Discovering Nonlinear Urban Heat Drivers Through Bayesian-Optimised Boosted Trees

Xiao Zhang, Tao Wu*

College of Architecture and Urban Planning, Tongji University, 1239 Siping Road, Shanghai, P.R. China

*taowu@tongji.edu.cn

Abstract. Urban thermal environments are shaped by complex interactions between built form and ecological elements, yet many studies still rely on linear assumptions that obscure threshold and stage-dependent behaviours. This study investigates the nonlinear drivers of daytime Land Surface Temperature (LST) in Beijing's central urban area using multi-source geospatial datasets and a four-season modelling strategy. We construct seasonal regression models with LightGBM and tune hyperparameters via Bayesian optimisation. Spatial dependence is assessed using Global Moran's I and partially addressed by including a spatial autocorrelation term. To improve interpretability of the ensemble models, Shapley Additive Explanation (SHAP) is applied to quantify both global contributions and local nonlinear effects of urban attributes. Results show that Fractional Vegetation Cover (FVC) shows the strongest cooling association with LST, especially in summer, while Water Density (WD) provides a stable cooling effect. Building Height (BH) is generally associated with lower LST within an effective range, whereas Building Footprint (BF) tends to be associated with increasing LST as density increases. Sky View Factor (SVF) and Building Plot Ratio (BP) contribute marginally. The findings provide interpretable, data-driven guidance for season-sensitive urban climate adaptation through ecological restoration and morphology optimisation.

Keywords: Urban Thermal Environment, Lightgbm, Bayesian Optimised, Additive Interpretation Algorithm, Urban Morphology

1 Introduction

Urban development is inherently intertwined with human civilization, yet rapid urbanization has led to increasing building densities and deteriorating environmental conditions [1]. The transformation of green spaces into impermeable artificial surfaces has exacerbated the Urban Heat Island (UHI) effect [2], contributing to economic losses, increased carbon emissions from cooling demands, and public health deterioration.

Traditionally, adaptations to urban ecological spaces (green and blue spaces) and adjustments to urban morphology have been primary strategies for mitigating UHI. However, early studies probing associations between urban form and UHI responses

© The Author(s) 2026
J. Zhang et al. (eds.), *Proceedings of the 2026 2nd International Conference on Engineering Management and Safety Engineering (EMSE 2026)*, Advances in Engineering Research 300,
https://doi.org/10.2991/978-94-6239-703-3_11

predominantly utilized classical statistical linear regression models. Due to simple algorithms and structural constraints, these models lacked the capability to capture the complex, nonlinear multidimensional interactions within urban microclimates, especially concerning 3D urban-form variables like Building Height (BH) and Sky View Factor (SVF).

Recent advancements have seen the adoption of machine learning methods capable of mapping nonlinear characteristics [3],[4]. Yet, advanced models often face the Black Box problem, lacking geographical and physical interpretability. To bridge this gap, this paper introduces a Bayesian optimized LightGBM model combined with the SHAP additive interpretation algorithm. Focusing on the central urban area of Beijing, China, this research disentangles the nonlinear impacts and threshold effects of both natural environments and built urban morphologies on near-surface temperatures across four seasons.

2 Methods

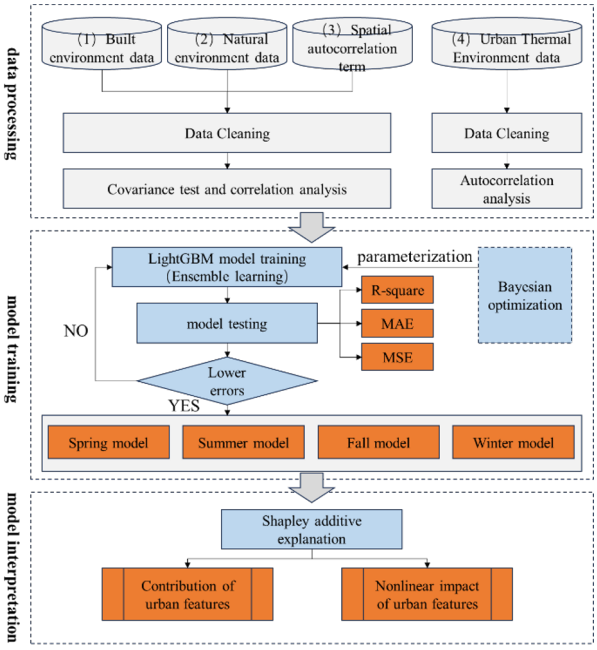

Fig. 1. Flowchart of the study

As shown in Fig. 1, the study consists of four main steps: data preparation, spatial autocorrelation detection, seasonal LightGBM modelling with Bayesian optimisation, and SHAP-based interpretation of variable effects.

2.1 Study Area and Datasets

The analysis covers the central districts of Beijing to capture representative urbanization patterns while minimizing mountainous climatic interference. Daytime LST is derived from the MODIS Near Land Surface Temperature product (MYD11A2.061, 1 km, 8-day average) for 2022, aggregated into spring, summer, autumn, and winter seasonal means.

The MODIS raw value T is converted to LST (°C) by:

$$LST = T \times 0.02 - 273.15$$

Explanatory variables include built-environment indicators (BH, Building Footprint (BF), SVF, Building Plot Ratio (BP)), natural-environment indicators (Altitude (ALT), Fractional Vegetation Cover (FVC), Water Density (WD), and a spatial autocorrelation term (LSTN) representing local neighbourhood temperature dependence. LSTN represents the mean LST of neighbouring grid cells within a defined spatial window, capturing local spatial context in the modelling process. FVC is calculated from NDVI as:

$$FVC = \frac{NDVI - NDVI_{Soil}}{NDVI_{veg} - NDVI_{Soil}}$$

Table 1 summarizes all modelling variables.

Table 1. Influencing factors for seasonal LST models

Dimension	Influencing Factors	Meaning	Unit
Built environment	building height (BH)	Reflects the average height of buildings within a certain area	m
	building floor footprint (BF)	Reflects the coverage of buildings within a certain area	%
	sky view factor (SVF)	Reflects the proportion of sky visible from the surface	%
	building plot ratio (BP)	Reflects the construction volume of buildings within a certain area	%
Natural environment	vertical projection area of vegetation(FVC)	Reflects the proportion of ground vegetation cover	%
	Altitude (ALT)	Reflects local topographic features at a certain resolution	m
	Water Density (WD)	Reflects the coverage of water within a certain area	%
Spatial autocorrelation term	LSTN	Reflects the temperature in the immediate area	°C

2.2 Spatial Autocorrelation Detection

Before regression modelling, Global Moran's I is computed for seasonal LST to test spatial dependence. Strong positive spatial autocorrelation is observed across seasons

(Moran's I approximately 0.802–0.946, with p-values < 0.001), motivating the inclusion of LSTN to partially account for spatial dependence in the modelling process.

2.3 LightGBM Regression and Bayesian Optimisation

Urban thermal processes reflect coupled effects of radiation, ventilation, and heat storage. Accordingly, we used LightGBM (a gradient-boosted decision-tree model) to represent these nonlinear relationships. LightGBM accelerates training on large geospatial datasets through Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) [5].

Given the high dimensionality of the predictors and the potential sensitivity of model performance to hyperparameter settings, Bayesian Optimization was used to tune the LightGBM hyperparameters. In implementation, the optimization procedure was conducted using Optuna, with the objective of minimizing the root mean square error (RMSE) on a validation subset. For each trial, the training data were randomly split into a training subset (80%) and a validation subset (20%) using a fixed random seed (random_state = 42). A LightGBM model was then trained on the training subset, and its predictive performance was evaluated on the validation subset. The objective function is defined as:

$$\min_{\theta \in \Theta} RMSE(\theta) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i(\theta))^2}$$

where θ denotes the hyperparameter combination, y_i is the observed LST, and $\hat{y}_i(\theta)$ is the predicted LST under hyperparameter setting θ.

The hyperparameter search space Θ was defined as Table 2:

Table 2. Hyperparameter search space for the LightGBM model in Bayesian optimization.

Hyperparameter	*Search space*
num_iterations	*integer in [100, 1000]*
reg_alpha	*log-uniform in [10^{-3}, 10]*
reg_lambda	*log-uniform in [10^{-3}, 10]*
colsample_bytree	*categorical in {0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0}*
subsample	*categorical in {0.4, 0.5, 0.6, 0.7, 0.8, 1.0}*
learning_rate	*categorical in {0.006, 0.008, 0.01, 0.014, 0.017, 0.02}*
max_depth	*categorical in {5, 7, 9, 11, 13, 15, 17, 20, 50}*
num_leaves	*integer in [100, 500]*
min_child_samples	*integer in [1, 300]*
cat_smooth	*integer in [1, 100]*

Within each trial, early stopping was applied during LightGBM training with stopping_rounds = 100, meaning that training terminated if the validation error did not

improve for 100 consecutive boosting rounds. At the optimization level, Bayesian Optimization was run for 20 trials (n_trials = 20), and the hyperparameter set yielding the lowest validation RMSE was selected as optimal. Therefore, the stopping mechanism of the overall tuning process was determined by a fixed optimisation budget (20 trials), while the stopping mechanism of each individual LightGBM model was determined by early stopping.

Bayesian Optimization constructs a surrogate probabilistic model over the hyperparameter space and uses an acquisition function to balance exploration and exploitation. In this study, the acquisition function follows the expected improvement criterion:

$$AF(x) = \begin{cases} \big(\mu(x) - f(x^+)\big)\Phi(Z) + \sigma(x)\phi(Z), & \sigma < 0 \\ 0, & \sigma \geq 0 \end{cases}$$

$$Z = \frac{\mu_x - f(x^+)}{\sigma}$$

In this formula, x denotes a candidate hyperparameter setting; $\mu(x)$ and σ(x) are the predicted mean and standard deviation from the surrogate model,; $f(x+)$ is the current best objective value,; $\Phi(Z)$ is the cumulative distribution function of the standard normal distribution, and $\phi(Z)$ is the corresponding probability density function.

2.4 SHAP Additive Interpretation Algorithm

To provide both local and global interpretability for the LightGBM models, we used Shapley Additive Explanations (SHAP). Based on Shapley values from cooperative game theory, SHAP quantifies each urban feature's contribution to the predicted outcome, thereby improving model transparency. This approach enables the interpretation of complex nonlinear relationships learned by the ensemble model while maintaining consistency with additive feature attribution theory.

$$SHAP_j = \sum_{S \subseteq \{V_1 + V_2 + \cdots + V_p\} \setminus \{V_j\}} \frac{|S|!\,(p - |S| - 1)!}{p!} \left(f_x\big(S \cup \{V_j\}\big) - f_x(S)\right)$$

$$y_i = y_{base} + \sum_{j=1}^{k} SHAP(x_{i\_j})$$

where $SHAP_j$ represents the Shapley value for variable j, S denotes a subset (coalition) of variables, p gives the number of variables, and y_i is the prediction function.

3 Results

3.1 Model Performance

All seasonal LightGBM models achieve stable predictive performance, with the best accuracy in summer and reduced performance in winter. Table 3 reports test-set met-

rics, demonstrating that the nonlinear ensemble approach captures seasonal LST variability effectively.

Table 3. Test-set performance of seasonal LightGBM models.

MODEL	*categories*	*R-square*	*MAE*	*MSE*
Spring	*Test Data Set*	*0.6594*	*0.8856*	*1.3649*
	Train Data Set	*0.6211*	*0.9503*	*1.5894*
Summer	*Test Data Set*	*0.8057*	*0.7056*	*0.8751*
	Train Data Set	*0.7774*	*0.7578*	*1.0286*
Fall	*Test Data Set*	*0.5936*	*0.6335*	*0.7304*
	Train Data Set	*0.5528*	*0.6653*	*0.8244*
Winter	*Test Data Set*	*0.4298*	*0.5701*	*0.6049*
	Train Data Set	*0.3821*	*0.5910*	*0.6701*

3.2 Global Contributions of Influencing Factors

Fig. 2. Global SHAP feature importance and contribution shares for the four seasonal models (panels a–d: spring to winter).

Fig. 2 illustrates the SHAP-based global contribution ranking for each season. FVC consistently emerges as the dominant factor in spring, summer, and autumn, accounting for over one-third of total contribution and approaching half in summer. In winter, FVC contribution decreases markedly (about 18.2%), consistent with sparse vegetation conditions. BH and BF remain important across seasons; BP and SVF contribute marginally (each typically below 5%), suggesting that building height by itself is unlikely to explain elevated daytime LST in this context.

3.3 The Effect of Environment on LST

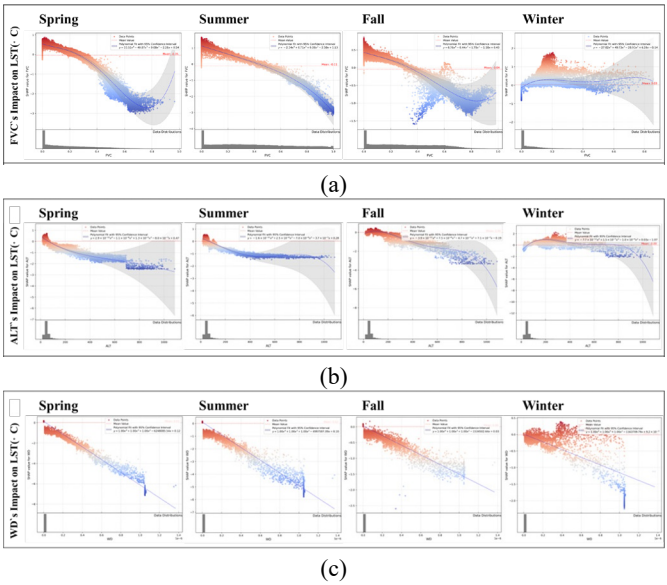

Fig. 3. Effects of FVC, ALT, and WD on LST across four seasons. (a) Effect of FVC on LST. (b) Effect of ALT on LST. (c) Effect of WD on LST

The study highlights the significant impact of environmental factors on Land Surface Temperature (LST), particularly focusing on Fractional Vegetation Cover (FVC), Altitude (ALT), and Water Density (WD).

Figure 3 showcases the overall impacts of FVC, ALT, and WD on LST in the four-season models, highlighting their varying degrees of influence.

Figure 3 illustrates the effects of FVC on LST across all four seasons, depicting temperature variations ranging from 0.9°C to -2.9°C during summer and 1.3°C to -3.1°C in spring. As the seasons transition, the effect of FVC diminishes, with autumn and winter showing smaller ranges of influence, indicating a decrease in the cooling effect of vegetation cover. Notably, a threshold of 0.5 FVC marks a significant turning point where the influence shifts from positive to negative effects on LST, with every 0.1 increase in FVC leading to an average decrease of 0.84°C in LST during the summer months.

ALT consistently affects LST across all seasons, reflecting its intrinsic nature as a spatial property.

Meanwhile, WD also exhibits a cooling effect, particularly in spring and summer, with average reductions of 0.66°C, 0.50°C, and 0.14°C for each 0.1 increase in WD across these seasons. This indicates that higher water density contributes significantly to urban cooling, emphasizing the importance of maintaining vegetation and water bodies to mitigate urban heat.

3.4 Impact of Urban Morphology on LST

The influence of urban morphology on LST is examined through Building Height (BH), Building Plot Ratio (BP), and Building Footprint (BF).

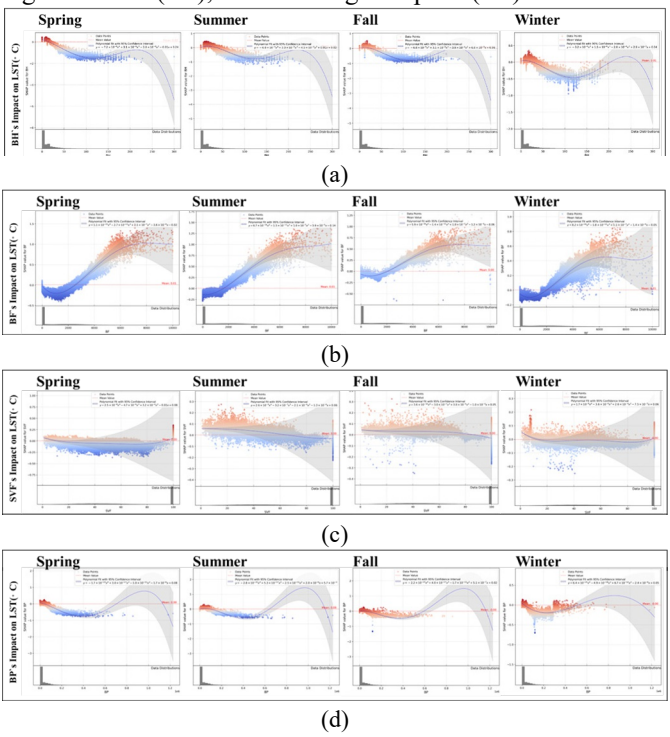

Fig. 4. Seasonal effects of urban morphology variables on LST: (a) BH, (b) BF, (c) SVF, (d) BP.

Figure 4 illustrates the seasonal variations in the impacts of BH and BP on LST, where BH shows a more extensive range of impact compared to BP, affecting LST by 0.35°C in spring and up to 1.7°C in summer. The effects of both BH and BP exhibit diminishing returns, where BP stabilizes at certain thresholds depending on the season, suggesting that beyond a specific ratio, additional increases do not yield proportional decreases in LST.

Additionally, the Sky View Factor (SVF) has a relatively minor influence on LST, indicating that while it is a factor in urban heat dynamics, its effect is less significant than that of building dimensions.

In contrast, the impact of BF displays a distinct pattern, initially showing slight cooling effects that shift to warming as BF increases, particularly when surpassing specific thresholds. As BF rises, every 100-unit increase correlates with a gradual rise in LST, averaging increases of 0.16°C in spring and 0.05°C in winter. This emphasizes the critical role of urban morphology in shaping thermal dynamics in urban environments, suggesting that careful urban planning and design can significantly mitigate

urban heat island effects. Overall, the research underscores the importance of both environmental and morphological factors in influencing urban thermal dynamics, providing valuable insights for effective urban planning and climate adaptation strategies.

4 Conclusion

The study's results indicate that ecological degradation is a primary factor in urban heat intensification. Specifically, in summer, when vegetation coverage surpasses 48%, every 10% increase in vegetation leads to a 0.8℃ reduction in surface temperature. In other seasons, the impact of vegetation coverage is still significant, albeit with varying degrees of marginal utility. Moreover, the influence of urban morphology characteristics, like ecological factors, displays seasonal variability and phased changes. Building height and footprint show stronger associations with LST variations than building plot ratio. During spring, summer, and autumn, building heights ranging from 15m to 75m exhibit a pronounced cooling effect on LST, which intensifies with increased height. By comparison, larger building footprints are associated with a near-linear rise in LST.

Despite the satisfactory predictive performance of the models, several limitations should be acknowledged. Although spatial autocorrelation was partially considered by introducing the neighbourhood temperature variable (LSTN), this approach cannot fully eliminate potential bias arising from spatial dependence in the observations. Machine-learning models such as LightGBM primarily optimise predictive accuracy rather than explicitly modelling spatial processes, and residual spatial autocorrelation may still influence the estimated feature contributions. Future research could further address this issue by integrating more rigorous spatial modelling strategies, such as spatial regression frameworks, spatial cross-validation schemes, or geographically weighted approaches. Incorporating these methods may help to better disentangle spatial dependence from the intrinsic effects of urban morphology and environmental variables.

In conclusion, for effective daytime temperature management, urban spatial planning should prioritize ecological restoration and consider raising building heights to enhance urban habitats. Focusing on Beijing, the study suggests maximizing building heights within the 15-75m range to minimize building footprints. Additionally, increasing vegetation coverage above 48% is identified as a key approach to addressing urban surface temperature challenges. These findings provide valuable, data-driven insights for developing sustainable urban environmental policies.

Acknowledgements

We acknowledge the open-source datasets and services that made this study possible, including MODIS products and digital elevation data.

References

1. A. Das, M. K. Annaqeeb, E. Azar, V. Novakovic, and M. B. Kjærgaard, “Occupant-centric miscellaneous electric loads prediction in buildings using state-of-the-art deep learning methods,” Applied Energy, vol. 269, p. 115135, Jul. 2020, doi: 10.1016/j.apenergy.2020.115135.
2. C. Molina, M. Kent, I. Hall, and B. Jones, “A data analysis of the Chilean housing stock and the development of modelling archetypes,” Energy and Buildings, vol. 206, p. 109568, Jan. 2020, doi: 10.1016/j.enbuild.2019.109568.
3. S. Liu, J. Zhang, J. Li, Y. Li, J. Zhang, and X. Wu, “Simulating and mitigating extreme urban heat island effects in a factory area based on machine learning,” Building and Environment, vol. 202, p. 108051, 2021, doi: 10.1016/j.buildenv.2021.108051.
4. Q. Cao, Q. Luan, Y. Liu, and R. Wang, “The effects of 2D and 3D building morphology on urban environments: A multi-scale analysis in the Beijing metropolitan region,” Building and Environment, vol. 192, p. 107635, Apr. 2021, doi: 10.1016/j.buildenv.2021.107635.
5. E. A. Daoud, “Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset,” Int. J. Comput. Inf. Eng., vol. 13, no. 1, pp. 6–10, Jan. 2019.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.