Traffic & Transportation

Integrating Spatial Heterogeneity and Nonlinear Dynamics for Interpretable Dockless Bike-Sharing Demand Prediction

Fan YANG[1], Guangji XU[2]

Original Scientific Paper
Submitted: 14 Feb 2025
Accepted: 3 Aug 2025
Published: 28 Apr 2026

[1] Corresponding author, fanyang@njit.edu.cn, School of Communication and Artificial Intelligence, School of Integrated Circuits, Nanjing Institute of Technology, Nanjing, China
[2] guangji_xu@seu.edu.cn, School of Transportation, Southeast University, Nanjing, China

Publisher:
Faculty of Transport and Traffic Sciences, University of Zagreb

ABSTRACT
Dockless bicycle-sharing systems (DBSS) have emerged as critical solutions for first/last-mile connectivity in urban transportation networks. While demand forecasting is fundamental to optimising system deployment and improving service efficiency, existing models often fail to account for the spatial heterogeneity of urban functional zones. In this study, we utilise social network-based check-in data to characterise the functional composition of urban areas, offering a dynamic and data-driven perspective on interpreting travel demand patterns. This study proposes a methodological framework integrating demographic, land-use and built environment variables to model DBSS demand patterns. The geographic detector method is employed to quantify the contributions of multidimensional factors, enhancing variable selection efficiency and improving model robustness. To address spatial heterogeneity and nonlinear dynamics, a high-precision hybrid model, GWRBoost, is proposed by combining the spatial heterogeneity-capturing capability of geographically weighted regression (GWR) with the nonlinear learning strength of extreme gradient boosting (XGBoost). Benchmark testing against conventional approaches demonstrates enhanced predictive accuracy and improved spatial explicability. The findings provide practical guidance for urban transportation planning and the optimisation of shared mobility systems.

KEYWORDS
dockless shared bicycle demand prediction; built environment; geographic detector; POI data analysis; spatial heterogeneity; hybrid modelling techniques.

1. INTRODUCTION

Dockless bicycle-sharing systems (DBSS) have become integral to sustainable urban transportation networks by addressing the critical first/last-mile connectivity challenge between transit hubs and destinations. Precise demand prediction forms the foundation for supporting fine-grained traffic planning and more adaptive resource allocation strategies in rapidly evolving urban environments.

Urban morphology, particularly land-use configurations and built environment characteristics, exerts substantial influence on DBSS demand patterns. Functional zoning, including residential, commercial, office and transit-oriented areas, significantly affects both temporal demand variations and spatial distribution. Recent methodological advancements demonstrate the efficacy of point-of-interest (POI) data from location-based services in quantifying urban spatial functions [1]. POI datasets enable granular spatial characterisation of urban activities, while location check-in statistics provide complementary metrics of human mobility intensity – a key determinant of bicycle-sharing demand. The multidimensional nature of cycling behaviour, shaped by demographic distributions and heterogeneous urban spatial configurations, necessitates advanced geospatial analytical approaches to disentangle spatial autocorrelation effects and identify fundamental demand determinants.

Current modelling paradigms for bicycle-sharing demand analysis present methodological trade-offs. While conventional linear regression models fail to capture complex nonlinear interactions among predictors adequately, machine learning approaches like gradient boosting decision trees (GBDT) [2, 3] demonstrate superior predictive capability but typically disregard spatial dependencies. Spatial regression techniques such as geographically weighted regression (GWR) [4, 5] address spatial heterogeneity effectively yet remain constrained in handling high-dimensional nonlinear relationships. This methodological dichotomy underscores the need for hybrid frameworks that synergistically integrate spatial econometric principles with machine learning's pattern recognition capabilities.

This study aims to estimate shared bicycle demand by integrating population, land use and built environment factors into a robust and interpretable framework that captures both spatial and nonlinear dynamics. To address key challenges, the study focuses on the following objectives: (1) employing the geographic detector method to identify and quantify the contributions of key influencing factors, thereby improving the efficiency of variable selection; (2) developing a high-precision hybrid model that combines the spatial heterogeneity-capturing capabilities of spatial regression models with the nonlinear learning strengths of machine learning techniques; and (3) evaluating the feasibility and advantages of the proposed framework through comparative analysis with existing methods, offering valuable insights into its applicability for urban transportation planning.

2. PREVIOUS STUDIES

The literature review is organised into four sections: Dockless shared bicycle demand prediction, Geographic detector, Geographically weighted regression and Gradient boosting decision trees. This structure is designed to comprehensively address the theoretical foundations and methodological advancements relevant to the study.

2.1 Dockless shared bicycles demand prediction

Identifying the factors influencing dockless shared bicycles demand is a critical topic in urban transportation planning. Numerous studies have highlighted the significant influence of land use characteristics on shared bicycle usage. For instance, a study in Budapest revealed that tourist attractions, educational facilities and public transport hubs significantly attract shared bicycle users, while residential and industrial zones exhibit a cyclical variation in demand distribution [6]. Similarly, research in Montreal using a spatially varying coefficient regression model demonstrated that land use and transportation infrastructure have varying local effects on shared bicycles demand, emphasising the sensitivity of specific regions to such factors [7]. Further analysis in Toronto indicated that road network density, cycling infrastructure and weather conditions substantially impact shared bicycle ridership, with cycling infrastructure playing a critical role in increasing demand [8].

Comprehensive reviews have also synthesised the connections between land use, transportation systems and urban design, providing a theoretical framework for future studies [9]. Moreover, non-linear relationships between land use and shared bicycles demand have been explored. A study in Chengdu identified population density and employment density as the most influential factors, with additional threshold effects from land use and accessibility [10]. Using deep learning methods, research in San Francisco quantified land use characteristics and their relationship with cycling behaviour, demonstrating the effectiveness of such approaches in identifying regional patterns [11]. A study in Beijing utilised data envelopment analysis (DEA) to assess the alignment of demand and land use, identifying redundancies in financial and living facilities [12]. Additionally, multiscale geographically weighted regression (MGWR) models have been employed to analyse the multi-layered impacts of land use intensity on shared bicycles demand, deepening the understanding of its spatial distribution [13].

Deep learning methods have been employed for short-term shared bicycle demand prediction. Residual neural networks and graph-based models improved short-term predictions, while convolutional LSTM models addressed spatial and temporal correlations [14-16]. Hybrid models further enhanced accuracy, with CNN-LSTM and irregular convolutional LSTM models excelling in regions with variable usage [17, 18]. Multi-view spatiotemporal networks provided grid-level demand insights [19]. Advanced graph-based methods, such as spatial-temporal graph attentional LSTM and deep clustering with group-based models, achieved high accuracy and generalisation [20-22]. A hybrid LSTM-sparse auto-encoder framework demonstrated robust performance with large datasets [23].

2.2 Geographical detector (geodetector)

The geographical detector (geodetector) has emerged as a prominent spatial analysis tool for examining spatial differentiation and identifying driving mechanisms across urban processes. Its core principle involves evaluating the spatial consistency between dependent variables (e.g. land use, transportation activities) and explanatory factors, where stronger consistency indicates higher explanatory power.

Applications of geodetector in urban development include its use in Zhengzhou to detect spatial differentiation and validate land resource allocation assessments [24]. Similarly, a study in Guangzhou on urban occupation mixtures found that land use diversity and cultural inclusiveness significantly shape spatial patterns, optimising indicators to better quantify the interactive effects of driving factors and advancing insights into urban dynamics [25].

In transportation research, geodetector has been applied to transit-oriented development (TOD), such as analysing metro ridership patterns by examining interactive effects of built environment factors like public open spaces and transport systems. This approach outperformed single-factor analyses in explaining traffic distribution variability [26]. Multi-scale analyses further revealed scale-dependent factors driving urban spatial changes, offering deeper insights into the integration of transportation dynamics and urban planning [27].

2.3 Geographically weighted regression (GWR)

Geographically weighted regression (GWR) has become a widely used tool for analysing spatial heterogeneity, offering insights into spatially varying relationships that traditional regression models often overlook. By enabling local regression with geographically varying coefficients, GWR effectively captures localised effects of explanatory variables, making it particularly relevant for transportation studies where spatial heterogeneity is essential.

In shared mobility, GWR has been applied to forecast shared bicycle demand, identifying localised factors that influence ridership. For example, multiscale GWR (MGWR) revealed significant spatially variable effects of non-motorised transport facilities and intersection density on shared bicycles demand, emphasising the importance of localised infrastructure in promoting cycling [13]. This fine-grained analysis informs targeted policy interventions based on urban form and transport infrastructure. Beyond shared bicycles, studies in Vermont identified spatially varying relationships between active mode trip volumes and built environment attributes such as building density and transit accessibility, underscoring the need for tailored urban planning strategies [28]. GWR's utility extends to flow-based transportation models, as demonstrated in the UK, where it captured spatial variability in rail demand elasticities, providing insights for integrating GWR into comprehensive forecasting frameworks [29]. It has also been employed to estimate paratransit service demand in the US, uncovering complex spatial effects linked to demographic and socio-economic variables, showcasing its superiority over traditional regression models [30].

2.4 Gradient boosting decision trees (GBDT)

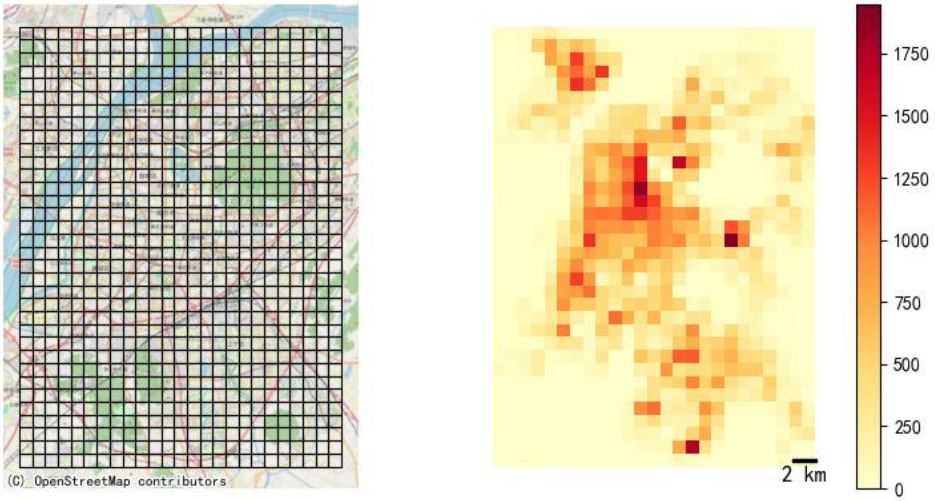
Gradient boosting decision trees (GBDT) and its derivatives, such as LightGBM and XGBoost, are powerful machine learning methods for modelling non-linear and complex relationships in transportation demand forecasting. These methods build sequential ensembles of decision trees, where each tree reduces the residual errors of the previous one, enhancing predictive accuracy and robustness. Their strengths include handling large-scale data, accommodating mixed data types and offering interpretable results through feature importance analysis.

In bike-sharing systems, GBDT-based models have demonstrated high effectiveness in demand forecasting. For example, a study integrating GBDT with LightGBM analysed shared bicycles data from Washington, D.C., achieving superior performance metrics like mean absolute error (MAE), root mean square error (RMSE) and R^2, while effectively capturing temporal and spatial ridership variations [31]. Another study utilised GBDT with nighttime light data to forecast shared bicycles demand, reducing mean squared error (MSE) and mean absolute percentage error (MAPE) without relying on traditional land use data, showcasing adaptability to unconventional datasets [32]. Beyond shared bicycles, GBDT has been applied to various transportation forecasting tasks. In automotive sales forecasting, GBDT outperformed traditional models by capturing non-linear and non-stationary relationships, improving predictions across multiple brands [33]. In supply chain demand forecasting, GBDT effectively models intricate demand fluctuations, enhancing predictive accuracy and financial optimisation [34].

3. DATA

This study integrates diverse data sources to analyse the determinants of shared bicycles demand, incorporating bike usage patterns, demographic characteristics, POI attributes and spatial features. The study area is an urban region in Nanjing, China, bounded by longitude 118.66–118.92 and latitude 31.87–32.17. As shown in *Figure 1(a)*, the region is systematically partitioned into 837 grid cells, each measuring 1 km × 1 km. This grid-based framework facilitates the consistent aggregation of variables such as population, POI activities and spatial features, providing a robust foundation for spatial modelling. Each grid cell serves as a localised analytical unit, capturing spatial variations in bike demand and the associated influencing factors.

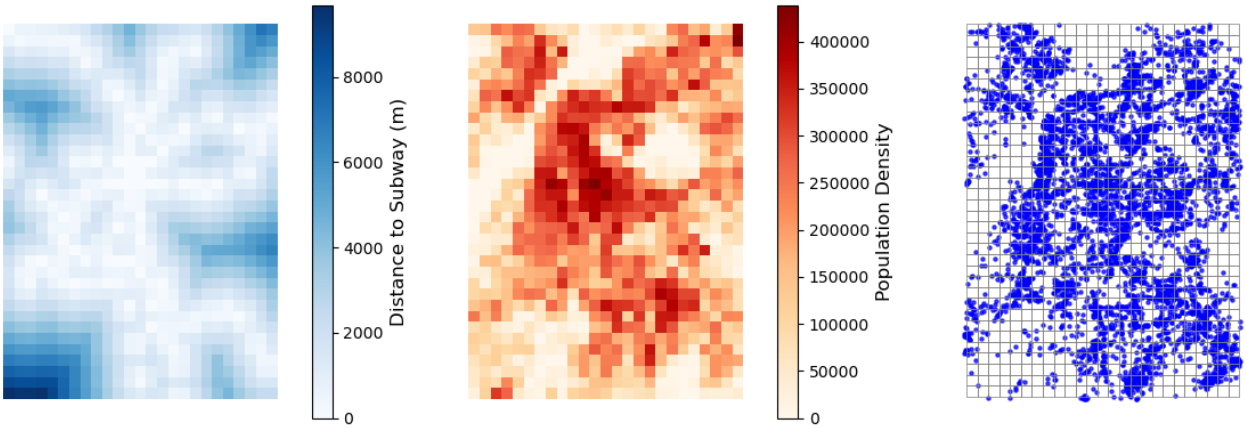
Figure 1(b) presents the dependent variable, shared bike demand, which represents the daily number of shared bike trips originating within each grid cell. These data were derived from Mobike usage records. The heatmap highlights significant spatial variation. High-intensity zones (red areas) are concentrated in the central urban core and along major transport corridors, likely due to high population density, diverse land use and strong transit connectivity. In contrast, low-intensity zones (yellow and white areas) are located in peripheral regions, possibly reflecting lower demand and greater reliance on private vehicles.



(a) *(b)*

Figure 1 – Research area and daily shared bike usage: a) Research area (OSM background); b) Daily bike usage intensity

To investigate the determinants of shared bike demand, 11 explanatory variables were selected to reflect the social, economic and geographic diversity of the study area. *Figure 2* presents the spatial distributions of the representative variables.



(a) *(b)* *(c)*

Figure 2 – Spatial characteristics of the study area: a) Distance to subway; b) Population density; c) POI distribution

These 11 variables can be classified into three categories: spatial characteristics, demographic attributes and POI-related features.

1) Spatial characteristics: The distance to the nearest subway entrance serves as a critical measure of public transit accessibility, derived from OpenStreetMap data. For grid cells containing subway entrances, the distance is zero. This variable plays a key role in modelling first- and last-mile bike usage, which is shown in *Figure 2(a)*. Darker blue areas indicate greater distances, primarily observed in peripheral regions, while lighter areas closer to the urban core represent shorter distances to subway access points. This gradient highlights the potential role of "last-mile" connectivity in influencing bike demand near transit hubs.
2) Population density: Population density data, extracted from the Global Human Settlement Layer (GHSL), quantifies the number of residents within each grid cell. As shown in *Figure 2(b)*, the heatmap demonstrates the distribution of population density, with darker red tones indicating higher densities. These densely populated zones are primarily located in the central urban core and closely align with regions exhibiting high levels of shared bike usage. In contrast, peripheral areas with lighter tones reflect lower population density and correspond to reduced bike demand. Population density thus serves not only as a key proxy for overall transport demand, but also as a reliable spatial indicator for identifying areas with concentrated or latent demand for bicycle-sharing services.
3) POI data: POI data are derived from Weibo check-in records, capturing activity patterns across various land-use types. As shown in *Figure 2(c)*, the scatter plot shows the spatial distribution of points of interest (POIs), represented as blue dots, overlaid on a grid. The clustering of POIs in central areas aligns with high urban activity, including commercial and recreational facilities, while sparse distributions in peripheral areas reflect lower urban activity levels.

To capture the functional heterogeneity of urban space, this study categorises the POI data into nine distinct types, following conventional classifications widely adopted in urban transportation planning activities. These categories include:

— Entertainment (ENTERTAIN): Locations such as cinemas and nightlife venues.
— Food (FOOD): Restaurants, cafes and other dining establishments.
— Work (WORK): Office buildings and business districts.
— General activity (CHECKIN): Aggregated POI activity across all categories.
— Outdoor activities (TOUR): Parks and recreational areas.
— Residential areas (RESIDENCE): Check-ins within residential neighbourhoods.
— Educational facilities (CAMPUS): Schools, universities and other academic institutions.
— Service locations (SHOP): Retail outlets and shopping venues.
— Public transport (TRANSP): Transit hubs, including subway stations.

This classification scheme facilitates the differentiation of urban subregions based on land-use functions and human activity patterns. It also enables the construction of a multi-dimensional feature space that reflects the diversity of activity opportunities within each spatial unit. Such a structure is essential for modelling shared bike demand, as various POI types are associated with different temporal and spatial mobility patterns.

4. METHODOLOGY

This study introduces a novel hybrid modelling framework, termed GWRBoost, as illustrated in *Figure 3*, which integrates geodetector, geographically weighted regression (GWR) and extreme gradient boosting (XGBoost) to analyse shared bicycle demand. The framework is designed to address three core challenges in spatial demand modelling: mitigating overfitting, capturing spatial heterogeneity and modelling nonlinear relationships while preserving interpretability.

The workflow begins with data preprocessing, involving the integration of population data, POI (Point of Interest) data, geographic data and shared bicycle usage data. These datasets undergo geographic coordinate system transformation to ensure spatial consistency, followed by gridding and aggregation to facilitate spatial analysis. Then all variables are normalised to eliminate unit inconsistencies and enhance model convergence.

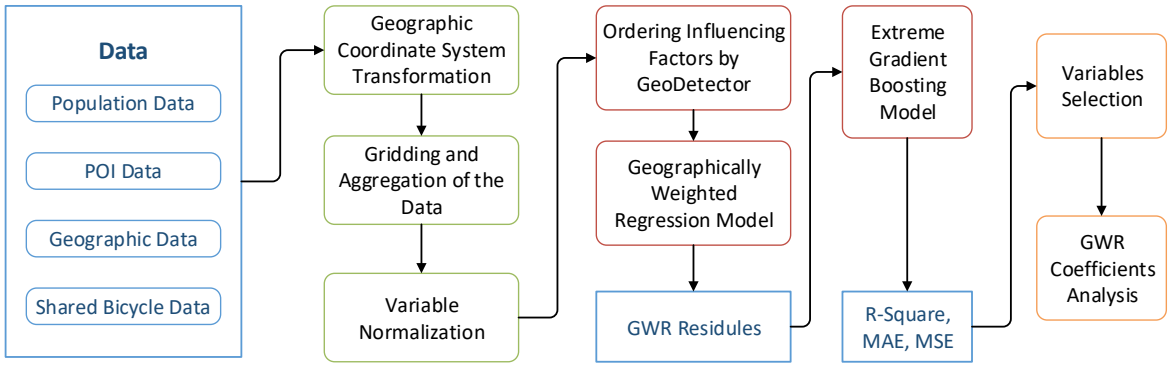


Figure 3 – Modelling framework of GWRBoost

To reduce the risk of overfitting and enhance model robustness, the geodetector method is employed to rank influencing factors based on their contributions to shared bicycle demand. Furthermore, by incrementally testing different subsets of top-ranked variables, the optimal number of input features is determined, balancing model complexity and predictive performance. Subsequently, the GWR model is employed to capture spatial heterogeneity by estimating localised coefficients and generating spatially varying residuals. These residuals are subsequently incorporated into the XGBoost model as supplementary input features, allowing the nonlinear learner to capture complex patterns that GWR may fail to explain. Combined, these components form the GWRBoost framework, which effectively integrates spatial structure and nonlinear learning.

Model performance is evaluated using metrics such as R-square, MAE (mean absolute error) and MSE (mean squared error), allowing for a comparison of the effects of different numbers of influencing factors and an assessment of the proposed framework against traditional methods.

To operationalise the proposed GWRBoost framework, the modelling process is structured into three sequential modules: factor detection with geodetector, modelling spatial heterogeneity with GWR and addressing nonlinear interactions with XGBoost. The following sections provide detailed descriptions of each module and its respective methodological roles within the GWRBoost framework.

4.1 Factor detection with geodetector

Geodetector [35] quantifies the explanatory power of each variable (q-statistic), identifying key factors that significantly influence shared bike demand. This step ensures a data-driven selection of relevant predictors.

The q-statistic is defined as:

$$q = 1 - \frac{\sum_{h=1}^{L} N_h \sigma_h^2}{N\sigma^2} \tag{1}$$

where:

N and σ^2: The total sample size and variance of the dependent variable across the entire study area.

N_h and σ_h^2: The sample size and variance of the dependent variable within subregion h, where $h = 1,2,\ldots,L$, and L is the number of subregions defined by the stratification of the explanatory variable.

The q-value ranges from 0 to 1, where higher values indicate stronger explanatory power. Specifically:

q=1: The explanatory variable perfectly accounts for the spatial variation in the dependent variable.

q=0: The explanatory variable has no relationship with the dependent variable.

Geodetector operates under the assumption that if an explanatory variable significantly influences the dependent variable, the variance within regions stratified by that variable will be smaller than the overall variance. This reduction in variance reflects the spatial consistency of the variable's influence.

By systematically calculating q-values for all candidate variables, this step enables a data-driven selection of predictors. Through incremental evaluation of different subsets of top-ranked variables, the optimal number of input features can be identified, achieving a balance between model complexity and predictive performance.

4.2 Modelling spatial heterogeneity with GWR

Geographically weighted regression (GWR) [36] is employed to capture localised variations in the relationships between bike demand and key factors. This spatially explicit modelling accounts for the inherent heterogeneity in urban environments. The hybrid framework uses GWR residuals to encapsulate localised

spatial effects not fully explained by the original features. These residuals are incorporated into the dataset as additional features, enriching the input for XGBoost. This integration is formalised as follows:

Let $X = \{x_1, x_2, \ldots, x_p\}$ represent the original set of explanatory variables.

GWR generates residuals ϵ_i for each observation i, computed as:

$$\epsilon_i = y_i - \hat{y_i}^{GWR} = y_i - \left[\hat{\beta}_0(u_i, v_i) + \sum_{k=1}^{p} \hat{\beta}_k(u_i, v_i) x_{ik}\right] \quad (2)$$

where:

y_i: Dependent variable at location i (e.g. shared bike demand).

$\hat{\beta}_0(u_i, v_i)$: Intercept term at location (u_i, v_i), capturing location-specific effects.

$\hat{\beta}_k(u_i, v_i)$: Coefficients for the k-th explanatory variable at location (u_i, v_i).

x_{ik}: Value of the k-th explanatory variable at location i.

To estimate localised coefficients, GWR uses a spatial weighting function, which assigns higher weights to observations closer to the focal location. The weight matrix $W(u_i, v_i)$ is defined as:

$$W(u_i, v_i) = \mathrm{diag}(w_{i1}, w_{i2}, \ldots, w_{in}) \quad (3)$$

where:

w_{ij}: Weight assigned to observation j relative to location i, determined by a spatial kernel function (e.g. Gaussian), the Gaussian kernel is defined as: $w_{ij} = \exp(-\frac{d_{ij}^2}{2b^2})$, where d_{ij} is the distance between locations i and j, and b is the bandwidth parameter controlling the influence range.

4.3 Addressing nonlinear interactions with XGBoost

GBDT models complex interactions among variables, providing a robust nonlinear analysis. XGBoost [37] is an improved version of traditional GBDT, incorporating regularisation terms, distributed computing and other optimisations, resulting in higher performance and faster execution. This step enhances predictive accuracy by accommodating intricate variable relationships that linear methods may overlook. XGBoost uses the enriched dataset X^+as input to learn a predictive function $\hat{f}(X^+)$.

The enriched dataset X^+ includes both original features and the residuals:

$$X^+ = \{x_1, x_2, \ldots, x_p, \epsilon\} \quad (4)$$

The final prediction is expressed as:

$$\hat{y_i} = \sum_{t=1}^{T} f_t\,({X_i}^+) \quad (5)$$

where $f_t({X_i}^+)$ is the t-th decision tree in the ensemble and T is the total number of trees. The decision trees $f_t({X_i}^+)$ belong to the function space of regression trees, and the algorithm sequentially learns these trees to optimise the objective function. The objective function in XGBoost consists of two components: a loss function and a regularisation term:

$$L^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y_i}^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i}^{(t)})^2 + \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \quad (6)$$

where:

l: Loss function measuring the difference between the true value y_i and prediction $\hat{y_i}^{(t)}$. A common choice is mean squared error (MSE).

$\Omega(f_k)$: Regularisation term controlling the complexity of the trees, where T is the number of leaf nodes, w_j is the weight of leaf j, and Υ,λ are regularisation parameters.

XGBoost improves predictions iteratively by minimising the loss function using gradient descent. The gradient and Hessian are computed for each observation to guide tree construction:

$$g_i^{(t)} = \frac{\partial l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)}}, h_i^{(t)} = \frac{\partial^2 l(y_i, \hat{y_i}^{(t-1)})}{\partial \hat{y_i}^{(t-1)2}} \quad (7)$$

The decision tree $f_t(X_i^+)$ at iteration t is optimised to fit the pseudo-residuals $g_i^{(t)}$ and weighted by $h_i^{(t)}$, ensuring rapid convergence.

This hybrid approach enhances the model's predictive accuracy while maintaining interpretability, making it well-suited for analysing complex urban mobility patterns. By combining the global predictive power of XGBoost with the localised spatial insights from GWR, this hybrid framework achieves a balanced and robust prediction model. The inclusion of spatially derived features ensures that both global nonlinearities and local spatial dependencies are effectively addressed, yielding superior predictive accuracy and interpretability.

5. RESULTS AND ANALYSIS

This section employs the aforementioned methodology to conduct the shared bicycle demand analysis in the study area in Nanjing. Firstly, the geodetector method was employed to conduct factor detection analysis on the determinants of shared bicycle travel demand, quantifying the explanatory power of each variable on shared bicycle demand. The variables were then ranked based on their q-values. Subsequently, GWR, OLS, XGBoost and GWRBoost methods were applied to rank the influence of variables, progressively reducing the number of factors from 11 to 1. The models' performance was evaluated and compared using R^2, MAE and MSE metrics, with the most suitable model being selected accordingly. Finally, an interpretative analysis of the selected model's results was conducted to gain insights into the underlying factors and spatial dynamics.

5.1 Factor contribution analysis using geodetector

To explore the spatial heterogeneity in shared bike demand, we employed the geodetector method to assess the explanatory power of 11 features. The q-values, representing the proportion of variance in the dependent variable explained by each feature, provide a quantitative measure of their relative importance. A higher q-value indicates a stronger relationship between the feature and shared bike demand. The results are illustrated in *Figure 4*.

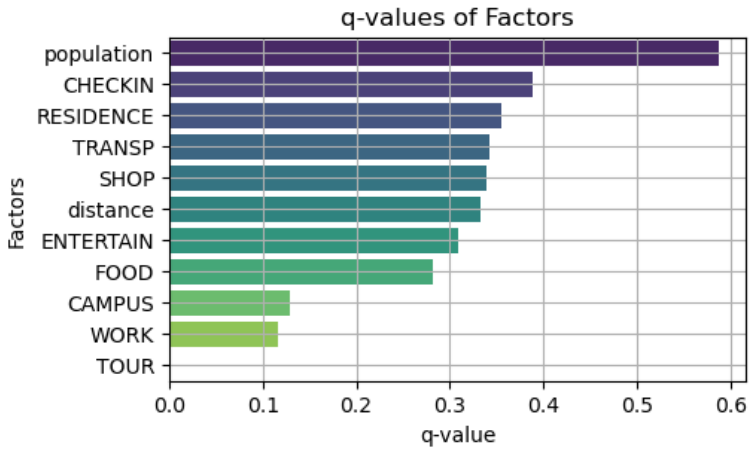


Figure 4 – Result of the factor contribution analysis using geodetector

The feature population demonstrated the highest q-value (0.587), accounting for 58.7% of the spatial heterogeneity in shared bike demand. This result indicates that population density is the primary determinant of bike-sharing usage, likely due to the higher demand for transportation alternatives in densely populated areas. This finding aligns with prior studies that highlight the central role of population in shaping urban mobility patterns.

Other features exhibited moderate q-values, suggesting a substantial but lesser influence on shared bike demand: CHECKIN (0.389) reflects location-based social network user activity, which directly correlates with shared bike demand; RESIDENCE (0.355) represents check-in density within residential neighbourhoods, indicating a potential source of trips, particularly for commuting purposes; TRANSP (0.342) and distance (0.332) denote the accessibility of transportation infrastructure, emphasising the role of seamless integration between bike-sharing systems and public transport. Additionally, SHOP (0.339) showed moderate q-values, implying shared bikes may serve as a convenient mode of transportation for short trips to shopping facilities or nearby locations.

5.2 Performance evaluation among different models

To evaluate the predictive performance of different methods (OLS, GWR, XGBoost and GWRBoost), we compared their R-squared (R^2), mean squared error (MSE) and mean absolute error (MAE) across different variable subsets, ranging from 11 to 1 variables. The data are divided into 80% for the training set and 20% for the test set. The results are listed in *Table 1*, which highlights significant differences among the methods, particularly in handling spatial heterogeneity and nonlinear dynamics.

Table 1 – Comparison across methods and variables

No. of variables	OLS			GWR			XGBoost			GWRboost		
	R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE	R^2	MSE	MAE
11	0.521	57082.4	159.1	0.444	66144.8	144.7	0.513	57940.7	142.6	0.77	27201.8	95.6
10	0.523	56815.3	158.4	0.63	44075.1	132.0	0.52	57184.6	147.8	0.85	18141.4	80.9
9	0.518	57391.6	160.0	0.603	47253.3	135.7	0.539	54901.8	142.1	0.83	20058.3	80.9
8	0.516	57602.9	161.8	0.639	43038.9	133.7	0.528	56201.0	140.4	0.88	13877.4	74.8
7	0.508	58559.9	164.0	0.642	42674.5	131.8	0.49	60720.0	147.2	0.9	11777.2	69.4
6	0.495	60171.3	166.8	0.624	44803.9	134.3	0.43	67894.5	151.6	0.91	10144.9	61.0
5	0.47	63104.7	168.4	0.612	46201.2	132.8	0.48	61885.8	150.1	0.87	14901.9	74.9
4	0.461	64148.1	171.1	0.589	48923.1	133.0	0.513	57950.5	151.9	0.89	12986.7	69.2
3	0.469	63272.8	169.8	0.565	51753.2	139.3	0.391	72544.0	172.6	0.91	11074.9	64.2
2	0.471	62986.9	172.1	0.571	51042.6	138.6	0.351	77282.5	179.1	0.9	12165.0	68.9
1	0.411	70129.0	189.8	0.608	46614.9	135.2	0.155	100645.1	207.9	0.83	20467.3	95.2

As shown in *Table 1*, GWRboost consistently outperformed the other methods, particularly when the number of variables was moderate (e.g. 6 variables). The R-square values for GWRboost were notably higher, while MSE and MAE values were markedly lower, indicating its superiority in both explaining variance and predicting outcomes with high precision. In contrast, OLS exhibited a stable but generally mediocre performance. The R-square values remained relatively low, and both MSE and MAE increased as the number of variables decreased, suggesting increased prediction errors with fewer variables. GWR showed promising results under specific variable conditions (e.g. the highest R-square at 7 variables), but its overall performance fluctuated considerably. Lastly, XGBoost displayed the largest variability in performance, particularly struggling with fewer variables, where R-square values dropped sharply and MSE and MAE values surged.

The comprehensive metrics demonstrate GWRBoost's particular robustness for the given dataset and evaluation criteria, highlighting its advantages in handling complex spatial-nonlinear relationships. These findings strongly support the hybrid approach's effectiveness for spatial predictive modelling tasks.

5.3 Analysis of the GWR coefficients and residuals

Based on the comparative analysis of model application results, the GWRBoost model demonstrated the best performance when incorporating six influencing factors. To provide a clearer understanding of the model parameters and outputs, we present the GWR coefficients and residual maps under the six-factor scenario in *Figure 5*. These visualisations offer valuable insights into the spatial variability of parameter estimates and the model's fit, further highlighting the effectiveness of the proposed approach. The geographically weighted regression (GWR) coefficients provide a detailed spatial insight into the relationships between explanatory variables and the dependent variable (e.g. shared bike demand). By normalising the variables before running the GWR model, the coefficients become directly comparable across the study area, offering a robust foundation for analysing spatial heterogeneity.

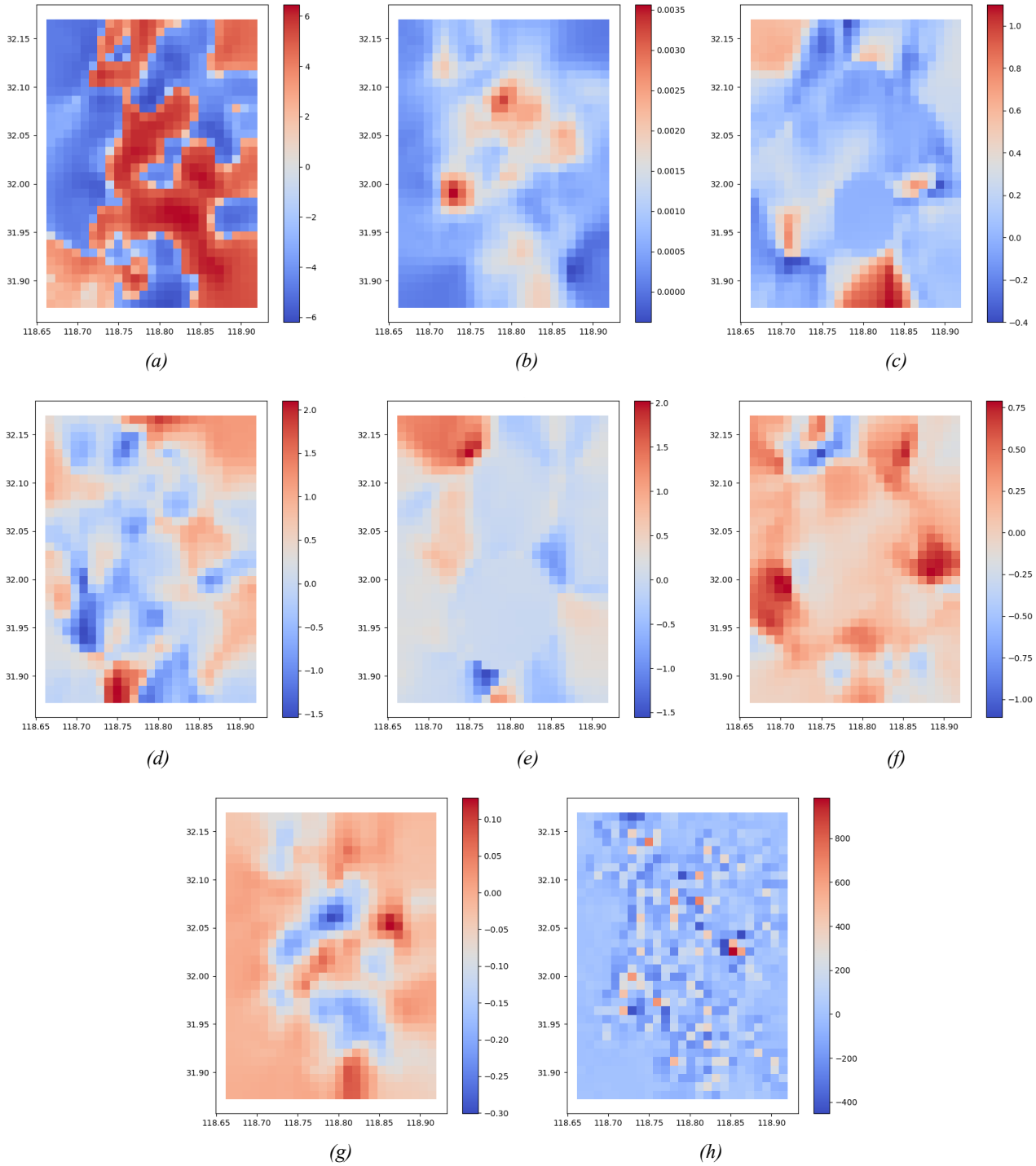


Figure 5 – The GWR coefficients and residuals: a) log(GWR coefficient) for intercept; b) log(GWR coefficient) for population; c) log(GWR coefficient) for CHECKIN; d) log(GWR coefficient) for RESIDENCE; e) log(GWR coefficient) for TRANSP; f) log(GWR coefficient) for SHOP; g) log(GWR coefficient) for distance; h) GWR residuals

The spatial variation in the intercept indicates the baseline level of shared bike demand independent of explanatory variables. Regions with higher intercept coefficients (red zones) are concentrated in the central areas, potentially reflecting high baseline demand due to the inherent characteristics of the urban core, such as accessibility or mixed land use. In contrast, blue zones in peripheral regions suggest lower baseline demand, which may be attributed to lower population density or reduced urban activity.

The GWR coefficients for population density exhibit clear spatial heterogeneity. Positive coefficients (red zones) in high-density areas, as evident in the central urban region, suggest that population density is a significant driver of bike demand, likely due to the increased need for short-distance mobility. Conversely, the negative or near-zero coefficients (blue zones) in peripheral areas indicate that population density has a limited

or inverse influence, potentially reflecting a stronger reliance on private vehicles or public transport in less dense residential zones.

The spatial distribution of the CHECKIN coefficients highlights the importance of POI activities in driving bike demand. Red zones in the central business district and along major transport corridors indicate that high POI activity strongly promotes shared bike usage, aligning with the concentration of commercial, entertainment and recreational facilities. Blue zones, where POI activity has a weaker or negative effect, suggest areas where high POI density may not translate into actual demand, possibly due to mismatches between land use and mobility patterns.

The coefficients for residential POI activity show that certain residential areas exhibit strong positive relationships with bike demand (red zones). These regions are likely characterised by high commuting activity or significant integration with nearby transit facilities. Conversely, negative coefficients (blue zones) are observed in low-demand or static residential areas, where private vehicles dominate short- and medium-distance travel.

The TRANSP variable captures the influence of public transport POIs on bike demand. Regions with high positive coefficients (red zones) are typically located near subway stations or bus hubs, reflecting strong "last-mile" demand for shared bikes. Negative coefficients (blue zones) indicate areas where public transport is less accessible or where alternative transport options reduce the role of shared bikes as a feeder service.

The coefficients for distance to subway stations reveal an expected "last-mile" effect. Negative coefficients (blue zones) in areas closer to subway stations suggest that proximity to transit hubs increases bike demand, particularly for short-distance connectivity. Positive coefficients (red zones) in distant areas may indicate either limited bike demand or alternative transit reliance in regions further away from subway infrastructure.

The residuals map highlights the spatial variation in model fit. Regions with small residuals (blue zones) indicate areas where the GWR model effectively captures the variation in bike demand. In contrast, high residuals (red zones) suggest areas where additional unmodeled factors, such as road conditions, weather or socio-economic variables, may play a significant role.

The GWR results emphasise the spatial heterogeneity of relationships between explanatory variables and shared bike demand. Population density and POI activity emerge as key drivers in urban cores, while public transport accessibility and the last-mile effect dominate near transit hubs. The observed spatial patterns provide actionable insights for optimising bike-sharing systems, such as prioritising infrastructure investments and dynamic bike allocation in high-demand zones while exploring alternative strategies to stimulate demand in low-demand areas. These findings underscore the importance of spatially explicit modelling in urban mobility research and policy planning.

6. CONCLUSIONS

This study proposes an innovative methodological framework for bicycle-sharing demand modelling by integrating geospatial analytics with machine learning techniques. The geographic detector method is employed to quantify the influence of multidimensional factors, providing a novel approach for identifying key variables and improving modelling efficiency. Furthermore, a hybrid model is introduced, which combines the spatial heterogeneity-capturing ability of geographically weighted regression (GWR) with the nonlinear learning capabilities of extreme gradient boosting (XGBoost). This model has shown substantial enhancements in predictive accuracy and interpretability. Such integration marks a methodological advancement in combining the strengths of different modelling approaches to tackle complex spatial problems, offering critical insights for urban transportation planning and aiding policymakers in designing more adaptive and efficient mobility strategies.

The research's practical value can be highlighted in two main aspects:

Initially, this study establishes a predictive model that incorporates location-based social network data with population statistics. This enables precise spatial-temporal mapping between urban functional characteristics and mobility demand patterns. Specifically, the model utilises Point of Interest (POI) data to determine the spatial distribution of high-demand hotspots, such as commercial areas and transit hubs, while population density metrics are used to estimate potential user bases. This multi-source data fusion method offers dynamic and detailed decision-support data for urban planning. It allows for (1) a quantitative evaluation of travel demand impacts under varying land-use planning scenarios and (2) the optimisation of bike-sharing resource allocation through refined deployment, strategic parking facility placement and the identification of service gaps.

Secondly, the proposed geospatial modelling framework shows significant potential for urban planning and transportation management. It is particularly useful for optimising bicycle-sharing systems in newly developed urban areas by enabling accurate demand prediction and data-driven deployment strategies. Moreover, beyond bicycle-sharing systems, the methodology shows remarkable adaptability to a broader range of transportation planning scenarios, including other shared mobility modes such as e-scooters, e-bikes and car-sharing services.

Future research could further enhance the proposed framework for shared bicycle demand estimation by incorporating dynamic factors such as weather, seasonal variations and real-time traffic data to estimate dynamic demand. Additionally, the framework's applicability to other cities and transportation modes warrants further exploration. Validating and expanding the framework across diverse urban contexts with more dynamic datasets could improve its robustness and generalisability.

ACKNOWLEDGEMENTS

This research was funded by Jiangsu Provincial Department of Education Fund of Philosophy and Social Science (2022SJYB0447), Humanities and Social Sciences Research Project of the Ministry of Education of China (25YJCZH330), Scientific Research Foundation of Nanjing Institute of Technology (YJK202114), the Basic Research Programs of Jiangsu Province (No. BK20232036) and Erasmus + project № 101176797 – "Enhancing Female Leadership in STEM in CHINA – FAITH" – ERASMUS-EDU-2024-CBHE. Funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them.

REFERENCES

[1] Yao Y, et al. Unsupervised land-use change detection using multi-temporal POI embedding. *International Journal of Geographical Information Science*. 2023;37(11):2392-415. DOI: 10.1080/13658816.2023.2257262.

[2] Lin PF, et al. Revealing spatio-temporal patterns and influencing factors of dockless bike sharing demand. *IEEE Access*. 2020;8:66139-49. DOI: 10.1109/ACCESS.2020.2985329.

[3] Wang BQ, et al. The impact of dockless bike-sharing and built environment on ride-sourcing trips. *Journal of Advanced Transportation*. 2024;2024. DOI: 10.1155/2024/2977018.

[4] Li ZT, et al. Exploring the multiscale relationship between the built environment and the metro-oriented dockless bike-sharing usage. *International Journal of Environmental Research and Public Health*. 2022;19(4). DOI: 10.3390/ijerph19042323.

[5] Tang J, et al. The association between travel demand of docked bike-sharing and the built environment: Evidence from seven US cities. *Sustainable Cities and Society*. 2024;106. DOI: 10.1016/j.scs.2024.105325.

[6] Jaber A, Csonka B. How do land use, built environment and transportation facilities affect bike-sharing trip destinations? *Promet - Traffic&Transportation*. 2023;35(1):119-32. DOI: 10.7307/ptt.v35i1.67.

[7] Wang X, et al. Modeling bike-sharing demand using a regression model with spatially varying coefficients. *Journal of Transport Geography*. 2021;93:103059. DOI: 10.1016/j.jtrangeo.2021.103059.

[8] El-Assi W, et al. Effects of built environment and weather on bike sharing demand: A station level analysis of commercial bike sharing in Toronto. *Transportation*. 2017;44(3):589-613. DOI: 10.1007/s11116-015-9669-z.

[9] Guo Y, et al. Bike share usage and the built environment: A review. *Frontiers in Public Health*. 2022;10:2296-565. DOI: 10.3389/fpubh.2022.848169.

[10] Chen E, Ye Z. Identifying the nonlinear relationship between free-floating bike sharing usage and built environment. *Journal of Cleaner Production*. 2021;280:124281. DOI: 10.1016/j.jclepro.2020.124281.

[11] Zhao J, et al. Identification of land-use characteristics using bicycle sharing data: A deep learning approach. *Journal of Transport Geography*. 2020;82:102562. DOI: 10.1016/j.jtrangeo.2019.102562.

[12] Sun C, Lu J. Coupling efficiency between bike-sharing demand and land use: data envelopment analysis. *Proceedings of the Institution of Civil Engineers - Transport*. 2023;177(5):316-25. DOI: 10.1680/jtran.23.00003.

[13] Wang Z, et al. Bike-sharing travel demand forecasting via travel environment-based modeling. *Applied Sciences*. 2024;14(16). DOI: 10.3390/app14166864.

[14] Lee J, Kim J. Evaluation of spatial and temporal performance of deep learning models for travel demand forecasting: Application to bike-sharing demand forecasting. *Journal of Advanced Transportation*. 2022. DOI: 10.1155/2022/5934670.

[15] Wang B, et al. Short-term traffic flow prediction in bike-sharing networks. *Journal of Intelligent Transportation Systems*. 2021;26:461-75. DOI: 10.1080/15472450.2021.1904921.

[16] Yang Y, et al. Using graph structural information about flows to enhance short-term demand prediction in bike-sharing systems. *Computers Environment and Urban Systems*. 2020;83:101521. DOI: 10.1016/j.compenvurbsys.2020.101521.

[17] Dastjerdi AM, Morency C. Bike-sharing demand prediction at community level under COVID-19 using deep learning. *Sensors*. 2022;22. DOI: 10.3390/s22031060.

[18] Li X, et al. Improving short-term bike sharing demand forecast through an irregular convolutional neural network. *Transportation Research Part C: Emerging Technologies*. 2022;147. DOI: 10.1016/j.trc.2022.103984.

[19] Chai J, et al. ST-bikes: Predicting travel-behaviors of sharing-bikes exploiting urban big data. *Ieee Transactions On Intelligent Transportation Systems*. 2023;24:7676-86. DOI: 10.1109/TITS.2022.3197778.

[20] Han L, et al. Short-term traffic prediction based on deepcluster in large-scale road networks. *Ieee Transactions On Vehicular Technology*. 2019;68:12301-13. DOI: 10.1109/TVT.2019.2947080.

[21] Ma X, et al. Short-term prediction of bike-sharing demand using multi-source data: A spatial-temporal graph attentional LSTM approach. *Applied Sciences*. 2022. DOI: 10.3390/app12031161.

[22] Xu C, et al. The station-free sharing bike demand forecasting with a deep learning approach and large-scale datasets. *Transportation Research Part C: Emerging Technologies*. 2018. DOI: 10.1016/J.TRC.2018.07.013.

[23] Chen C, et al. Traffic flow prediction based on deep learning in internet of vehicles. *Ieee Transactions On Intelligent Transportation Systems*. 2021;22:3776-89. DOI: 10.1109/TITS.2020.3025856.

[24] Wang H, et al. Evaluating the suitability of urban development land with a Geodetector. *Ecological Indicators*. 2021;123:107339. DOI: 10.1016/j.ecolind.2021.107339.

[25] Deng X, et al. Spatial distribution and mechanism of urban occupation mixture in Guangzhou: an optimized geodetector-based index to compare individual and interactive effects. *International Journal of Geo-Information*. 2021;10:659. DOI: 10.3390/ijgi10100659.

[26] Deng X, et al. Interactive impacts of built environment factors on metro ridership using geodetector: From the perspective of TOD. *International Journal of Geo-Information*. 2022;11:623. DOI: 10.3390/ijgi11120623.

[27] Yuan S, Yu S. Geodetector-based analysis of spatial variation and drivers in urban scale. 2024;12980:129802D. DOI: 10.1117/12.3020992.

[28] Yang H, et al. Spatial variations in active mode trip volume at intersections: a local analysis utilizing geographically weighted regression. *Journal of Transport Geography*. 2017;64:184-94. DOI: 10.1016/J.JTRANGEO.2017.09.007.

[29] Blainey S, Preston J. Extending geographically weighted regression from points to flows: A rail-based case study. *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*. 2013;227:724-34. DOI: 10.1177/0954409713496987.

[30] Kuo P, et al. Modeling the spatial effects on demand estimation of americans with disabilities act paratransit services. *Transportation Research Record*. 2013;2352:146-54. DOI: 10.3141/2352-17.

[31] Han X, et al. Research on construction and contribution analysis of demand forecasting model based on GBDT-LightGBM algorithm. *2023 IEEE International Conference On Image Processing and Computer Applications (ICIPCA)*. 2023:359-64. DOI: 10.1109/ICIPCA59209.2023.10257726.

[32] Sun C, Lu J. Investigating the potential of nighttime light data to estimate travel demand. *Transactions in GIS*. 2024. DOI: 10.1111/tgis.13240

[33] Zhang Z, Chai H, et al. Automobile-demand forecasting based on trend extrapolation and causality analysis. *Electronics*. 2024. DOI: 10.3390/electronics13163294.

[34] Metawa N. Forecasting business demand to enhance supply chain financial optimization: A predictive modeling approach. *American Journal of Business and Operations Research*. 2019. DOI: 10.54216/ajbor.000201.

[35] Wen QQ, et al. Sources and risk characteristics of heavy metals in plateau soils predicted by geo-detectors. *Remote Sensing*. 2023;15(6). DOI: 10.3390/rs15061588.

[36] Wu DC. Spatially and temporally varying relationships between ecological footprint and influencing factors in China's provinces using geographically weighted regression (GWR). *Journal of Cleaner Production*. 2020;261. DOI: 10.1016/j.jclepro.2020.121089.

[37] Su W, et al. An XGBoost-based knowledge tracing model. *International Journal of Computational Intelligence Systems*. 2023;16(1). DOI: 10.1007/s44196-023-00192-y.