Energy Proceedings
Vol 50, 2025
ISSN 2004-2965

Estimation of urban land use implication on energy-related carbon emissions based on machine learning methods

Ping He[1,2], Helen Xiaohui Bao[2*], Geoffrey Qiping Shen[1*]

1 Department of Building and Real Estate, The Hong Kong Polytechnic University, Hong Kong SAR, China

2 Department of Land Economy, University of Cambridge, Cambridge, United Kingdom
(Corresponding Author: hxb20@cam.ac.uk; geoffrey.shen@polyu.edu.hk)

ABSTRACT
Understanding the impact of urban land use patterns on energy-related carbon emissions is critical for developing effective climate change mitigation strategies. This study employed machine learning techniques to model the relationship between multidimensional urban land use characteristics and city-scale carbon emissions. Urban land use was characterized across four dimensions: scale, structure, mixture, and intensity. Machine learning algorithms, including CART, Random Forest, and XGBoost, were trained to quantify the relative importance of these land use features in predicting carbon emissions. The machine learning models demonstrated strong predictive performance, outperforming traditional linear regression. The feature importance analysis revealed that urban land use indicators collectively account for over one-quarter of the models' predictive power, with land use scale, structure, and intensity features exhibiting greater importance than socioeconomic variables. These findings underscore the value of data-driven, nonparametric modeling approaches in elucidating the complex, multifaceted links between urban form and greenhouse gas emissions.

Keywords: urban land use, carbon emissions, machine learning, sustainable urban planning

1. INTRODUCTION

Cities are major contributors to global energy-related carbon emissions, accounting for over 70% of worldwide emissions [1]. This urban emissions burden has rapidly escalated due to the concurrent trends of accelerated urbanization and economic development, which have driven up energy consumption across the urban built environment [2]. Optimized urban land use planning has emerged as a promising strategy to mitigate these substantial city-based carbon emissions [3]. Urban land use patterns fundamentally shape the energy profiles of transportation, buildings, and socio-economic activities within cities [4-6]. Carefully orchestrated land use planning that promotes energy efficiency, renewable energy integration, and low-carbon urban lifestyles can thus serve as a crucial pathway for emissions reduction. Robust scientific evidence directly linking specific urban land use characteristics to energy-related emissions is needed to inform urban planning policies for climate change mitigation.

Existing approaches to modeling the drivers of urban carbon emissions have faced significant limitations. Quantifying the multifaceted impacts of diverse land use characteristics has proven challenging, given the nonlinear and context-dependent nature of the land use-emissions relationship [7]. Traditional statistical regression methods have struggled to capture the full complexity of urban systems and generate accurate emissions predictions [8]. In response, machine learning techniques have emerged as a promising alternative for urban sustainability analysis [9]. These data-driven, predictive modeling approaches hold the advantage of uncovering hidden patterns and relationships within large, multidimensional urban datasets. Studies have demonstrated improved predictive performance of machine learning models compared to conventional statistical methods in modeling urban energy demand, greenhouse gas emissions, and optimizing planning and infrastructure decisions [10]. However, the existing literature lacks a comprehensive investigation of the implications of urban land use patterns on carbon emissions using advanced modeling techniques. This study aims to address this research gap by employing machine learning to elucidate the land use determinants of urban carbon emissions.

This study would like to explore the following research questions: (1) How to describe urban land use patterns and what are the urban land use patterns in

This is a paper for the 16th International Conference on Applied Energy (ICAE2024), Sep. 1-5, 2024, Niigata, Japan.

China's cities? (2) To what extent the machine learning methods can better predict the implication of urban land use patterns on carbon emissions? (3) What implications does urban land use have on carbon emissions?

To answer the above questions, this study has the following objectives: (1) Analyze the spatial-temporal urban land use patterns in China's cities. (2) Develop machine learning models incorporating urban land use indicators for carbon emission prediction. (3) Identify the feature importances of urban land use indicators in the emission prediction models.

2. MATERIALS AND METHODS

2.1 Selection and calculation of urban land use indicators

The analysis of urban land use patterns in this study examines four key dimensions: scale, intensity, structure, and mixture [6,11,12]. The urban land use scale is represented by the total land area of each city. Land use intensity is indicated by population density. The structural composition of land uses is captured by the proportion of different land use types. The degree of land use mixture is quantified using the entropy index. The entropy index, originally developed to measure the measure economic inequality, has become a widely adopted metric for assessing the land mixed use degree [13]. The urban land use entropy score for each city is calculated using the following formulas:

$$H = \sum_{i=1}^{n} -P_i \, ln \, P_i \tag{1}$$

$$P_i = \frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i} \tag{2}$$

where H represents the entropy score; i means the land use type; n denotes the number of urban land use types in each city; P_i denotes the proportion of urban land use type i in each city; λ_i denotes the area of land use type i in each city.

2.2 Machine learning methods

2.1.1 CART model

The Classification and Regression Trees (CART) approach is a non-parametric solution which can form regression trees. Starting with the full data set, the CART tree is built by iteratively separating subsets of the data set using all predictor variables to build two child nodes. A range of diversity or impurity metrics are used to select the best predictor. Producing data subsets that are as homogenous as feasible with regard to the target variable is the aim. Every predictor in the CART algorithm is assessed for each split in order to determine the optimal cut point based on improvement score or impurity reduction. The predictor with the greatest improvement is then chosen for the split after the predictors are compared.

Detailed mathematical and statistical information for the CART algorithm can be found in previous studies such as Bel et al., (2009) [14].

2.1.2 Random Forest model

As a tree algorithm as well, Random Forest performs better when working with complicated data than other nonlinear methods. Specifically, Random Forest approaches provide the greatest gain above CART methods, which connect inputs to grouped outputs utilizing layers through a tree-like process. Two additional layers of randomness distinguish Random Forest from standard CART analysis: first, a larger set of trees can be obtained through multiple bootstrapping iterations to create a forest; second, each tree in a Random Forest is fitted using a random set of predictors.

Detailed mathematical and statistical information for the Random Forest algorithm can be found in previous studies such as Jaiswal et al., (2017) [15].

2.1.3 XGBoost model

eXtreme Gradient Boosting (XGBoost) is a scalable tree boosting algorithm. Boosting is an ensemble technique that iteratively combines weak learning models to create a more resilient and powerful estimator. To learn and optimize the loss function at each iteration, the residual of the preceding estimator will be employed.

Detailed mathematical and statistical information for the XGBoost algorithm can be found in previous studies such as Santhanam et al., (2017) [16].

2.3 Linear regression model

To better understand the direction of each urban land use indicator's impact on carbon dioxide emissions, the paper constructs an empirical model:

$$y_{it} = \alpha + \lambda_t + \mathbf{x}_{it}\beta + \epsilon_{it} \tag{3}$$

where i represents the city and t represents the year. y represents carbon emissions. λ denotes the time effect. $\mathbf{x}$ denotes the independent variables. β represents the coefficient of all variables. ϵ represents the random error.

2.4 Data sources

The data of 258 cities in China from 2002 to 2019 is used in this study based on availability. The controlled variables in the estimation models involve

representative socio-economic indicators, including population (POP), per capita GDP (PGDP), GDP of the secondary industry (GDP2), industrial structure (IS2), and energy consumption (ENE). They are obtained from China City Statistical Yearbook. The urban land use indicators include land area (ALL), land use mixture degree (MIX), proportion of each land use type area (RES to GRE), and intensity (INT). They are obtained from China City Construction Statistical Yearbook. The carbon emission data are from Carbon Emission Accounts and Datasets (CEADS). The detailed variable description is shown in Table 1. The descriptive statistics of the data are presented in Table 2.

Table 1 Variables description and source.

Abbreviation	Detailed description	Data source
Social economic indicators		
YEAR	Year of the data	China City Statistical Yearbook
POP	Population	
PGDP	Per capita GDP	
GDP2	Secondary industry GDP	
IS2	Proportion of Secondary industry GDP	
ENE	Energy consumption	
Urban land use indicators		
ALL	All land area	China City Construction Statistical Yearbook
MIX	Land use mixture	
RES	Residential land	
PUB	Public service land	
IND	Industrial land	
STR	Storage land	
TRA	Transportation land	
MUN	Municipal administrative land	
GRE	Green land and open space	
Dependent variable		
CE	Carbon emissions	CEADS

Table 2 Data description.

	Mean	Std	Min	Max	Unit
YEAR	2010.99	4.73	2002	2019	/
POP	474.20	330.03	15.97	3416.00	10^4 ppl
PGDP	4.11	3.28	0.01	46.77	10^4 yuan
GDP2	977.03	1232.49	18.12	10298.30	10^8 yuan
IS	48.67	10.26	14.95	89.34	%
ENE	1392.32	1366.67	73.08	21585.19	10^6 tce
RES	31.71	7.07	10.90	77.40	%
PUB	14.30	4.97	0.00	48.71	%
IND	19.83	6.99	0.00	42.97	%
STR	3.37	2.10	0.00	28.52	%
TRA	15.90	4.73	0.00	55.73	%
MUN	3.79	2.56	0.00	24.27	%
GRE	11.10	5.70	0.15	43.76	%
ALL	174.03	235.93	4.99	2915.56	km^2
MIX	1.68	0.10	0.89	1.89	/
INT	4.61	3.93	0.14	43.81	10^4 ppl/km^2
CE	36.81	38.97	1.17	457.76	10^6 tons

3. RESULTS

3.1 Assessment of urban land use patterns

The spatial distribution of ALL and MIX in the years 2002 and 2019 is presented in Fig. 1. The figure shows that the urban land scale expanded during the year 2002 to 2019. The land use mixture degree shows decrease in several specific areas such as east costal area and northwestern area.

3.2 Performance of machine learning models

The overall performance of the machine learning models is given in Fig. 2. The metrics including R-squared, root mean squared error (RMSE), and mean absolute percentage error (MAPE) are used to make comparison of the performance among the models. Through

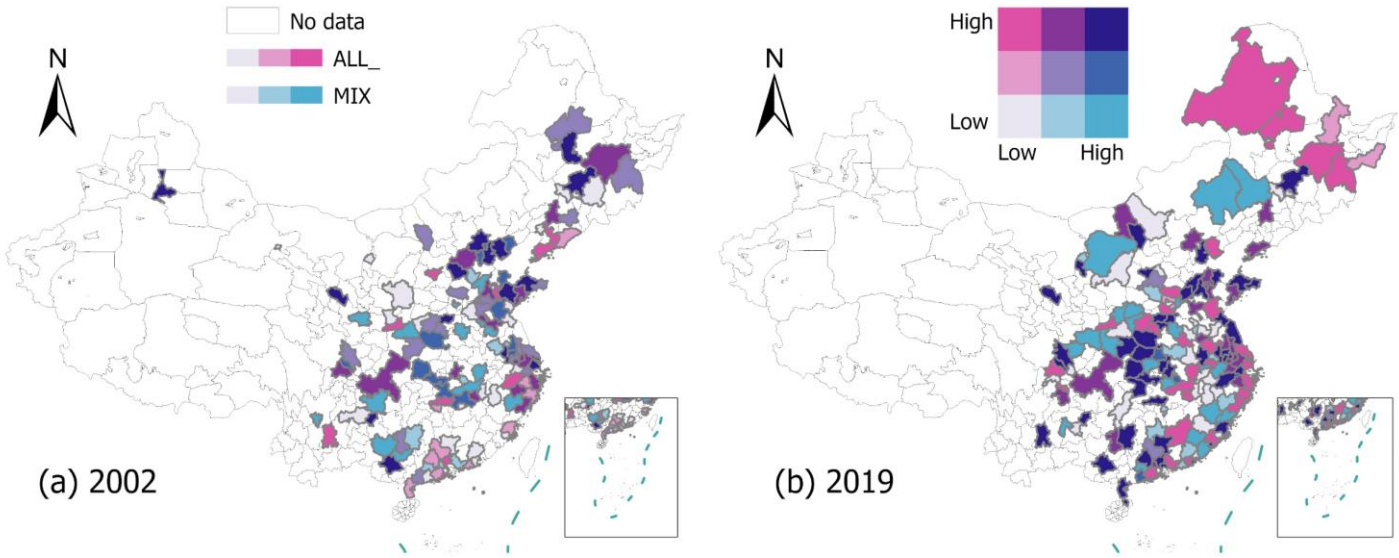

Fig. 1 Spatial distribution of ALL and MIX in 2002 and 2019

comparing the above metrics, the XGBoost model has the best estimation accuracy. This model has the highest coefficient of determination with R^2 = 0.822 and lowest error with RMSE = 18.085. The CART and Random Forest models showed relatively lower goodness of fit and larger errors. The CART model has R^2 = 0.791 and RMSE = 19.596, and the Random Forest model shows R2 = 0.813 and RMSE = 18.541.

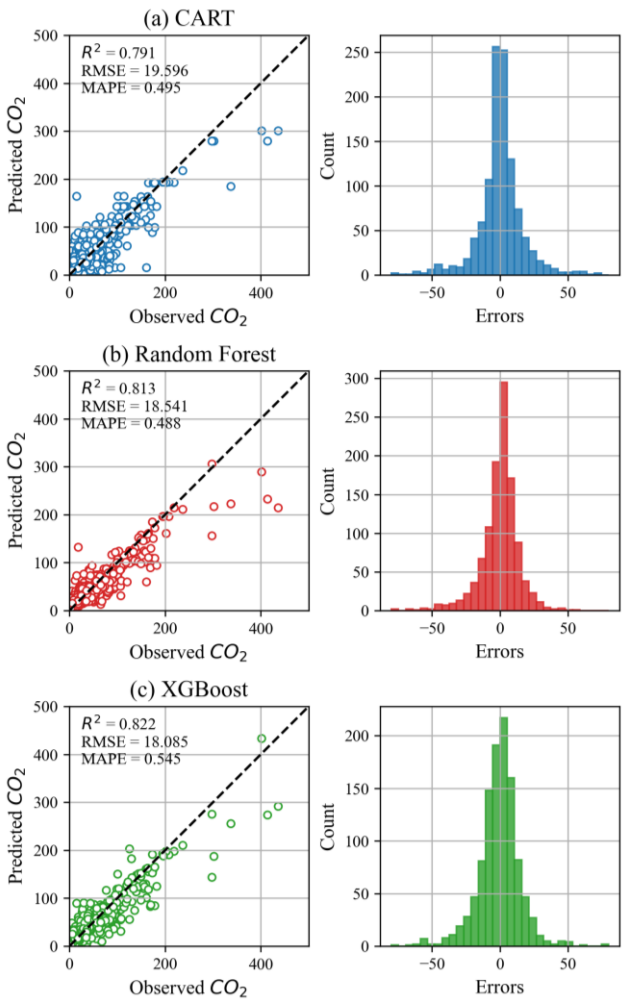

Fig. 2 Fitting results of the machine learning models for CE prediction

3.3 Driving factors of carbon emissions

3.3.1 Machine learning results

The feature importances of the machine learning models for CE prediction is presented in Fig. 3. In general, the importances of land use indicators exceeds 25% in all machine learning models. The estimation results of the CART model show 25.78% feature importances of the land use indicators, Random Forest model 30.68%, and XGBoost model 46.38%. Among the land use indicators, GRE shows the highest importance, with the value of 7.90% in the CART model, 6.45% in the Random Forest model, and 10.32% in the XGBoost model. Besides, IND and STR both have relatively high importance. IND shows 3.18%, 3.38%, and 6.77% in the CART, Random Forest, and XGBoost models respectively. STR shows 3.65%, 3.85%, and 4.37% in the three models respectively.

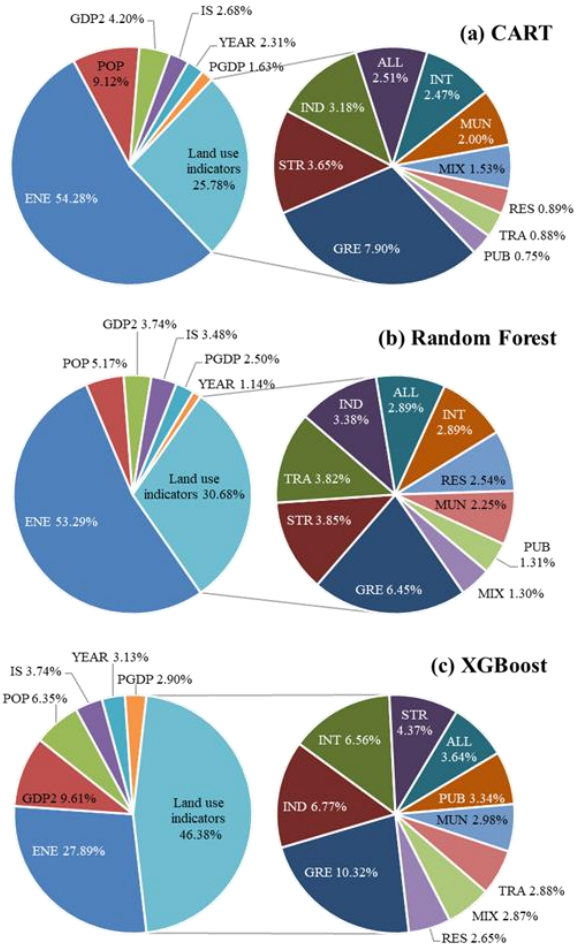

Fig. 3 Feature importances in the machine learning models for CE prediction

3.3.2 Linear regression results

The Ordinary Least Squares (OLS) linear regression results are presented in Table 3. The OLS model achieves an R-squared value of 0.423, which is lower than the performance of the machine learning approaches. Variance Inflation Factor (VIF) analysis indicates low multicollinearity among the predictor variables, as all VIF values are less than 10. The regression coefficients reveal that RES, IND, STR, and TRA land uses exhibit significant negative associations with carbon emissions, while MIX and ALL have significant positive relationships with carbon emissions.

Table 3 Regression analysis results (OLS model).

Variables	Coefficient	Std. Error	t value	p value	VIF
POP	-0.012*	0.007	-1.660	0.098	2.849
PGDP	0.000***	0.000	3.780	0.000	4.334
GDP2	0.000	0.000	1.430	0.154	6.013
IS	-0.022	0.039	-0.550	0.583	1.287
ENE	0.016***	0.001	23.750	0.000	4.031
RES	-0.302***	0.064	-4.720	0.000	2.943
IND	-0.210***	0.059	-3.540	0.000	2.413
STR	-0.506***	0.174	-2.900	0.004	1.227
TRA	-0.157**	0.071	-2.200	0.028	1.869
MUN	0.018	0.116	0.150	0.878	1.404
GRE	0.016	0.067	0.240	0.814	2.220
ALL	0.012**	0.005	2.410	0.016	5.050
MIX	6.964*	3.190	2.180	0.029	1.266
INT	0.151	0.137	1.100	0.270	2.052
Constant	20.603**	8.253	2.500	0.013	

Notes: Adjusted R-squared: 0.423, F-statistic: 180.72;
*** p<0.01, ** p<0.05, * p<0.1.

4. DISCUSSION

4.1 Estimation models

Based on the results, the machine learning models have better performance for estimation than the traditional linear regression model. The R-squared was improved from 0.423 in the traditional linear regression model to 0.816 in the XGBoost model, nearly doubled, indicating the powerful prediction ability of the machine learning approach. Machine learning approaches have the ability to capture the complex, nonlinear relationships between land use and emissions.

Besides, it is important to incorporate diverse land use characteristics into the prediction model. The results show that the urban land use indicators contribute over 1/4 and even nearly 1/2 total feature importances. It implies the contribution of different land use features to model accuracy and the advantages of the incorporated models over simplified models with only socio-economic factors.

4.2 Driving factors of carbon emissions

The urban land use indicators have more than 1/4 feature importances among all variables. Among the indicators, the land use structure and scale have high feature importance in predicting carbon emissions, indicating they are the most influential factors. Combining the results obtained from the traditional linear regression model, the land use scale contributes to the increase of carbon emissions, implying the importance of control of urban sprawl.

Besides, the proportion of green land and open space, industrial land, and storage land contribute most among the land use structure indicators. Based on the linear regression results, the increase of the proportion of industrial land and storage land can reduce carbon emissions. The possible reason might be that larger, centralized industrial and storage facilities may be able to take advantage of economies of scale, allowing for the implementation of more advanced and efficient energy-saving technologies that may not be feasible for smaller, dispersed facilities. Additionally, concentrated industrial and storage areas can enable more efficient logistics and supply chain management. This can reduce the need for unnecessary transportation of goods and materials, leading to a decrease in fuel consumption and lower carbon emissions from the transportation sector.

The impacts of land use mixture on carbon emissions investigated in this study also support the above inference. The higher the mixture degree, the higher carbon emissions. A higher land mixture degree can lead to increased travel distances and fragmented transportation networks, which results in higher fuel consumption and carbon emissions from transportation.

To summarize, from the carbon reduction perspective, curbing urban land expansion, optimizing land use structure and careful planning of mixed land use are helpful in achieving low-carbon urban development.

4.3 Limitation and future research

Due to the inherent characteristics of the complex mechanism inside the machine learning models, it is difficult to intuitively present the association between factors and emissions when using the machine learning approaches. For example, though the feature importance results demonstrate the proportion of the importance of each variable, it cannot tell us the detailed relationship as what the traditional linear regression can tell us. This can cause confusion when making land use planning decisions, e.g., knowing which factors are important but don't know what to do.

However, the machine learning models have powerful advantages in prediction. Given the importance of land use in achieving the climate goal, future research is suggested to apply machine learning techniques in predicting carbon emissions under multiple urban land use scenarios.

Additionally, because of the data limitation, several dimensions of land use patterns have not been considered, such as concentration, centrality, proximity [17].

Moreover, given the huge heterogeneity among cities, the implications might differ from region to region. In-depth investigations fully considering the geographical characteristics of cities are suggested.

5. CONCLUSIONS

This study applies machine learning methods including CART, Random Forest, and XGBoost to estimate the impact of urban land use on energy-related carbon emissions. The major conclusions are as follows:

Firstly, the machine learning-based modeling approach significantly outperformed traditional linear regression, doubling the accuracy of carbon emissions predictions. This demonstrates the importance of utilizing data-driven, nonlinear techniques to capture the complex relationships between land use and emissions.

Secondly, urban land use indicators contributed over 25% of the total feature importance in the emissions prediction models. This underscores the critical role that diverse land use characteristics play in shaping a city's carbon footprint, beyond just socioeconomic factors.

Thirdly, the analysis revealed that both the scale and structure of urban land use have substantial effects on energy-related carbon emissions. Specifically, a higher degree of land use mix was found to lead to increased emissions, highlighting the need to carefully consider the spatial configuration and integration of different land uses within a city.

ACKNOWLEDGEMENT

This work was supported by the PhD Student Stipend and PhD Scholars International Collaborative Research Fellowship from The Hong Kong Polytechnic University.

REFERENCE

[1] Ang, Y. Q.; Berzolla, Z. M.; Letellier-Duchesne, S.; Reinhart, C. F. Carbon Reduction Technology Pathways for Existing Buildings in Eight Cities. Nat Commun 2023, 14 [1], 1689.

[2] Wang, W.-Z.; Liu, L.-C.; Liao, H.; Wei, Y.-M. Impacts of Urbanization on Carbon Emissions: An Empirical Analysis from OECD Countries. Energy Policy 2021, 151, 112171.

[3] Cheng, C.; Yan, X.; Fang, Z.; Zhou, Q.; Tang, Y.; Li, N.; Tang, D. Proposing Carbon Reduction Strategies for Mega-Urban Agglomerations − a Cluster Analysis Based on Carbon Emission Intensity. Ecological Indicators 2024, 166, 112336.

[4] Wegener, M. Land-Use Transport Interaction Models. In Handbook of Regional Science; Fischer, M. M., Nijkamp, P., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2021; pp 229–246.

[5] He, P.; Xue, J.; Shen, G. Q.; Ni, M.; Wang, S.; Wang, H.; Huang, L. The Impact of Neighborhood Layout Heterogeneity on Carbon Emissions in High-Density Urban Areas: A Case Study of New Development Areas in Hong Kong. Energy and Buildings 2023, 287, 113002.

[6] Xia, C.; Yeh, A. G. O.; Zhang, A. Q. Analyzing Spatial Relationships between Urban Land Use Intensity and Urban Vitality at Street Block Level: A Case Study of Five Chinese Megacities. Landscape and Urban Planning 2020, 193.

[7] Liu, H.; Alharthi, M.; Atil, A.; Zafar, M. W.; Khan, I. A Non-Linear Analysis of the Impacts of Natural Resources and Education on Environmental Quality: Green Energy and Its Role in the Future. Resources Policy 2022, 79, 102940.

[8] Jin, Y.; Sharifi, A.; Li, Z.; Chen, S.; Zeng, S.; Zhao, S. Carbon Emission Prediction Models: A Review. Science of The Total Environment 2024, 927, 172319.

[9] Li, F.; Yigitcanlar, T.; Nepal, M.; Nguyen, K.; Dur, F. Machine Learning and Remote Sensing Integration for Leveraging Urban Sustainability: A Review and Framework. Sustainable Cities and Society 2023, 96, 104653.

[10] Abbasabadi, N.; Ashayeri, M. Urban Energy Use Modeling Methods and Tools: A Review and an Outlook. Building and Environment 2019, 161, 106270.

[11] Xu, F.; Wang, Z.; Chi, G.; Zhang, Z. The Impacts of Population and Agglomeration Development on Land Use Intensity: New Evidence behind Urbanization in China. Land Use Policy 2020, 95, 104639.

[12] Lei, W.; Jiao, L.; Xu, G. Understanding the Urban Scaling of Urban Land with an Internal Structure View to Characterize China's Urbanization. Land Use Policy 2022, 112, 105781.

[13] O'Driscoll, C.; Crowley, F.; Doran, J.; McCarthy, N. Land-Use Mixing in Irish Cities: Implications for Sustainable Development. Land Use Policy 2023, 128, 106615.

[14] Bel, L.; Allard, D.; Laurent, J. M.; Cheddadi, R.; Bar-Hen, A. CART Algorithm for Spatial Data: Application to Environmental and Ecological Data. Computational Statistics & Data Analysis 2009, 53 [8], 3082–3093.

[15] Jaiswal, J. K.; Samikannu, R. Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression. In 2017 World Congress on Computing and Communication Technologies [WCCCT]; 2017; pp 65–68.

[16] Santhanam, R.; Uzir, N.; Raman, S.; Banerjee, S. Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets; 2017.

[17] Sarzynski, A.; Galster, G.; Stack, L. Evolving United States Metropolitan Land Use Patterns. Urban Geography 2014, 35 (1), 25–47.