boosting method is a method of extracting data and continuously reducing errors and has high predictive performance. Various tree-based prediction methods are compared in this study. The GBM model was suitable as an analysis model. The current research in Abdulhammed et al. and Leem et al. affirms that DNN and GBM are the best methods to predict performance [42,43]. The results indicate that GBM tends to be faster and more potent than DNN due to its lower processing requirements. GBM is hence the preferred method for credit scoring prediction [44]. The result of a GBM single model that can reduce errors in data for each grid and consider all various characteristics is not a high-ranking model, so to improve prediction performance, DNN model was combined. Deep learning models have high prediction performance and are usually evaluated with similar predictive performance to GBM models. In this study, the results were derived by combining DNN and GBM models to utilize total data collected for one month for analysis. Best model was used as default, and combination model was used as modified.

2.1.6. Model validation

The RMSE, MAE, and R^2 indicators validate regression-based data mining analysis approaches and are used to evaluate the prediction accuracy of models. The RMSE is statistic difference between an estimated or projected value and actual value. The MAE measures the average magnitude of the errors in a set of forecasts. The lower the MAE and RMSE, the better the regression model's prediction accuracy. R^2 describing the prediction performance, is calculated by dividing the sum of squared residuals (SSR) by the sum of squared total (SST). RMSE, MAE, and R^2 are calculated as **Equation (10)–(12)**. Where y_i is the prediction result, $\widehat{y_i}$ is the observed value (real value), and n is the number of samples:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n^2} \frac{(y_i - \widehat{y_i})^2}{n}} \tag{10}$$

$$\mathrm{MAE} = \frac{1}{n} * \sum_{i=1}^{n^2} |y_i - \widehat{y_i}| \tag{11}$$

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} \tag{12}$$

2.1.7. K-means clustering

The k-means clustering model was first introduced by MacQueen [45]. This model partitions a set of data into k clusters in such a way that the sum of squared errors between the mean of each cluster and the existing data in the cluster is minimized. clusters are created by repeating the process of moving the center point to the center of the classified cluster.

Table 1
Hyperparameter list of models.

Models	Libraries	Hyperparameters	100 m grid	250 m grid	500 m grid
CART	rpart, fit.tree, prune	bestcp(nsplit)	9	12	13
RF	System.time, trainControl, rf, RandomizedSearchCV	n_estimators	250	300	250
		max_depth	4	4	4
SVM	SVR, RandomizedSearchCV	C	5.76	21.83	38.84
		Gamma	0.34	0.09	0.10
		Kernel	rbf	rbf	rbf
MARS	caret, earth	degree	2	2	2
AutoML_Default	H2O(h2o.automl)	max_model	50	50	50
AutoML_Modified		nfolds	10	10	10
		fold_assignment	Modulo	Modulo	Modulo
		distribution	gaussian	gaussian	gaussian
K-means clustering	wss, kmeans	NbClust	–	5	-

2.1.8. Hyperparameter list for bus demand prediction models

To predict bus demand, the analysis process was conducted in Python and R programming language. the libraries and hyperparmeters implemented for each individual model are presented in Table 1. Random search method or the best model was selected. The CART model considered 'bestcp', and the RF model considered 'n_estimator' and 'max_depth'. The SVM model adjusted 'C', 'Gamma', and 'Kernel', and the MARS model considered the 'degree of interaction'. AutoML adjusted 'max_model', 'nfolds', 'fold_assignment', and 'distribution for deep learning', and the K-means model considered 'NbClust'.

2.2. Data preparation

Gyeongsangnam-*do* was chosen as the analysis region in this study except Changwon-*si*, Geoje-*si*, and Geochang-*gun* due to issues with data collection. Data were composed by referring to Foda and Osman [46], which collected data such as population, income quintile, and bus stop to predict the number of bus passengers, and cases that classified data by various characteristics such as demographic characteristics, service characteristics, land use and socioeconomic index characteristics [14,47–50]. As population-related variables affect bus demand in rural areas, the number of single-person families and the number of people aged sixty-five and older were collected. Demand data by bus stops were collected by web data crawling method with the number of passengers and routes served. Bus data was derived from 0:00 to 24:00 for one month from April 1, 2022. The number of bus passengers was the dependent variable indicating public transportation demand, whereas the other variables served as explanatory variables. Table 2 lists all the data used in the analysis. As shown in Fig. 2, data was mapping at GIS map by grid sizes.

3. Results

3.1. Descriptive statistics of prediction data with models

We performed correlation analysis on the processed data, and analysis was performed using significant variables derived through correlation analysis for each grid size. Outlier data that was different from the general bus stop demand was removed, and scaling was performed to equalize the scale of the data. We attempted to perform analysis and compare the performance of each analysis result. Table 3 lists the descriptive statistics of the predicted number of bus passengers. Descriptive statistics is consisted of mean, standard deviation, minimum, maximum values. The prediction results were examined by grid sizes and models (CART, RF, SVM, MARS, and AUTOML). The higher the value of the statistics, the more bus demands are projected. The standard deviation (SD), which is the square root of the positive variance and represents a value for addressing the overestimation problem, is used to calculate the difference between variances based on the mean. The bigger SD, the easier it is to compare the visualization impacts [24]. The 100 m *100 m grid consists of 877,169 grids. The 250 m*250 m grid is composed of 208,592 grids. The 500 m*500 m grid consists of 37,694 grids. Since this results are normalized, the maximum value is 1 and the minimum value is 0. Since SD is the smallest and mean is the largest, the visualization will not be clear in SVM model. In the MARS model, both SD and mean values are small. Therefore, the visualization results will be ambiguous.

3.2. Validation

AutoML_Modified model applied 1 DNN and 1 GBM. Table 4 shows the analytical model's prediction accuracy for each grid size. When the prediction performance of all models was compared with models in existing literature, it was confirmed that they had similar

Table 2
List of variables and four data types of local characteristics.

Data type	Variable	Source
Socioeconomic Data	Population(grid)	National Spatial Information Portal
	No. of single person household	Statistical Geographic Information Service
	No. of people aged 65 and over	
	No. of employees	
	No. of businesses	
	Income	National Spatial Information Portal
	No. of hospital	
	No. of high school	
	No. of middle school	
	No. of elementary school	
Locational Data	Location of bus stop	Address Information Portal
	Location of DRT bus stop	
Public transportation Data	No. of bus passenger	Transportation Card Big Data Integrated Information System
	No. of bus Route	
	Walking satisfaction	National Statistical Office
	Link/Node	KTDB
	Distance matrix	GIS based data
Land Data	Slope	National Spatial Information Portal
	Special-Purpose Area	

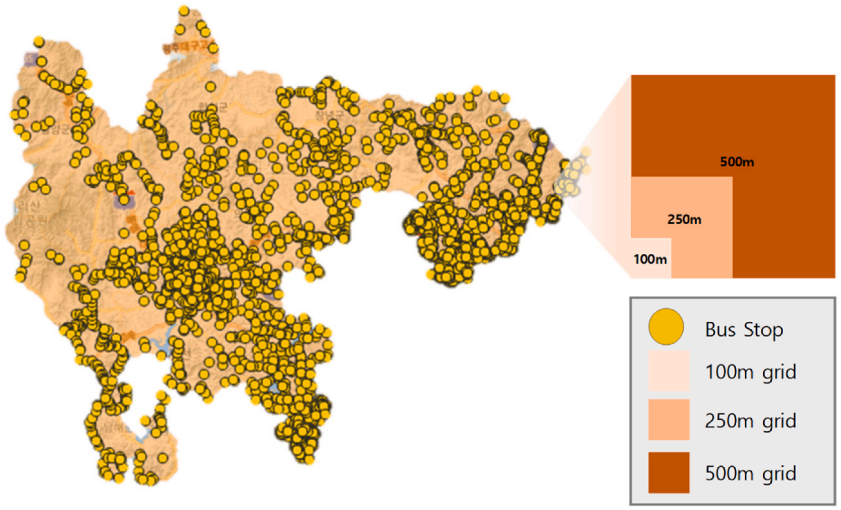


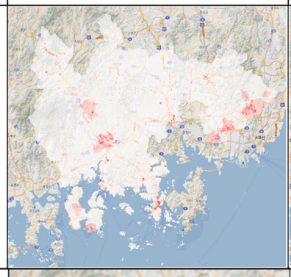
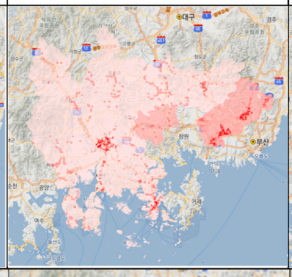
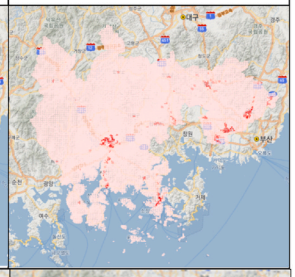
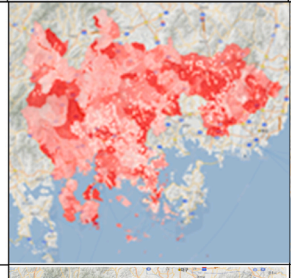
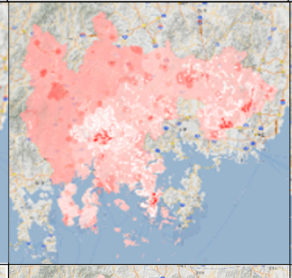
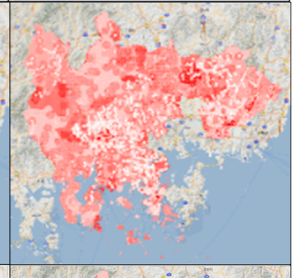
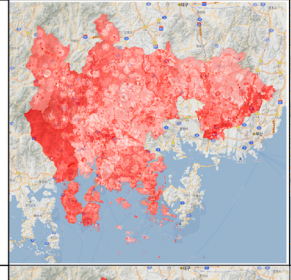
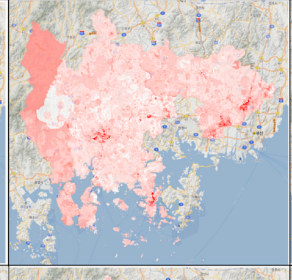
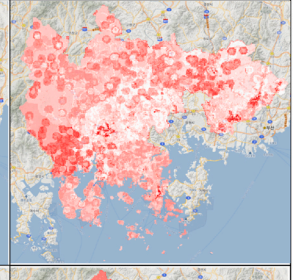
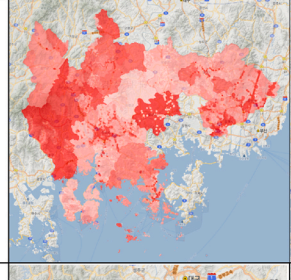
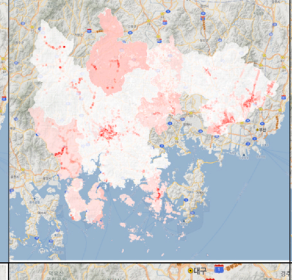
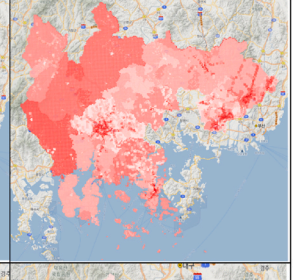
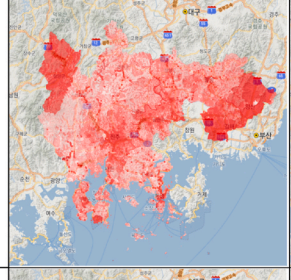
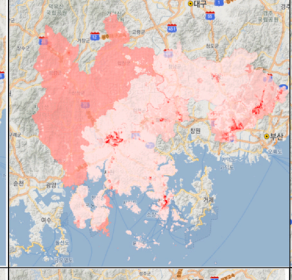
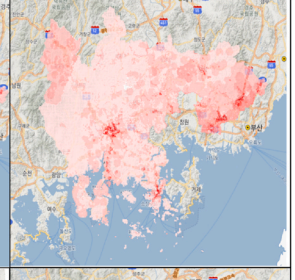
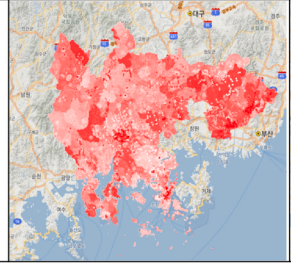
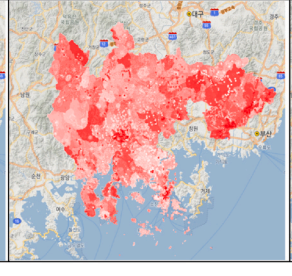
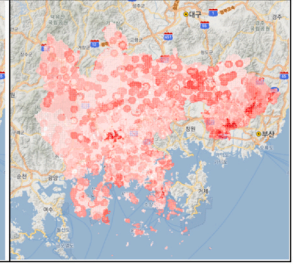
Fig. 2. Mapping bus stop location and depicted analysis grid sizes.

Table 3
Descriptive Statistics of Prediction results by grid sizes.

Grid	Models	Statistic Indicator				
		Mean	Standard deviation	Minimum	Maximum	Observation
100(m)*100(m)		0.061	0.051	0.000	1.000	877,169
	RF	0.086	0.071	0.000	1.000	
	SVM	0.180	0.036	0.000	1.000	
	MARS	0.160	0.042	0.000	1.000	
	AutoML_Default	0.063	0.047	0.000	1.000	
	AutoML_Modified	0.073	0.043	0.000	1.000	
250(m)*250(m)	CART	0.217	0.217	0.000	1.000	208,592
	RF	0.143	0.136	0.000	1.000	
	SVM	0.187	0.103	0.000	1.000	
	MARS	0.037	0.037	0.000	1.000	
	AutoML_Default	0.136	0.126	0.000	1.000	
	AutoML_Modified	0.149	0.120	0.000	1.000	
500(m)*500(m)	CART	0.050	0.042	0.000	1.000	37,694
	RF	0.129	0.081	0.000	1.000	
	SVM	0.140	0.055	0.000	1.000	
	MARS	0.370	0.033	0.000	1.000	
	AutoML_Default	0.111	0.063	0.000	1.000	
	AutoML_Modified	0.041	0.047	0.000	1.000	

Table 4
Model Precision Accuracy evaluated with indicators.

Grid	Evaluation Indicator	CART	RF	SVM	MARS	AutoML	
						Default	Modified
100*100(m^2)	RMSE	0.200	0.167	0.154	0.178	0.162	0.149
	MAE	0.126	0.099	0.077	0.109	0.093	0.087
	R^2	0.437	0.606	0.439	0.550	0.625	0.685
250*250(m^2)	RMSE	0.290	0.171	0.151	0.178	0.163	0.146
	MAE	0.221	0.100	0.076	0.109	0.095	0.085
	R^2	0.482	0.588	0.475	0.554	0.623	0.700
500*500(m^2)	RMSE	0.197	0.168	0.157	0.176	0.161	0.149
	MAE	0.122	0.098	0.079	0.107	0.094	0.089
	R^2	0.452	0.602	0.421	0.564	0.624	0.685

Model		100 m*100 m	250 m*250 m	500 m*500 m
CART				
RF				
SVM				
MARS				
Auto ML	Default			
	Modified			

(caption on next page)

Fig. 3. Visualization of predicted demand result by models.

performance. AutoML_Modified model showed best RMSE performance in 100 m*100 m grid. The value is 0.149. MAE performances were good at the SVM model. AutoML_Modified and AutoML_Default model outperformed other models with R2. If predicted results showed larger variance, the visualization expected to be more pronounced [24]. The standard deviation of bus passengers (dependent variable) is more than 2300 and the value must be scaled due to affecting final results. Normalization was performed on the entire data, and the data was used for analysis. The SVM model is sensitive on the distribution of data. The variation of the results is smaller than other models. The CART model tended to vary depending on the number of branches of the data, and the accuracy was the highest in the 500 m grid, but the accuracy was the lowest in the overall model's results. In addition, when comparing the accuracy of MARS and CART models based on tree techniques, the prediction performance of the MARS model was higher, and the error rate was lower. The AutoML model outperformed other models and RF model performed well than CART, SVM, MARS. This is the same as the results of Sun et al. comparing SVM, RF, and LR models in previous studies [51]. When comparing grid-specific analysis from 50 m to 400 m in various previous studies, the 300 m grid showed the highest accuracy [12,51]. In this study, the 250 m grid was most accurate.

3.3. Prediction results and visualization

The prediction results for each model are visualized in Fig. 3. As a result, Gimhae-*si*, Yangsan-*si*, and Hamyang-*gun* appeared in dark colors. Gimhae-*si* and Yangsan-*si* have the highest population and are classified as metropolitan areas in Gyeongsangnam-*do*. Hamyang-*gun* has small the population, but transportation infrastructure facilities were well established. In the region where is the well-equipped transportation facilities, prediction performance was accurate [52]. In CART model result of the 100 m*100 m grid, the branch conducted nine times but it is not suitable to classify about 880,000 grids. In RF model, 100 m *100 m grid result was visually clarified than 250 m*250 m grid. When comparing results between AutoML models, the visualization results from modified model reflect the distribution of darker colors. In case of modified model, there is a difference in metropolitan cities compared with other regions. In county with a small population, such as Uiryeong-*gun*, the population density compared to metropolitan cities is more than twenty times. This finding was similar to previous studies showing that high population density affects travel [53]. The distribution of population showed a difference in demand forecasting, and most regions in Yangsan-*si* and Gimhae-*si* have high prediction performance in visualization.

3.4. Characterizing the public transportation vulnerable areas

The DRT bus stop area installed in Yangsan-*si* is characterized by K-means clustering model. The appropriate number of clusters was derived through Within-Cluster-Sum of Squares (WSS), and the number of clusters was found to be 5. The results are visualized in

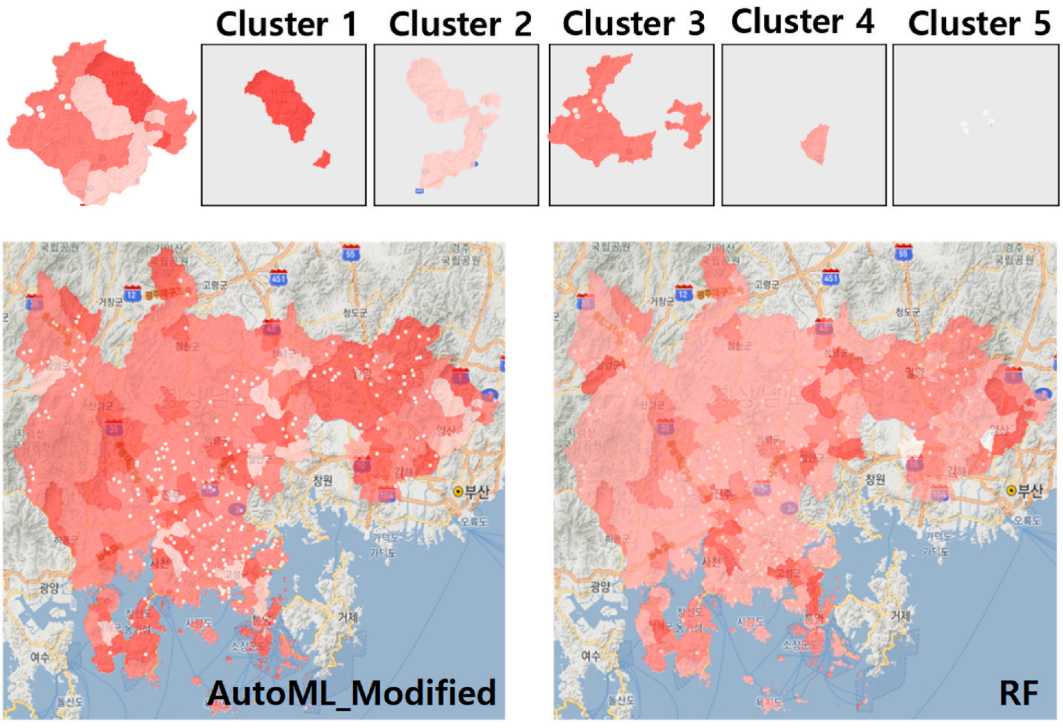


Fig. 4. Visualization of K-means clustering analysis results using 250 m*250 m grid prediction values.

Fig. 4. The cluster 3 includes non-urban or rural areas, and these results overlapped with The DRT Routes 1, 2, 7, and 8. The cluster 2 included industrial complexes or urban areas and overlapped with DRT routes 3, 4, 5, and 6.

Table 5 represent urban and suburban DRT bus stops characteristics. Urban DRT bus stops have nice transportation accessibility compared to underdeveloped areas. but compared to convenient transportation environments, they are allocation areas to resolve inconvenience. Its economic affordability is higher than that of Suburban DRT bus stops, and it is in a commercial or residential area. Suburban DRT bus stops have a small population, negative economic affordability, and inconvenience to access, so they need to be improved. They need to install the result of the area near the apartment complexes, or a company located at the end of an industrial complex that is little far from the bus stop. And they are in an area with a high slope and inconvenient to move on foot, so an urgent installation is recommended rather than an Urban DRT bus stop. The number of population was found to be the most important spatial and local characteristic to predict DRT demand. Increasing number of population per grids lead to higher demand predictions. This may be explained as increasing the number of population increases the number of potential users, underscoring the principle of the "rural mobility problem" caused by low population size and density [28,54]. The finding on the interrelation between population density and demand for trips is in line with previous research on urban flexible transport services [28,55,56].

3.5. Selection of public transportation vulnerable areas

This section shows the results obtained when the model was applied to the entire area of Gyeongsangnam-*do*. The higher the predicted value, the higher the demand. There are 36 predicted locations which urban DRT bus stops can be installed, and all of area are in Yangsan-*si* and Gimhae-*si*. In the case of suburban bus stops, it needs to be considered to secure mobility even when there is little. The maximum value of urban bus stop allocation area is 0.822, and the minimum value is 0.022. The highest demand was predicted in apartment complexes near Beomeo-*ri* of Yangsan-*si*. The least predicted area was found to be located near a shopping complex near Jinyeong-*ri* of Gimhae-*si*. Beomeo-*ri*, Gachon-*ri* of Yangsan-*si* and Jinyeong-*ri,* Yeorae-*ri* of Gimhae-*si* were classified as urban DRT bus stop sites. There are 108 predicted locations where suburban DRT bus stops can be installed. The areas are mountainous areas, but include cases where residents are lived in. Vulnerable areas were consisted of 71 areas in Gyeongsangnam-*do*. The areas that require installation with the highest priority are in *gun* areas. The maximum value of suburban DRT bus stop area is 0.058, and the minimum value is 0.004. Fig. 5 shows areas where have predicted Top 5 values of each bus stop type. Although the bus stop is currently located, it was selected as an area that requires a bus stop because it was not included in the data used in this study. This means that the prediction has been carried out appropriately. This results are similar to Lee et al. in that PM service areas are installed in places with high population density and developed commercial districts [24].

4. Discussion and conclusions

In this study, we attempted to predict the demand of bus passengers in by selecting the optimal model, and to identify vulnerable areas of study area for public transportation by comparing with areas where demand-responsive bus stops are located. It is important to solve problems by quickly introducing transportation in areas where public transportation is vulnerable, and there are implications in that the research results were based on actual data. We collected data from web-data crawling method. Real bus demand data have been collected to train the regression models, and their performances are comparatively measured based on evaluation metrics described in Section 3.2. These metrics were also used to interpret the results of each applied machine learning model and based on which the best performing regression model is identified. Social demographic, location, public transportation, and land use data were collected and processed. Conducting GIS analysis, 100 m, 250 m, and 500 m rectangle grids were created, and data were combined. Various data mining models were used to predict the number of bus passengers. Among the AutoML models, the modified model consisted of one DNN and one GBM, with the highest prediction performance. The best model was 'AutoML_Modified', and analysis grid size was selected by 250 m. According to previous results, clustering analysis was performed and the characteristics were derived. The allocation areas of DRT bus stop can be divided into urban DRT bus stops and suburban DRT bus stops. The characteristics of the urban DRT bus stops location were identified as commercial areas and residential areas, 126 to 254 people, two lanes (one-way), 0–2° of slope, within 300 m of bus stops, and two to three quintiles of income. The characteristics of the suburban DRT bus stops location were identified in agricultural and forestry areas, 8 to 15 residents, 1 to 2 lanes (one-way), 1 to 2 quartiles of income, side roads, 30–60° of slope, and more than 500 m of bus stops. This result suggests that the operation of the DRT system is more urgent, with most people engaged in the primary industry and several high-slope areas residing in agricultural and forestry areas.

From a social perspective, DRT systems should improve mobility, minimize public transport travel and waiting times, increase public transport satisfaction, and increase public transport mode sharing. In addition, economic effects can be expected such as replacing public transportation (buses), reducing subsidies, revitalizing the local economy, generating profits for transportation companies, and inducing labor. If local governments implement it and create the possibility of development, the welfare of public transportation vulnerable areas will be improved. *gun* areas frequently experience lack of data and are tough to create raw data. A processing system that creates an accurate database using public transportation and regional data is required. When collecting and processing regional data becomes difficult, usage will be low and excluded from the analysis. Therefore, *gun* areas must supply raw and processed data on public transit and geographical characteristics. There is a disparity in public transportation supply by *si* and *gun* area, and further studies are needed to refer this. Management strategies such as establishing a DRT system must be adopted in locations as public transportation vulnerable areas. Various decision processes, such as region, branch, and route selections, are necessary to establish the DRT system. If a national policy based on this approach is formed, various regions across the country can consider for implementation. This study has a limitation in that data were compared for places with good access to public transportation. Further

Table 5
List of characteristics by the DRT bus stop types in the reference region.

Type	Characteristic	Contents
Urban DRT bus stop	Special-Purpose Area	Commercial and residential area
	Population	126 to 254 persons
	Number of lanes	2 to 3 lanes
	Road type	Segregation road of pedestrian and vehicle
	Slope	0–2°
	Bus stop	A distance within 300 m
	Income	2nd to 3rd quartile
Suburban DRT bus stop	Special-Purpose Area	Agricultural and forestry area
	Population	8 to 15 persons
	Number of lanes	1 to 2 lanes
	Road type	Side road(or sharing road of pedestrian and vehicle)
	Slope	30–60°
	Bus stop	A distance farther than 500 m
	Income	1st to 2nd quartile

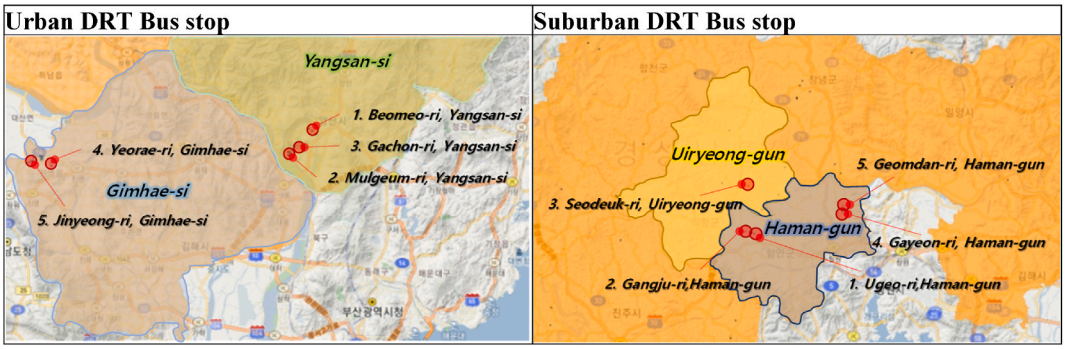


Fig. 5. Visualization of Urban and Suburban DRT Bus stop allocation that fulfilled the characteristics.

research is needed to increase the mobility of residents in *gun* areas. Accuracy is improved by segmenting the analysis, such as securing monthly data and separating characteristics by time zone. The results can be used as policy data to reduce social exclusion and promote civic welfare by successfully introducing and settling demand-responsive bus stops in the future.

Data availability statement

The authors do not have permission to share data.

CRediT authorship contribution statement

Donghee Oh: Writing – original draft, Validation, Software, Methodology. **Sangjae Lee:** Validation, Methodology. **Juneyoung Park:** Writing – review & editing, Supervision, Conceptualization. **Jaehong Park:** Writing – review & editing. **Chang-Gyun Roh:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research was supported by research project “Development of Sustainable MaaS(Mobility as a Service) 3.0+ Technology in Rural Areas” funded by the Korea Institute of Civil Engineering and Building Technology (KICT).

Appendix tables

Appendix A
Descriptive statistics of training data

Grid	Variable Population(grid)	Statistic Indicator				
		Mean	Standard deviation	Minimum	Maximum	Observation
100 m	Population(grid)	51.328	116.498	0.000	1253.000	4,075.000
*100 m	Slope	8.809	18.491	0.000	80.000	
	Income	2.350	0.559	0.000	4.467	
	No. of single person household	1480.559	2390.278	0.000	11374.000	
	No. of people aged 65 and over	402.317	443.966	0.000	1928.000	
	Walking satisfaction	3.194	1.471	1.500	6.800	
	Public transportation satisfaction	6.576	1.594	0.900	8.700	
	Distance matrix	214.960	493.905	0.000	10842.975	
	Link	1.439	0.755	1.000	4.000	
	No. of hospital	0.814	0.389	0.000	1.000	
	No. of high school	0.449	0.497	0.000	1.000	
	No. of middle school	0.429	0.495	0.000	1.000	
	No. of elementary school	0.574	0.494	0.000	1.000	
	No. of bus Route	5.444	8.121	1.000	95.000	
	Bus passenger	1315.889	3406.517	1.000	6000.000	
250 m	Population(grid)	376.516	619.408	0.000	3548.000	3,964.000
*250 m	Slope	11.206	20.498	0.000	80.000	
	Income	2.344	0.556	0.000	4.467	
	No. of people aged 65 and over	1868.939	2180.265	0.000	9793.000	
	Walking satisfaction	3.208	1.478	1.500	6.800	
	Public transportation satisfaction	6.564	1.609	0.900	8.700	
	Distance matrix	215.254	499.831	0.000	10842.975	
	Link	1.438	0.747	1.000	5.000	
	No. of hospital	0.827	0.379	0.000	1.000	
	No. of high school	0.459	0.498	0.000	1.000	
	No. of middle school	0.442	0.497	0.000	1.000	
	No. of elementary school	0.600	0.490	0.000	1.000	
	No. of bus Route	5.250	7.713	1.000	95.000	
	Bus passenger	1281.860	3331.882	1.000	6000.000	
500 m	Population(grid)	1320.145	1990.799	0.000	10288.000	3,983.000
*500 m	Slope	15.802	23.846	0.000	80.000	
	Income	2.349	0.560	0.000	4.467	
	No. of single person household	1480.823	2387.131	0.000	11374.000	
	Walking satisfaction	3.202	1.479	1.500	6.800	
	Public transportation satisfaction	6.564	1.605	0.900	8.700	
	Distance matrix	212.141	497.313	0.000	10842.975	
	Link	1.469	0.766	1.000	5.000	
	No. of hospital	0.484	0.500	0.000	1.000	
	No. of high school	0.849	0.358	0.000	1.000	
	No. of middle school	0.491	0.500	0.000	1.000	
	No. of elementary school	0.651	0.477	0.000	1.000	
	No. of bus Route	5.278	7.723	1.000	95.000	
	Bus passenger	1314.907	3405.429	1.000	6000.000	

Appendix B
Urban DRT vulnerable areas lists

Rank	Administrative area	Maximum prediction value
1	Beomeo-*ri*, Yangsan-*si*	0.822
2	Mulgeum-*ri*, Yangsan-*si*	0.723
3	Gachon-*ri*, Yangsan-*si*	0.631
4	Yeorae-*ri*, Gimhae-*si*	0.158
5	Jinyeong-*ri*, Gimhae-*si*	0.123

Appendix C
Top20 of suburban DRT vulnerable areas lists

Rank	Administrative area	Maximum prediction value
1	Ugeo-*ri*, Haman-*gun*	0.058
2	Gangju-*ri*, Haman-*gun*	0057
3	Seodeuk-*ri*, Uiryeong-*gun*	0.055

(*continued on next page*)

Appendix C (*continued*)

Rank	Administrative area	Maximum prediction value
4	Gayeon-*ri*, Haman-*gun*	0.047
5	Geomdan-*ri*, Haman-*gun*	0.048
6	Mijo-*ri*, Namhae-*gun*	0.041
7	Hyangyang-*ri*, Jinju-*si*	0.039
8	Doyo-*ri*, Gimhae-*si*	0.038
9	Neukdo-*dong*, Sacheon-*si*	0.036
10	Gyesan-*ri*, Hapcheon-*gun*	0.036
11	Songjeong-*ri*, Namhae-*gun*	0.036
12	Daecheon-*ri*, Jinju-*si*	0.035
13	Sanghyeon-*ri*, hapcheon-*gun*	0.033
14	Jeokgok-*ri*, Uiryeong-*gun*	0.032
15	Baegya-*ri*, Uiryeong-*gun*	0.030
16	Songjin-*ri*, Changnyeong-*gun*	0.030
17	Ugang-*ri*, Changnyeong-*gun*	0.029
18	Songrim-*ri*, Hapcheon-*gun*	0.029
19	Seongsan-*ri*, Uiryeong-*gun*	0.029
20	Hyojeong-*ri*, Changnyeong-*gun*	0.029

References

[1] J.-I. Kim, S.-K. Kang, J.-H. Kwon, The spatial characteristics of transit-poors in urban areas, Journal of the Korean Association of Geographic Information Studies 11 (No. 2) (2008) 1–12.

[2] Z. Sultana, S. Mishra, C.R. Cherry, M.M. Golias, S.T. Jeffers, Modeling frequency of rural demand response transit trips, Transport. Res. Pol. Pract. 118 (2018) 494–505.

[3] F. Hodgson, J. Turner, Participation not consumption: the need for new participatory practices to address transport and social exclusion, Transport Pol. 10 (No. 4) (2003) 265–272.

[4] R. Witter, Public urban transport, travel behaviour and social exclusion–the case of Santiago de Chile, in: XII World Conference on Transportation Research, July), Lisbon, 2010.

[5] L. Cheng, F. Caset, J. De Vos, B. Derudder, F. Witlox, Investigating Walking Accessibility to Recreational Amenities for Elderly People in Nanjing, China, vol. 76, Transportation research part D: transport and environment, 2019, pp. 85–99.

[6] S. Kenyon, G. Lyons, J. Rafferty, Transport and social exclusion: investigating the possibility of promoting inclusion through virtual mobility, J. Transport Geogr. 10 (No. 3) (2002) 207–219.

[7] S. Zhang, Y. Yang, F. Zhen, T. Lobsang, Z. Li, Understanding the travel behaviors and activity patterns of the vulnerable population using smart card data: an activity space-based approach, J. Transport Geogr. 90 (2021), 102938.

[8] H.M. Badland, J.N. Rachele, R. Roberts, B. Giles-Corti, Creating and applying public transport indicators to test pathways of behaviours and health through an urban transport framework, J. Transport Health 4 (2017) 208–215.

[9] G. Higgs, R. Zahnow, J. Corcoran, M. Langford, R. Fry, Modelling spatial access to general practitioner surgeries: does public transport availability matter? J. Transport Health 6 (2017) 143–154.

[10] J.-H. Son, D.-G. Kim, E. Lee, H. Choi, Investigating the spatiotemporal imbalance of accessibility to demand responsive transit (drt) service for people with disabilities: explanatory case study in South Korea, J. Adv. Transport. 2022 (2022) 1–9.

[11] S. Ma, H.S. Kim, Accessibility to welfare facilities for the aged through GIS network analysis: focused on inland areas in Incheon, The Korea spatial planning review 70 (2011) 61–75.

[12] Q. Ma, Y. Xin, H. Yang, K. Xie, Connecting metros with shared electric scooters: comparisons with shared bikes and taxis, Transport. Res. Transport Environ. 109 (2022), 103376.

[13] S.N. Sadeek, A.J.M.M.U. Ahmed, M. Hossain, S. Hanaoka, Effect of land use on crime considering exposure and accessibility, Habitat Int. 89 (2019), 102003.

[14] X. Yao, Where are public transit needed–Examining potential demand for public transit for commuting trips, Comput. Environ. Urban Syst. 31 (No. 5) (2007) 535–550.

[15] C. Xue, Y. Ju, S. Li, Q. Zhou, Research on the sustainable development of urban housing price based on transport accessibility: a case study of Xi'an, China, Sustainability 12 (No. 4) (2020) 1497.

[16] Y.J. Yi, E.J. Kim, The effects of accessibility to medical facilities and public transportation on perceived health of urban and rural elderly: using generalized ordered logic model, J Korean Reg Dev Assoc 27 (No. 1) (2015) 65–88.

[17] Y.-S. Lee, H. Jung, H.J. Yoo, K.-M. Kim, Urban characteristics affecting obesity of elementary, middle and high school students, Journal of the Korean Regional Science Association 31 (No. 3) (2015) 113–130.

[18] A. Rattan, A. Campese, C. Eden, Modeling walkability, Arc. User. Winter (2012) 30–33, 2012.

[19] L.-Y. Chang, W.-C. Chen, Data mining of tree-based models to analyze freeway accident frequency, J. Saf. Res. 36 (No. 4) (2005) 365–375.

[20] M.S. Tehrany, B. Pradhan, S. Mansor, N. Ahmad, Flood susceptibility assessment using GIS-based support vector machine model with different kernel types, Catena 125 (2015) 91–101.

[21] B. Pradhan, 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. 51 (2013) 350–365.

[22] O. Rahmati, D.D. Moghaddam, V. Moosavi, Z. Kalantari, M. Samadi, S. Lee, D. Tien Bui, An automated python language-based tool for creating absence samples in groundwater potential mapping, Rem. Sens. 11 (No. 11) (2019) 1375.

[23] J. Chen, J. Ni, C. Xi, S. Li, J. Wang, Determining intra-urban spatial accessibility disparities in multimodal public transport networks, J. Transport Geogr. 65 (2017) 123–133.

[24] S. Lee, S.O. Son, J. Park, J. Park, Ensemble-based methodology to identify optimal personal mobility service areas using public data, KSCE J. Civ. Eng. 26 (7) (2022) 3150–3159.

[25] Y. Cong, J. Wang, X. Li, Traffic flow forecasting by a least squares support vector machine with a fruit fly optimization algorithm, Procedia Eng. 137 (2016) 59–68.

[26] T. Stadler, A. Sarkar, J. Dünnweber, Bus demand forecasting for rural areas using XGBoost and random Forest algorithm, in: Computer Information Systems and Industrial Management: 20th International Conference, CISIM 2021, Elk, Poland, September 24–26 vol. 20, Springer International Publishing, 2021, pp. 442–453, 2021, Proceedings.

[27] M.F. Khan, S. Asghar, M.I. Tamimi, M.A. Noor, Multi-objective transport system based on regression analysis and genetic algorithm using transport data, IEEE Access 7 (2019) 81121–81131.
[28] S. Imhof, K. Blättler, Assessing spatial characteristics to predict DRT demand in rural Switzerland, Res. Transport. Econ. 99 (2023), 101301.
[29] J.D. Caicedo, M.C. González, J.L. Walker, Public Transit Demand Prediction during Highly Dynamic Conditions: A Meta-Analysis of State-Of-The-Art Models and Open-Source Benchmarking Infrastructure, 2023 arXiv preprint arXiv:2306.06194.
[30] X. Ma, Y. Yin, Y. Jin, M. He, M. Zhu, Short-term prediction of bike-sharing demand using multi-source data: a spatial-temporal graph attentional LSTM approach, Appl. Sci. 12 (3) (2022) 1161.
[31] L. Yang, S. Fei, H. Jia, J. Qi, L. Wang, X. Hu, Study on the relationship between the spatial distribution of shared bicycle travel demand and urban built environment, Sustainability 15 (18) (2023), 13576.
[32] Y. Cui, W. Lv, Q. Wang, B. Du, Usage demand forecast and quantity recommendation for urban shared bicycles, , October, in: 2018 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), IEEE, 2018, pp. 238–2388.
[33] S.L. Handy, D.A. Niemeier, Measuring accessibility: an exploration of issues and alternatives, Environ. Plann. 29 (No. 7) (1997) 1175–1194.
[34] K.T. Geurs, B. Van Wee, Accessibility evaluation of land-use and transport strategies: review and research directions, J. Transport Geogr. 12 (No. 2) (2004) 127–140.
[35] T.L. Lei, R.L. Church, Mapping transit-based access: integrating GIS, routes and schedules, Int. J. Geogr. Inf. Sci. 24 (No. 2) (2010) 283–304.
[36] L. Breiman, Bagging predictors Machine Learning 24 (2) (1996) 123–140, 10.1023. *A: 1018054314350*, 1996.
[37] L. Breiman, J. Friedman, R. Olshen, C. Stone, Classification and regression trees, in: Wadsworth. Inc., Chapman & Hall, Monterey, Calif.In, USA, 1984. Belmont, CA.
[38] W.S. Noble, What is a support vector machine? Nat. Biotechnol. 24 (No. 12) (2006) 1565–1567.
[39] O.S. Azeez, B. Pradhan, H.Z. Shafri, Vehicular CO emission prediction using support vector regression model and GIS, Sustainability 10 (No. 10) (2018) 3434.
[40] J.H. Friedman, Estimating Functions of Mixed Ordinal and Categorical Variables Using Adaptive Splines, Department of Statistics, Stanford Univ., 1991.
[41] J. Ma, S. Jiang, Z. Liu, Z. Ren, D. Lei, C. Tan, H. Guo, Machine learning models for slope stability classification of circular mode failure: an updated database and automated machine learning (AutoML) approach, Sensors 22 (No. 23) (2022) 9166.
[42] R. Abdulhammed, H. Musafer, A. Alessa, M. Faezipour, A. Abuzneid, Features dimensionality reduction approaches for machine learning based network intrusion detection, Electronics 8 (3) (2019) 322.
[43] S. Leem, J. Oh, J. Moon, M. Kim, S. Rho, Enhancing multistep-ahead bike-sharing demand prediction with a two-stage online learning-based time-series model: insight from Seoul, J. Supercomput. (2023) 1–34.
[44] S. Sakri, Assessment of deep neural network and gradient boosting machines for credit risk prediction accuracy, in: 2022 14th International Conference on Computational Intelligence and Communication Networks (CICN), IEEE, 2022, December, pp. 1–7.
[45] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, No. 1, Oakland, CA, USA, 1967, pp. 281–297.
[46] M.A. Foda, A.O. Osman, Using GIS for measuring transit stop accessibility considering actual pedestrian road network, Journal of Public Transportation 13 (No. 4) (2010) 23–40.
[47] J. Mageean, J.D. Nelson, The evaluation of demand responsive transport services in Europe, J. Transport Geogr. 11 (No. 4) (2003) 255–270.
[48] A.M. Bento, M.L. Cropper, A.M. Mobarak, K. Vinha, The effects of urban spatial structure on travel demand in the United States, Rev. Econ. Stat. 87 (No. 3) (2005) 466–478.
[49] D. Brownstone, K.A. Small, Valuing time and reliability: assessing the evidence from road pricing demonstrations, Transport. Res. Pol. Pract. 39 (No. 4) (2005) 279–293.
[50] J.D. Harford, Congestion, pollution, and benefit-to-cost ratios of US public transit systems, Transport. Res. Transport Environ. 11 (No. 1) (2006) 45–58.
[51] D. Sun, Q. Gu, H. Wen, J. Xu, Y. Zhang, S. Shi, M. Xue, X. Zhou, Assessment of landslide susceptibility along mountain highways based on different machine learning algorithms and mapping units by hybrid factors screening and sample optimization, Gondwana Res. (2022).
[52] W. Liu, Q. Tan, W. Wu, Forecast and early warning of regional bus passenger flow based on machine learning, Math. Probl Eng. 2020 (2020) 1–11.
[53] I. Sanaullah, N. Alsaleh, S. Djavadian, B. Farooq, Spatio-temporal analysis of on-demand transit: a case study of Belleville, Canada, Transport. Res. Pol. Pract. 145 (2021) 284–301.
[54] R. Mounce, M. Beecroft, J.D. Nelson, On the role of frameworks and smart mobility in addressing the rural mobility problem, Res. Transport. Econ. 83 (2020), 100956.
[55] C. Weckström, M.N. Mladenović, W. Ullah, J.D. Nelson, M. Givoni, S. Bussman, User perspectives on emerging mobility services: ex post analysis of Kutsuplus pilot, Research in transportation business & management 27 (2018) 84–97.
[56] F. Zwick, K.W. Axhausen, Ride-pooling demand prediction: a spatiotemporal assessment in Germany, J. Transport Geogr. 100 (2022), 103307.