ENVIRONMENTAL RESEARCH
HEALTH

OPEN ACCESS

RECEIVED
24 October 2025

REVISED
11 January 2026

ACCEPTED FOR PUBLICATION
11 February 2026

PUBLISHED
2 March 2026

PAPER

Nonlinear relationships between urban morphology, socioeconomic conditions, and infectious disease risk: evidence from COVID-19 in Tokyo

Yangguang Xiao[1], **Kojiro Sho**[2,*], **Yuya Shibuya**[3], **Kimihiro Hino**[2], **Shichen Zhao**[4] and **Ronita Bardhan**[5,6]

[1] Graduate School of Human-Environment Studies, Kyushu University, Fukuoka, Japan
[2] Department of Urban Engineering, School of Engineering, The University of Tokyo, Tokyo, Japan
[3] The University of Tokyo Interfaculty Initiative in Information Studies, Tokyo, Japan
[4] Faculty of Human-Environment Studies, Kyushu University, Fukuoka, Japan
[5] Sustainable Design Group, Department of Architecture, University of Cambridge, Cambridge, United Kingdom
[6] Current address: London School of Hygiene and Tropical Medicine, London, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: shokojiro@g.ecc.u-tokyo.ac.jp

Keywords: urban morphology, nonlinear relationship, COVID-19, Tokyo, public health, machine learning

Supplementary material for this article is available online

Abstract

Urban infectious disease outbreaks pose critical challenges to public health in rapidly urbanizing cities. The COVID-19 pandemic provided a natural experiment to examine how area-level environmental and socioeconomic contexts are associated with infection risk over time. This study analyzed seven pandemic stages in Tokyo, Japan (2020–2022), across 53 municipalities using random forest-based interpretable machine learning models with SHapley Additive exPlanations and partial dependence plots diagnostics. Feature selection and nonlinear modeling identified key drivers among 49 candidate variables. The results revealed significant spatial clustering of COVID-19 infection rates, with persistent hotspots in central wards like Shinjuku, Minato, and Shibuya, while suburban regions of Western Tokyo maintained lower infection rates. Key built environment factors exhibited stage-specific, nonlinear, and threshold-like associations with infection rates, including road density (25 km km^{-2}), FAR (0.25–0.3), and population density (400 and 600 people/km^2). These associations were predominantly positive beyond the identified thresholds, indicating elevated infection risk under higher density and connectivity conditions. Socio-demographic factors also showed temporal specificity: the percentage of foreigners displayed a threshold around 1.5%, and construction worker density emerged as a relevant correlate during the Omicron-dominated phase. Overall, the relative importance and marginal patterns of these associations varied across intervention stages, highlighting temporal instability in area-level risk correlates. Importantly, these findings are associational rather than causal, reflecting contextual exposure conditions at the municipal scale. From an environmental epidemiology perspective, the results suggest that stage-sensitive and spatially explicit interpretation of area-level indicators may enhance infectious disease surveillance, compared with approaches assuming temporally invariant or linear effects.

1. Introduction

As urbanization has accelerated, the impacts of urban morphology—defined by network characteristics and parameters of the built environment such as floor area ratio (FAR; Guo *et al* 2025) and network density (Wang and Zha 2024), i.e. man-made spatial settings that shape human activity and interaction patterns—on population health have received increasing attention (Azzopardi-Muscat *et al* 2020,

Guida and Carpentieri 2021, Marselle *et al* 2021). During the COVID-19 pandemic, studies related to the characteristics of urban morphology on the spread of infectious diseases became a crucial research topic in the fields of urban planning and public health (Lai *et al* 2020). The role of nonpharmacological interventions is increasingly recognized and prioritized alongside the search for pharmacological treatments and vaccines to treat and prevent viral transmission (Alidadi and Sharifi 2022). Studies have demonstrated that urban morphology, health, socioeconomic and personal factors, and the interactions between them have profound effects on the transmission and control of infectious diseases (Sharifi and Khavarian-Garmsir 2020).

Importantly, urban morphology is generally understood to influence infection risk indirectly by shaping population mobility patterns and contact intensity, rather than acting as a direct causal determinant.

Recent studies have explored how urban morphology and socioeconomic factors impact health, particularly through the lens of the transmission dynamics of COVID-19. Key factors—such as density, land use, and access to public infrastructure—significantly influenced the spread of the virus in urban settings (You *et al* 2020, Hu *et al* 2021). Density exhibits dual effects, as higher contact rates in dense areas may increase transmission (Ren *et al* 2020, You *et al* 2020), while superior access to healthcare in these areas might enhance compliance with preventive measures (Hamidi *et al* 2020, Liu 2020). Moreover, while mixed land use (i.e. the spatial coexistence of residential, commercial, and service functions within the same area) can reduce travel and infection risk, it may draw diverse populations to centralized facilities, which could potentially increase the risk of infection (Lak *et al* 2021). Furthermore, green spaces have generally helped to mitigate the spread of the virus (Sharifi 2022), whereas increased outdoor activity in these areas has sometimes elevated the risk of exposure (Huang *et al* 2020). Additionally, findings regarding public transport have been mixed, with some studies noting increased transmission risks (Zheng *et al* 2020) and others observing no significant correlations (Hamidi and Hamidi 2021).

Nevertheless, evidence from other global high-density metropolises—such as New York, Hong Kong and Tehran—suggests that similar associations between urban structure, mobility, and COVID-19 transmission exist. This allows Tokyo to be examined as a representative yet institutionally distinct case within this broader metropolitan context.

In parallel, Many studies have also examined the effects of socioeconomic factors. When an individual is infected, overcrowded housing increases susceptibility (Alidadi and Sharifi 2022). Additionally, certain socioeconomic factors like one's occupation, income, family structure, travel behaviors, age, educational level, social class, and housing prices have been influential in explaining the variability and intensity of COVID-19 transmission in urban areas (Ali *et al* 2021, Kashem *et al* 2021, Liao *et al* 2021). For instance, individuals who work in healthcare, transportation, entertainment, retail, and other life-related services are often more susceptible due to their physical proximity to others (Wang *et al* 2021).

Methodologically, research on the relationships between urban morphology and COVID-19 has typically employed linear models (Liu *et al* 2021, Wang *et al* 2021, Alidadi *et al* 2023). Additionally, spatial models like geographically weighted regression have been used to explore the effects of spatial autocorrelation and heterogeneity (Gaisie *et al* 2022, Liu *et al* 2021, Ulubaş Hamurcu and Yılmaz 2023). However, linear models may not accurately capture the complex relationships between environmental variables and COVID-19, as some factors of urban morphology could exhibit nonlinear and threshold effects on viral transmission (Ma *et al* 2021).

While relevant studies have provided critical insights into the impact of socioeconomic and urban morphology factors on COVID-19 transmission, they have reported limitations in two key areas. First, many of these studies have assumed a consistent relationship due to their focus on a single phase of the pandemic. However, epidemiological evidence suggests that viruses at different mutation stages can lead to varied epidemic outcomes, with the same factors potentially having different effects in the same location over time (Twohig *et al* 2022). Few studies have addressed the impacts across multiple stages of the pandemic (Liao *et al* 2021, Wang *et al* 2021, Gaisie *et al* 2022). Moreover, research has predominantly employed correlation and regression analyses to explore linear relationships and often failed to uncover the nonlinear patterns of how urban morphology factors influenced COVID-19's transmission. A few studies have adopted machine learning approaches to analyze these relationships, such as Ma *et al* (2021), who used random forest models to objectively assess the key factors of urban morphology that influence the early spread of COVID-19 in townships in China, and Pan *et al* (2021), who applied random forest models to capture epidemiological dynamics and make time-series projections; however, these studies have typically focused on single-stage data analyses without considering the dynamics of the pandemic across its multiple stages.

Given that the COVID-19 pandemic in Tokyo unfolded under successive states of emergency and priority preventive measures, the urban environment and social behavior of residents evolved significantly over time. Each stage reflected distinct combinations of viral characteristics, government interventions,

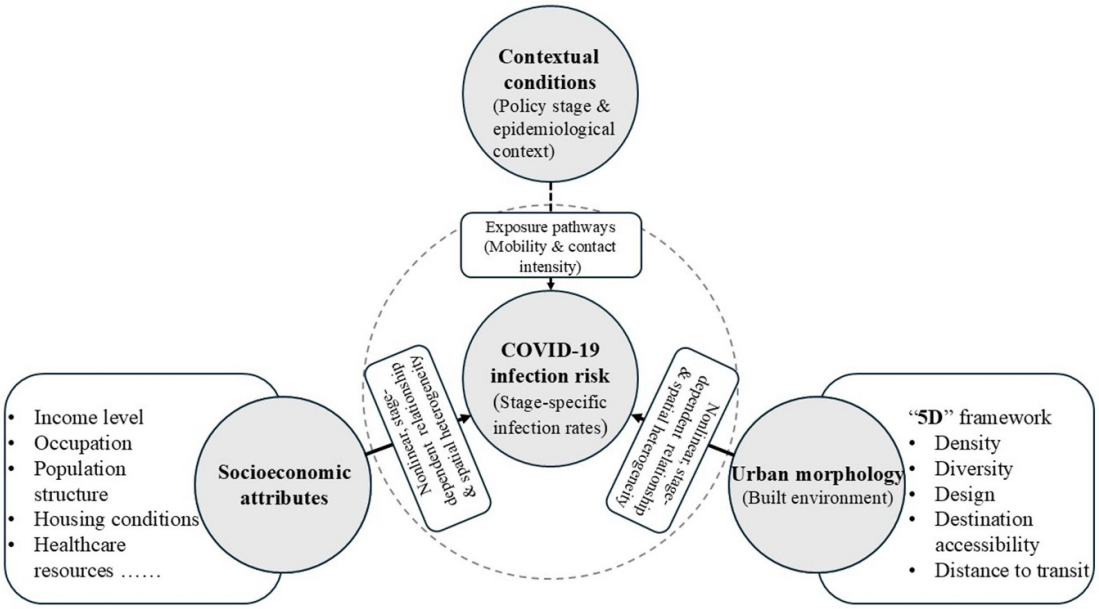

Figure 1. Theoretical framework. The framework depicts the nonlinear and stage-dependent associations between socioeconomic attributes, urban morphology, and COVID-19 infection risk, mediated by exposure pathways under different contextual conditions.

and mobility restrictions, leading to different spatial transmission mechanisms. Examining multiple stages therefore enables the identification of stage-specific drivers-such as density and mobility in the early stages, and social composition or occupational exposure in later stages-that influence infection risk. Such understanding provides an evidence-based foundation for developing adaptive, phase-sensitive strategies to mitigate disease transmission and strengthen urban resilience.

To fill the aforementioned research gaps, this study aimed to identify the key socioeconomic and urban morphology factors that influenced the spatial distribution of infection rates across different stages of the COVID-19 pandemic in Tokyo, Japan, to contribute to contemporary discussions on public health resilience. Urban morphology does not directly determine infection rates; rather, it exerts indirect effects through population mobility, contact intensity, and accessibility patterns that shape human interactions in space. Importantly, these relationships are expected to be nonlinear and stage-dependent, with threshold effects that vary across epidemiological contexts. Understanding these behavioral and spatial pathways is essential for disentangling the structural determinants of infection risk beyond government interventions and healthcare capacity.

Figure 1 presents the theoretical framework of this study, synthesizing these indirect, nonlinear, and stage-specific relationships between urban morphology, socioeconomic conditions, and COVID-19 infection risk under different policy and epidemiological contexts.

2. Methods

2.1. Research design

This study aimed to leverage geospatial data and machine learning models to explore the spatial distribution patterns of COVID-19 infection rates per 100 000 people across 53 wards and municipalities of Tokyo during seven stages of pandemic control measures (table 1). It also aimed to understand the relationship between socioeconomic and urban morphology factors during different stages of COVID-19 infection. First, a dataset of 49 explanatory variables related to COVID-19's impact was constructed from socioeconomic and urban morphology perspectives, by integrating publicly available municipal-level statistics, point-of-interest densities, built-environment indicators, and population structure measures commonly used in urban health and environmental epidemiology studies. Second, the support vector machine recursive feature elimination (SVM-RFE) algorithm was used to extract the optimal subset of explanatory variable features for each of the seven stages. Third, three machine learning models were established, and Shapley Additive exPlanation (SHAP) and partial dependence plot (PDP) models (Xiao *et al* 2021, Kim and Lee 2023) with high interpretability were employed to reveal the exact thresholds of the nonlinear dependent effects of various factors that influenced COVID-19 infection rates. Finally,

Table 1. Summary of key features of the seven pandemic stages.

Intervention phases	First state of emergency	Second state of emergency	First priority measures	Third state of emergency	Second priority measures	Fourth state of emergency	Third priority measures
Date	April 7, to 25 May 2020	January 8, to 21 March 2021	April 12, to 24 April 2021	April 25, to June 2021	June 21, to 11 July 2021	July 12, to 30 September 2021	January 21, to 21 March 2022
Total number of cases	4,020	44 914	7,675	30 963	11 801	180 001	690 675
Key features	First state of emergency; school and business closures; public activity restrictions; public transportation reductions; limited human contact	In response to the winter surge; shortening of business hours and residents urged to minimize outings; increased hospitalizations	Imposition of stricter regulations in specific areas; focus on public awareness and community prevention	Ahead of the Tokyo Olympics, comprehensive lockdown measures implemented, with accelerated vaccine distribution to reduce infection rates	As vaccination rates increased, restrictions were gradually relaxed, allowing some economic recovery, though concerns about the Delta variant influenced control strategies	During the Olympics, strict border controls were enforced to manage risks from the Delta variant, leading to a strain on hospitals and further tightened restrictions on nighttime economic activities	Emergence of the Omicron variant prompted booster vaccination efforts, the phased relaxation of restrictions, and gradual lifting of restrictions as cases declined

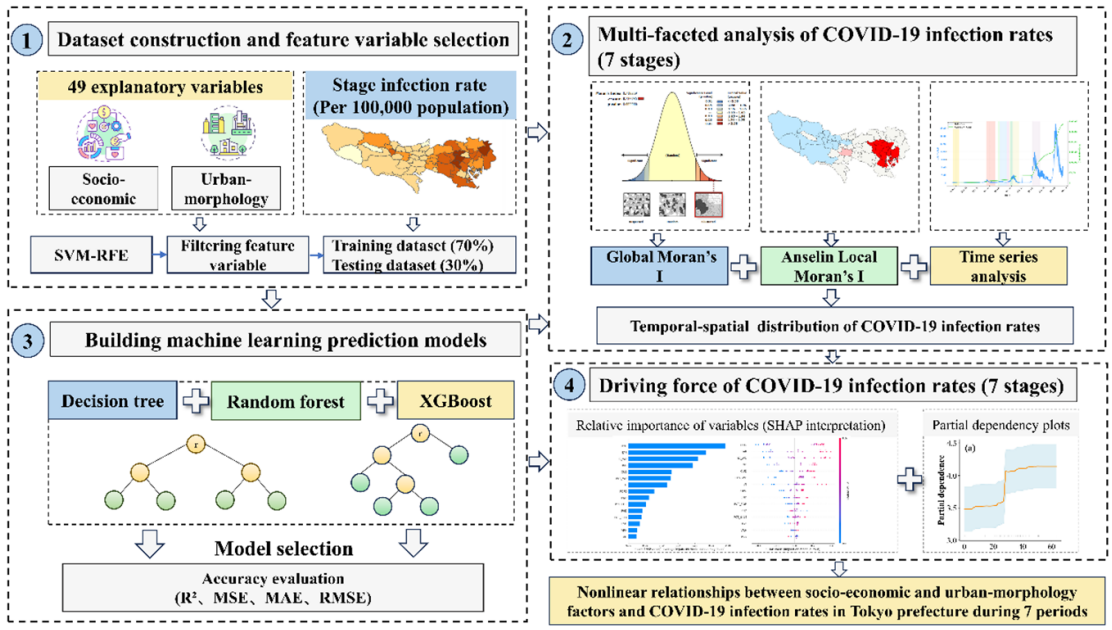

Figure 2. Research framework of this study.

based on the results of quantitative analyses, new perspectives were developed to understand how factors of urban morphology influence the prevalence rates of infectious diseases. Figure 2 presents the research framework.

2.1.1. Basis for the seven pandemic stages

The seven stages of the COVID-19 pandemic in Tokyo were delineated according to the official declarations of states of emergency and priority preventive measures issued by the Japanese government between April 2020 and September 2022. These phases represent discrete shifts in intervention intensity, rather than uniform or continuous temporal segmentation.

Each stage corresponds to a distinct public health response framework, characterized by changes in population mobility, business operation limits, and public compliance levels. Such policy-induced behavioral shifts have been shown to substantially influence transmission dynamics in previous studies and official mobility reports (Yabe *et al* 2020, Nagata *et al* 2021, Tokyo Metropolitan Government 2022).

Although fluctuations in infection counts may coincide with seasonal transitions, the present stage delineation is explicitly policy-based rather than season-based, This design aims to capture temporal changes in transmission mechanisms driven primarily by government interventions and collective behavioral responses. While transitions in dominant viral variants (e.g. the emergence of the Omicron variant) may temporally overlap with certain policy phases, the stages are intended to serve as proxies for intervention-driven behavioral contexts rather than virological characteristics.

The detailed timeline and definitions of the seven pandemic stages are provided in table 1. Recognizing that alternative temporal aggregations may also be plausible, sensitivity analyses using coarser stage delineations were conducted to assess the robustness of the main findings. Detailed results of these analyses are reported in the appendix A.1.

2.2. Study area

This study focused on Tokyo's 23 special wards and the Tama area (i.e. Western Tokyo's 26 cities, three towns, and one village in Nishitama District) under the jurisdiction of the Tokyo Metropolitan Area (figure 3). As of 2023, the resident population of the Tokyo Metropolitan Area was approximately 14 040 625, making it one of the most densely populated areas of the world, with an average density of 6306 people/km^2, rising to 15 606 people/km^2 in the city center (Tokyo Metropolitan Government Bureau of General Affairs 2023). The center also hosts administrative agencies, enterprises, commercial facilities, and an extensive transportation network.

Japan reported its first COVID-19 case in February 2020 and experienced multiple waves of infection. From April 2020 to October 2022, the government implemented various measures, including the declaration of states of emergency in April 2020 and several times in 2021. Schools were closed and

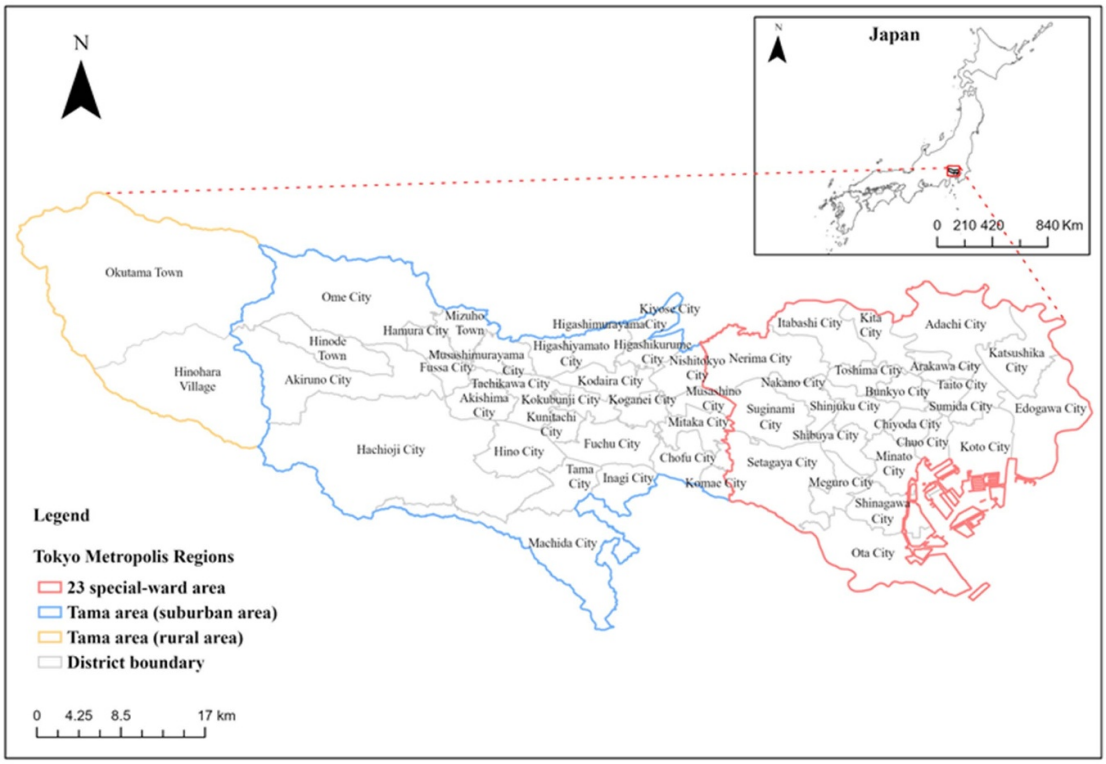

Figure 3. Map of the study area.

many employees switched to working remotely. Border restrictions were imposed on travelers from over 100 countries and, in March 2021, the entry of nonresident foreign nationals was temporarily banned.

Among all regions of Japan, Tokyo was the most severely affected by COVID-19, with 2866 240 confirmed cases in total during the study period (Infectious Disease Data and Medical/Health Information of NHK, 2022). While Tokyo accounts for 11% of Japan's population, it accounted for 24% of all confirmed cases. Although the current legal system does not permit the government to enforce mandatory lockdowns (Yabe *et al* 2020), the government declared a state of emergency during the Tokyo Olympics (23 July–5 August 2021), accompanied by mobility restrictions and a mass vaccination campaign, resulting in an unexpectedly low COVID-19 infection rate by the end of October 2021 (Karako *et al* 2022). Note that the reported infection rate reflects confirmed PCR cases only and does not represent the actual infection prevalence. The policy differences and high infection rates make Tokyo an interesting case for conducting an international comparative analysis during the COVID-19 pandemic.

2.3. Data collection and variables

Socioeconomic and urban morphology data for the study area were collected from various official statistical agencies along with daily COVID-19 infection data. All data used in this study were obtained from publicly available, aggregated sources at the municipal level. No individual-level or identifiable information was involved; therefore, ethical approval was not required and data privacy concerns are minimal.

2.3.1. Outcome variables

COVID-19 infection data were derived from daily confirmed case numbers collected by Japan's Ministry of Health, Labor and Welfare, which covered the spread of COVID-19 across 53 districts of the Tokyo Metropolitan Area from April 2020 to October 2022 (Tokyo COVID-19 Task Force website n.d.). The primary outcome variable was the municipal-level COVID-19 infection rate (cases per 100 000 population), computed for each of the seven policy-defined pandemic stages. Although Tokyo Metropolitan Area administratively consists of 62 municipalities, the present analysis was restricted to 53 municipalities for which complete and temporally consistent COVID-19 infection data and explanatory variables were available across all seven pandemic stages. Several municipalities—primarily small towns and island areas—were excluded due to limited population size, incomplete infection surveillance records, or lack of comparable built-environment indicators. This restriction was applied to ensure stability and comparability in infection rate estimation across spatial units. The COVID-19 infection data were processed

according to the seven designated infection stages. Then, by combining them with population data (Tokyo Metropolitan Government 2023), the COVID-19 infection rates per 100 000 people for each district were calculated during each infection stage. All infection rates were standardized per 100 000 residents to ensure comparability across districts with different population sizes. Prior to model estimation, all variables were examined for missing values, and no missing observations were identified after data aggregation. To reduce skewness and stabilize variance, continuous explanatory variables were transformed using a Box–Cox transformation before model training.

2.3.2. Explanatory variables

For the explanatory variables, socioeconomic variables as well as urban morphology variables were selected in Appendix B (table B.1). Urban morphology refers to man-made settings that provide the setting for human activity, which range from buildings and infrastructure to parks. Based on the 5D framework (Ewing and Cervero 2010), which is currently widely used in relevant research, the following 19 urban morphology attributes were selected: population density (POPD), FAR, and building density (*Density*), which increase contact frequency and airborne transmission risk (Hamidi *et al* 2020, Chen *et al* 2021); the Shannon diversity index (*Diversity*) measuring land use mixing, which generates complex mobility patterns increasing cross-neighborhood exposure (Huang *et al* 2021a); road density (ROD) and intersection density (IND; *Design*) characterizing walkability and pedestrian encounters (Yang *et al* 2024); and the following number densities: accommodation, catering, company, cultural (CUD), educational, entertainment (END), hospital, public service, shopping, sports, parks (*Destination accessibility*), representing congregation sites where transmission risk concentrates (Hamidi *et al* 2020, Li *et al* 2020, Lak *et al* 2021, Liu *et al* 2021, Ma *et al* 2021, Yip *et al* 2021, Huang *et al* 2021b, Nasiri *et al* 2022). Parks were operationalized as discrete counts rather than continuous area because they function as gathering locations with concentrated visitor flows, creating localized transmission hotspots in dense contexts (Shoari *et al* 2020, Venter *et al* 2020). Bus stops, and railway stations (*Distance to transit*) quantifying exposure in confined transit environments (Xiao and Liu 2023). The socioeconomic variables captured vulnerability through income level, educational level, employment status, population structure, housing conditions, healthcare resources, and others, reflecting differential exposure risk and protective capacity (Karaye and Horney 2020, Williamson *et al* 2020, Alidadi *et al* 2023).

While this study primarily focused on socioeconomic and urban morphology attributes, variations in government interventions and healthcare capacity were indirectly controlled through the seven-stage framework, which reflects differences in policy stringency, mobility restrictions, and medical system stress across time.

To mitigate potential overfitting and multicollinearity issues inherent in meso-scale administrative-level analyses, Pearson correlation screening ($|r| > 0.7$ threshold) and feature selection were performed prior to model estimation. The final explanatory variables were selected to ensure both conceptual soundness and empirical robustness for spatial and machine learning analyses.

2.4. Analytical methods

Although urban morphology indicators such as FAR, ROD, and IND are often analyzed at micro or pixel levels, this study aggregates them to the administrative unit level, comprising 53 wards and municipalities in the Tokyo Metropolitan Area, to align with the spatial resolution of publicly available COVID-19 infection data. This meso-scale approach ensures analytical consistency between infection data and explanatory variables, capturing spatial disparities in infection dynamics at a policy-relevant scale while maintaining the interpretability of urban morphology indicators. Similar administrative-scale analyses have been applied in metropolitan COVID-19 studies (e.g. Liu *et al* 2021, Gaisie *et al* 2022). This scale selection allows the findings to remain comparable with government statistics and directly applicable to urban health policy and planning.

2.4.1. Spatial statistical methods

To examine the spatial heterogeneity of COVID-19 infection rates per 100 000 people across the seven phases in the Tokyo Metropolitan Area, the following two spatial statistical approaches were employed: Global Moran's I and Anselin Local Moran's I. Each method provides a unique perspective on spatial autocorrelation and clustering patterns. Global Moran's I was used to assess whether COVID-19 infection rates exhibited significant overall spatial autocorrelation across the entire study area, indicating a general tendency toward spatial clustering or dispersion. In contrast, Anselin Local Moran's I was applied to identify the specific locations and types of localized spatial clusters and spatial outliers by comparing

each municipality with its neighboring areas. All spatial statistical analyses were performed in ArcGIS Pro 2.5.2. Detailed descriptions of these methods are provided in the following subsections.

2.4.1.1. Global Moran's I

We evaluated the strength of spatial autocorrelation using Moran's I statistics (Anselin 1995). The Global Moran's I statistic provides insights into the degree of clustering or the dispersion of infection rates across the study area (Dadashpoor and Alidadi 2017). The equation used to calculate Global Moran's I was as follows:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\right)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \tag{1}$$

where n denotes the number of regions in the study area; x_i represents the attribute value in the i th region; x_j represents the attribute value in the j th region; $\bar{x}$ is the mean attribute value across all regions; and w_{ij} denotes the spatial weight matrix, where i is not equal to j.

2.4.1.2. Cluster and outlier analysis (Anselin Local Moran's I)

Local Moran's I identifies local spatial clusters and outliers, which enhance global statistics. It reveals significant clustering and spatial outliers using a Moran scatterplot and identifies statistically significant hot and cold spots through exploratory spatial data analysis (Dall'erba 2005). In this study, the cluster types identified by Anselin Local Moran's I are interpreted as follows: A 'high–high' cluster indicates a municipality with a high COVID-19 infection rate that is spatially surrounded by neighboring municipalities also exhibiting high infection rates, reflecting a localized concentration of elevated risk. Conversely, a 'low–low' cluster represents an area with consistently low infection rates among spatially adjacent municipalities. The 'high–low' and 'low–high' categories denote spatial outliers, where a municipality's infection rate contrasts with that of its surrounding neighbors. The Local Moran's I statistic of spatial association was calculated using the following equation:

$$I_i = \frac{x_i - \bar{x}}{S_i^2}\sum_{j=1, j\neq i}^{n} w_{ij}\left(x_j - \bar{x}\right) \tag{2}$$

$$S_i^2 = \frac{\sum_{j=1, j\neq i}^{n}\left(x_j - \bar{x}\right)}{n-1} \tag{3}$$

where x_i is an attribute for feature i; $\bar{x}$ is the mean of the corresponding attribute; w_{ij} is the spatial weight between feature i and j; and n is the total number of features.

2.4.2. Machine learning models

Three advanced machine learning models were trained and validated—namely a random forest, decision tree, and XGBoost—using spatial datasets. Through comprehensive comparisons and validations, the optimal model was selected to identify the key influencing factors and nonlinear driving mechanisms underlying the COVID-19 infection rates per 100 000 population across the seven stages in Tokyo. Random forest was adopted as the primary model due to its strong performance in small-sample settings, robustness to nonlinear relationships, and compatibility with SHAP-based interpretation. Decision Tree models were included as a transparent baseline, while XGBoost was incorporated to evaluate whether boosting-based ensemble learning could further improve predictive performance. Hyperparameters for all machine learning models were optimized using GridSearchCV with five-fold cross-validation, applying a consistent parameter grid and a 70/30 training–validation split (random_state = 42) across all pandemic stages.

2.4.2.1. Feature variable selection

The SVM-RFE algorithm is a widely used feature selection method that ranks predictors based on support vector weights and iteratively removes less informative variables, thereby reducing dimensionality and mitigating overfitting (Chandrashekar and Sahin 2014, Cruz *et al* 2021).

Given the relatively small sample size ($n = 53$ municipalities) compared with the initial pool of 49 explanatory variables, explicit feature selection was necessary to improve model generalizability. SVM-RFE was therefore applied to identify parsimonious, stage-specific subsets of predictors while preserving the original feature structure.

Unlike variance inflation factor (VIF) diagnostics, which are primarily designed for coefficient-based inference in linear regression, SVM-RFE evaluates predictors according to their contribution to predictive

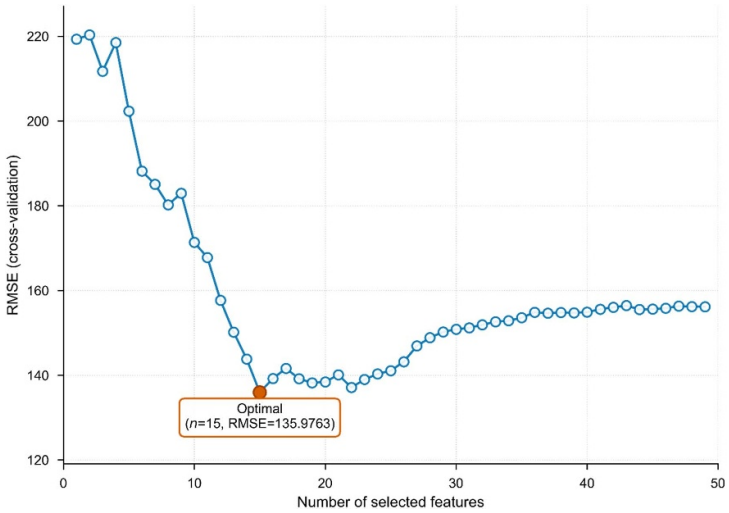

Figure 4. SVM-RFE result output in SIR1.

Table 2. Summary of SVM-RFE feature selection results across seven pandemic stages.

Stage	Initial features	Optimal features	RMSE (CV)	Reduction rate (%)
SIR1	49	15	135.98	69.4
SIR2	49	14	4626.32	71.4
SIR3	49	5	218.16	89.8
SIR4	49	16	2775.63	67.3
SIR5	49	6	433.27	87.8
SIR6	49	23	71 627.43	53.1
SIR7	49	35	521 265.23	28.6

performance and does not require the exclusion of correlated variables *a priori*. This makes it well suited to urban morphology datasets, where many indicators capture overlapping structural characteristics. VIF diagnostics were therefore treated as complementary robustness checks rather than as selection criteria.

SVM-RFE was implemented within a cross-validation framework, and the optimal number of features at each pandemic stage was determined by minimizing the cross-validated root mean square error (RMSE). The resulting feature subsets were subsequently used as inputs for all tree-based machine learning models. To ensure that the selected feature sets did not bias model results, additional robustness analyses-including sensitivity tests for extreme multicollinearity and alternative temporal stage delineations-were conducted and are described in section 2.4.2.5, with detailed results reported in the appendix A. Figure 4 presents the SVM-RFE results for the SIR1 stage, while results for the remaining stages are shown in Appendix C (figures C.1–C.6). A summary of optimal feature counts and corresponding RMSE values across all seven stages is provided in table 2.

2.4.2.2. Decision tree model

Decision trees automatically model training data to classify or predict test data. They are highly interpretable, which enables the discovery of features and extraction of patterns in large databases. The nonparametric C4.5 algorithm used in this study sequentially partitions data based on classification criteria (Prajwala 2015, Alita *et al* 2021). The formula is as follows:

$$\text{Entropy}(S) = \sum_{i=1}^{n} -p_i \times \log_2 p_i \tag{4}$$

where S represents the case set; n represents the number of partitions; and p_i represents the proportion of S_i to S.

The gain value was calculated using the following equation

$$\text{Gain}(S) = \text{Entropy}(s) - \sum_{i=1} \frac{S_i}{S} \times \text{Entropy}(S_i) \tag{5}$$

where A represents a feature or attribute; n represents the number of attribute partitions A; $|S_i|$ represents the proportion of S_i to S; $|S|$ represents the number of cases in S; and Entropy(S_i) represents the entropy for samples with the i value.

Based on the optimization of the model parameters, a decision tree model was constructed using the seven optimal feature variables for the infection stages derived from the SVM-RFE algorithm. Then, the predictive accuracy of the model was verified with an independent sample (30%).

2.4.2.3. Random forest model

Random forests are a versatile machine learning method developed by Breiman (2001) to address classification and regression challenges. They construct multiple uncorrelated decision trees from diverse data subsets and combine their predictions. This enhances model accuracy through statistically aggregating individual tree decisions.

The bootstrap aggregating (bagging) method is used to randomly select samples to build a subset of data through put-back sampling. Based on observed samples, the predicted response variable (y) can be obtained from the set of explanatory variables (y). D bootstrap samples can be derived by resampling the original data with replacements. After fitting is performed for each bootstrap sample, the final prediction of the model is obtained by averaging all predictions as follows:

$$\hat{y} = \hat{f}(z) \tag{6}$$

$$\hat{f}_{\text{bagging}}(z) = \frac{1}{D}\sum_{d=1}^{D}\hat{f}^{*d}(z) \tag{7}$$

where $\hat{f}^{*d}$ is the predicted response variable of bootstrap sample d. Bagging reduces the variance of the final prediction.

In this study, the optimal explanator variables for each of the seven research phases were selected to construct a random forest model based on SVM-RFE. The parameter *ntree* was set to the default value of 500 while *mtry* was set as 5 (the number of conditioning factors divided by 3), as confirmed by numerous studies.

2.4.2.4. Extreme gradient boosting model

The XGBoost algorithm enhances the original gradient boosting decision tree algorithm (Friedman 2001) by addressing challenges related to efficiency and complexity. XGBoost employs CPU multithreading and second-order Taylor expansion to improve the training speed and prevent overfitting (Si *et al* 2021). Unlike random forests, which focus on hyperparameter optimization, XGBoost prioritizes functional space to reduce model cost (Didavi *et al* 2021). XGBoost is detailed as follows (Chen and Guestrin 2016):

The tree model is integrated with the addition method, assuming a total of K trees, and F is used to represent the basic tree model. Then,

$$\widehat{y_i} = \sum_{k=1}^{K} f_k(x_i), f_k \in F. \tag{8}$$

The objective function is as follows:

$$L = \sum_i l(\widehat{y_i}, y_i) + \sum_k \Omega(f_k) \tag{9}$$

where l is the loss function, which represents the error between the predictive value and the true value, while Ω is the function used for regularization to prevent overfitting.

$$\Omega(f) = \gamma T + \frac{1}{2}\|\lambda w\|^2 \tag{10}$$

where T represents the number of leaves per tree, while w represents the weight of each tree's leaves.

After the second-order Taylor expansion of the objective function and other calculations, which are detailed in the supplementary materials, the information gain of the objective function after each split was obtained as follows:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum_{i\in IL} g_i)^2}{\sum_{i\in IL} h_i + \lambda} + \frac{(\sum_{i\in IR} g_i)^2}{\sum_{i\in IR} h_i + \lambda} + \frac{(\sum_{i\in I} g_i)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma. \tag{11}$$

To reveal the driving mechanism of various types of feature variables on the infection rate of COVID-19 in the seven stages, XGBoost was constructed using the optimal subset of features derived from SVM-RFE.

2.4.2.5. Accuracy evaluation

The dataset was randomly split into 70% for training and 30% for validation. Model accuracy and stability were evaluated by comparing the predicted and actual values using the R^2, mean absolute error (MAE), and RMSE. These metrics assess predictive accuracy and reliability, as previous studies have detailed (Chicco *et al* 2021). The mathematical expressions for these metrics are provided in Appendix D.

Beyond predictive accuracy, additional robustness analyses were conducted to assess the sensitivity of model results to feature multicollinearity and temporal stage delineation, as well as to evaluate spatial reliability using alternative statistical specifications. These robustness checks are described in sections 2.1.1 and 2.4.2.1, with detailed results reported in Appendix A.

2.4.3. Robustness validation using ordinary least squares (OLSs) and geographically weighted regression (GWR)

To verify the robustness and spatial interpretability of the machine learning results, OLS and GWR analyses were conducted for each infection stage. The OLS model served as a global benchmark, whereas the GWR model captured spatial non-stationarity by allowing coefficients to vary across locations. An adaptive bandwidth optimized using the golden-section search was applied to minimize the corrected Akaike information criterion (AICc), ensuring model parsimony and comparability among stages. Model evaluation focused on adjusted R^2, AICc, and ΔR^2 (GWR − OLS) to assess improvements in explanatory performance and spatial fit. All analyses were performed in ArcGIS Pro 2.5.2 using the key explanatory variables identified from the machine learning models. This procedure provided complementary evidence of global–local consistency and spatial heterogeneity, thereby supporting and extending the nonlinear relationships identified by the machine learning models. Detailed results are presented in appendix E (tables E.1–E.2; figures E.1–E.2).

2.4.4. Model interpretation approach

This study employed the following two complementary methods to interpret the model's results: SHAP and PDPs. SHAP is based on cooperative game theory. It quantifies each feature's contribution to predictions through Shapley values. By contrast, PDPs visualize the direction and magnitude of feature impacts on model predictions, revealing nonlinear interactions between features (Lundberg and Lee 2017, Greenwell 2017). The mathematical foundations of both methods are detailed in appendix F. The algorithm used in this study was based on the Model Interpretation Package developed in Python.

3. Results

3.1. Spatial patterns of urban morphology and socioeconomic factors

Before analyzing COVID-19 infection rates, the spatial characteristics of key explanatory variables were examined to contextualize the built-environment and socioeconomic background of the Tokyo Metropolitan Area. Built-environment indicators such as FAR and ROD exhibited clear core–periphery gradients across the study area, with higher values concentrated in municipalities located within the 23 special wards of central Tokyo—administrative units that constitute a subset of the municipalities analyzed in this study, particularly in Shinjuku, Minato, and Chuo, reflecting Tokyo's compact and multifunctional urban structure. In contrast, suburban municipalities in western Tokyo showed lower built-environment intensities but higher residential land shares. Socioeconomic indicators also displayed spatial polarization, with income and education levels being higher in the urban core, whereas the proportion of elderly residents and housing crowding were more prominent in peripheral and Tama areas. These spatial disparities establish the underlying urban morphology that shaped subsequent infection dynamics. Representative spatial distributions of FAR and ROD are illustrated in appendix G (figures G.1 and G.2).

3.2. Temporal-spatial distribution of COVID-19 infection rates per 100 000 population

According to COVID-19 infection data released by the Tokyo Metropolitan Government, 2866 240 positive cases had been recorded as of 25 September 2022. Figure 5 depicts the daily number of positive counts of COVID-19 in the Tokyo Metropolitan Area. The maximum number of daily infections exceeded 35 000 per day in August 2022. Between 1 April 2020 and 25 September 2022, the number of infections increased steadily. The spread of COVID-19 in Tokyo was highly volatile, with various policy

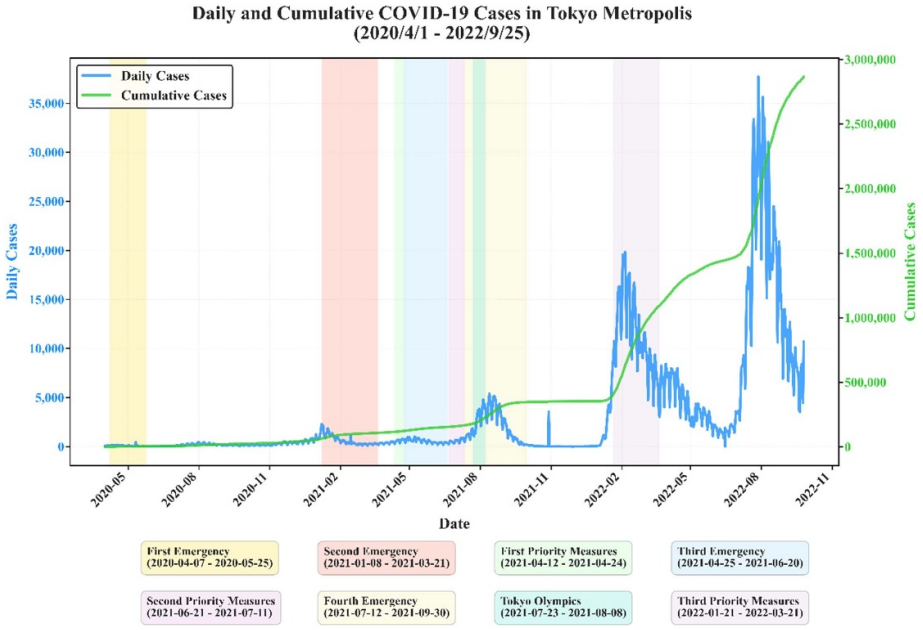

Figure 5. Temporal dynamics of the COVID-19 pandemic in the Tokyo Metropolitan Area. The line plot reveals new daily infection cases and government-imposed restriction policies in Japan (2020/4/1–2022/9/25).

Table 3. Global spatial autocorrelation report of COVID-19 infection rates in Tokyo.

Intervention phases	Moran's index	*Z* score	*P* value
First state of emergency (SIR1)	0.718 33	8.738	0.000
Second state of emergency (SIR2)	0.452 33	5.509	0.000
First priority measures (SIR3)	0.695 98	8.278	0.000
Third state of emergency (SIR4)	0.660 78	8.033	0.000
Second priority measures (SIR5)	0.809 43	9.585	0.000
Fourth state of emergency (SIR6)	0.694 03	8.334	0.000
Third priority measures (SIR7)	0.719 93	8.644	0.000

interventions impacting control of the pandemic; however, infection rates repeatedly rebounded significantly, which highlights the challenges of managing the pandemic.

To confirm the existence of spatial clustering, Global Moran's I analysis was conducted for each stage of the pandemic. Although spatial autocorrelation is expected given the contagious nature of COVID-19, this step provides a statistical baseline for subsequent multivariate analysis. The results in table 3 demonstrate significant global spatial autocorrelation of COVID-19 infection rates across all phases of the pandemic in Tokyo. High positive *Z*-Scores (ranging from 5.509 to 9.585) and *p* values of 0.000 indicate statistically significant spatial clustering throughout the study period. Notably, the *Z*-score during the second state of emergency (SIR2) is relatively lower than those observed in other stages, suggesting a temporary weakening—rather than an absence—of spatial clustering. This pattern likely reflects a more spatially dispersed distribution of infection rates across municipalities during this phase, compared with the more pronounced core–periphery concentration observed in later stages. The strongest spatial autocorrelation was observed during the second period of priority measures (SIR5), indicating a reintensification of spatial clustering during subsequent intervention phases.

The spatial patterns of COVID-19 infection rates per 100 000 population were analyzed across the seven infection phases in the Tokyo Metropolitan Area using Anselin Local Moran's I cluster maps (figure 6). Consistently, the analysis revealed significant spatial clustering throughout all seven phases, with distinct patterns observed between central and peripheral areas. High-high clusters (Here, 'high–high' clusters indicate municipalities with high COVID-19 infection rates that are surrounded by neighboring municipalities also exhibiting high infection rates, reflecting localized spatial autocorrelation rather than associations with specific explanatory variables.) were predominantly concentrated in the center of Tokyo's 23 special wards, particularly in areas like Shinjuku, Shibuya, and Minato, while low-low clusters were consistently observed in the northern and western suburban Tama areas as well as the mountainous Tama region. While the specific locations of infection hotspots exhibited some temporal

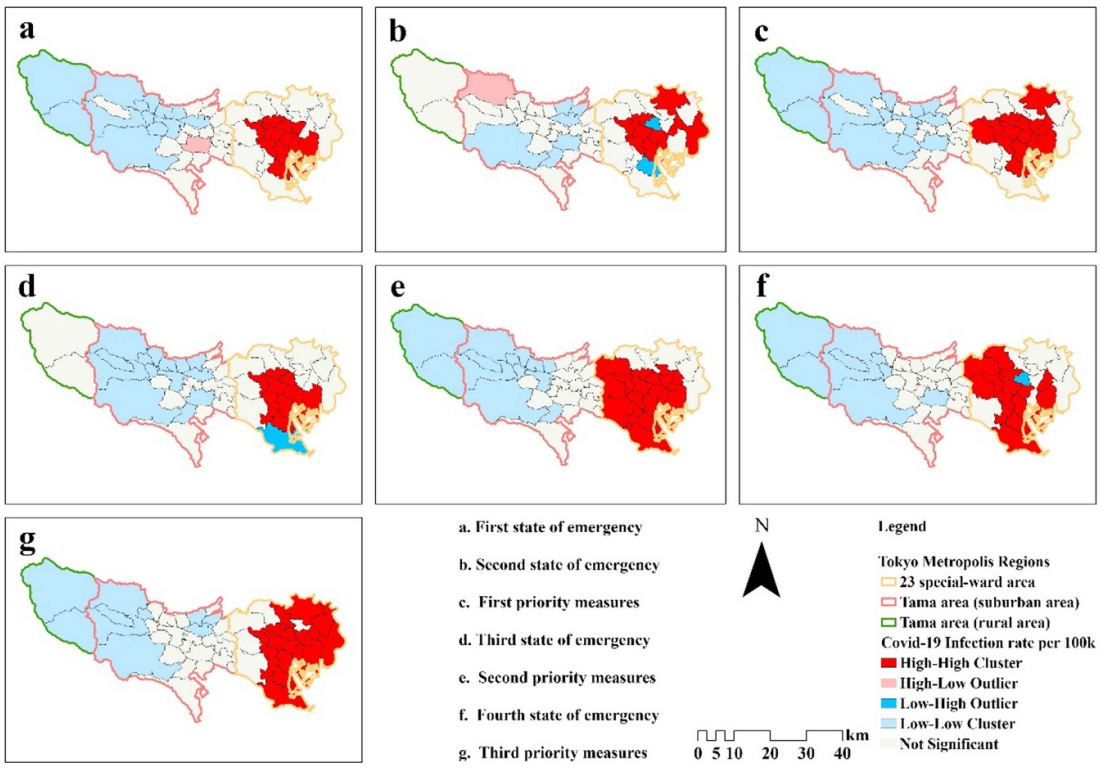

Figure 6. Anselin Local Moran's I of COVID-19 infection rates in the Tokyo Metropolitan Area during the seven phases.

variations across the different intervention phases (figures 6(a)–(g); detailed infection rate distributions presented in figure F.1), the general pattern of the central concentration remained stable. This indicates the strong influence of urban structural characteristics on infection transmission.

3.3. Influencing factor dynamics on COVID-19 infection rates during the seven phases

3.3.1. Evaluation of the three models

The average reciprocal RMSE and MAE values and R^2 were calculated for each model (detailed results in table I.1). The average scores for random forest, decision tree, and XGBoost were 0.404, 0.402, and 0.397, respectively; thus, as the random forest model had the best accuracy, it was selected to analyze the relationship between socioeconomic and urban morphology factors and COVID-19 infection rates in the seven stages.

3.3.2. Relative importance of explanatory variables

Figure 7 presents the key explanatory variables across the seven stages of the pandemic and their relative contributions, as calculated by the mean absolute SHAP values for each variable. On the left side, the variables are ranked by their global importance from highest to lowest, while the right side describes the contribution of each variable's value to the prediction of infection rates. In the right plot, the color of the point represents the high or low value of the variables, while the direction indicates whether a variable has a positive (SHAP $>$ 0) or negative (SHAP $<$ 0) value.

To assess the reliability of SHAP-based importance rankings, their stability was evaluated through multiple sensitivity analyses, including alternative model specifications and variable exclusion strategies. These analyses demonstrated a high degree of consistency in top-ranked variables across stages, with more than 90% overlap in leading predictors (appendix A tables A.3–A.4), supporting the robustness of the reported contribution patterns.

During the first state of emergency (SIR1), the primary influential factors were ROD, FAR, information and communication worker density (IC_WD), and IND. The SHAP value distributions indicated predominantly positive correlations for these variables in high-value regions (red dots).

During the SIR2, the percentage of foreigners (PCT_FOR), household crowding (HHC), percentage of the population aged 15–64 (PCT_15-64), and END emerged as dominant factors. Notably, the scatterplot for PCT_FOR exhibited a polarized SHAP distribution, with stronger positive contributions observed at higher values.

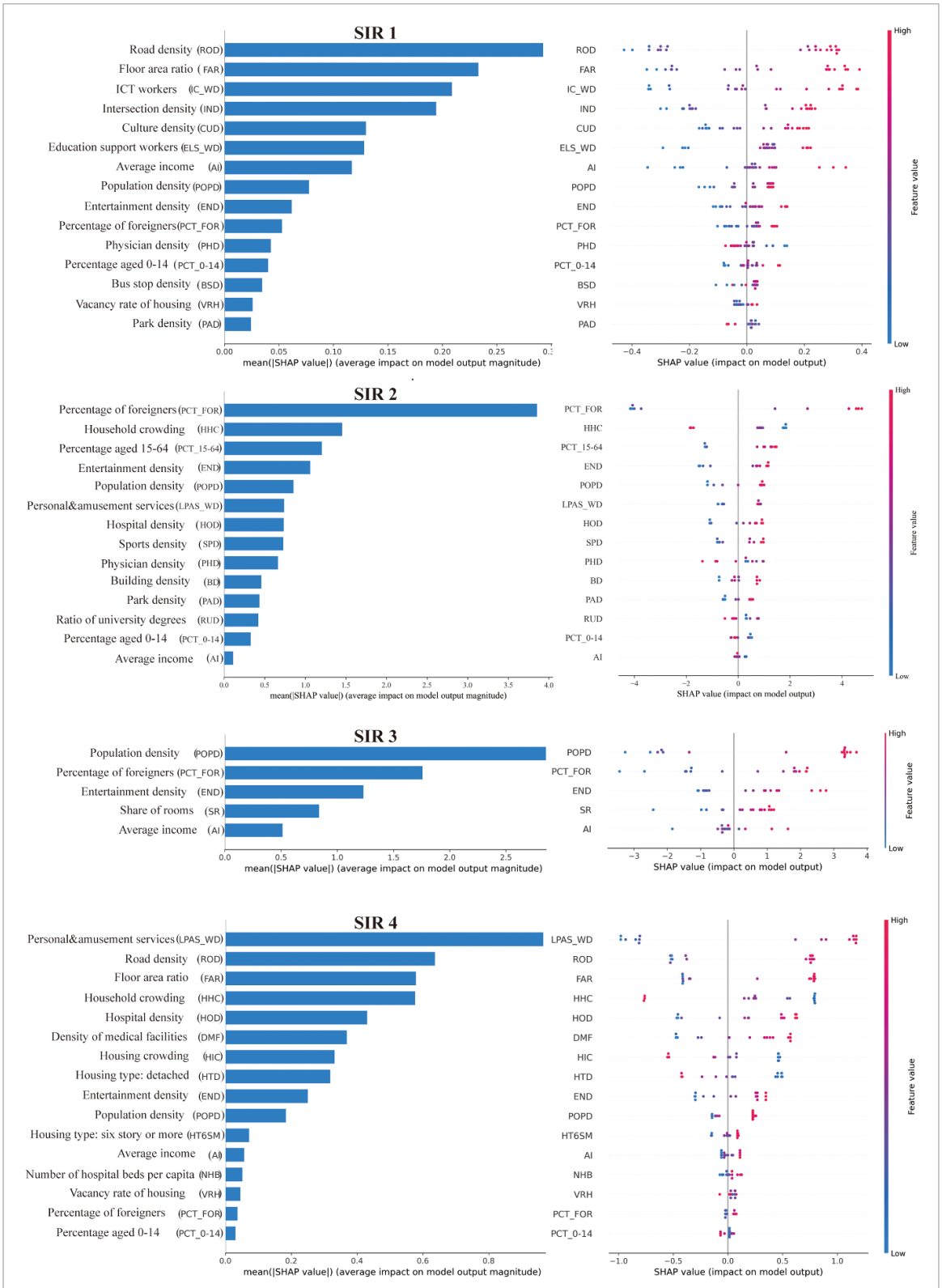

Figure 7. Relative importance of the explanatory variables and summary of local explanations. Variable labels use concise descriptive names with abbreviations; full definitions are provided in appendix B (table B.1).

During the first priority measures (SIR3), POPD, percentage of foreigners (PCT_FOR), END, and share of rooms (SR) ranked among the key contributors. The SHAP scatterplot demonstrated strong positive correlations for POPD in high-value regions (red dots).

The third state of emergency (SIR4) was characterized by the dominance of living-related and personal services as well as amusement service worker density (LPAS_WD), ROD, FAR, and HHC. The SHAP distribution for LPAS_WD exhibited relatively stronger positive SHAP contributions in high-value regions.

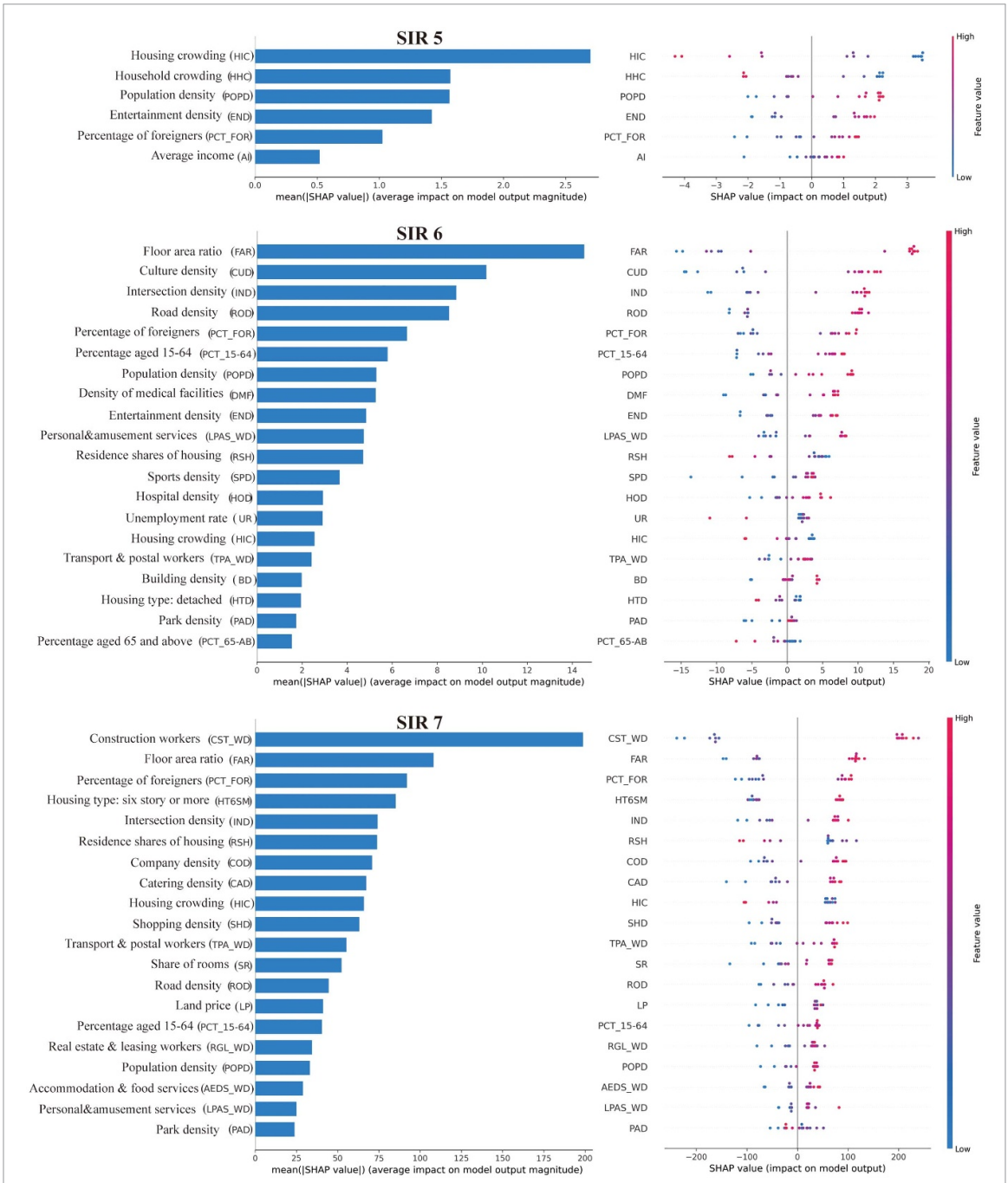

Figure 7. (Continued.)

During the second priority measures (SIR5), housing crowding (HIC), HHC, POPD, and END emerged as the key critical factors. Moreover, HIC's SHAP scatterplot exhibited substantial dispersion, which indicated the considerable variability of its influence.

The fourth state of emergency (SIR6) was dominated by FAR, CUD, IND, and ROD. FAR's scatterplot featured significantly positive effects in high-value regions.

During the third priority measures (SIR7), construction worker density (CST_WD), FAR, percentage of foreigners (PCT_FOR), and housing of six stories or more (HT6SM) ranked among the leading contributors. CST_WD's scatterplot distribution revealed the pronounced positive effects in high-value regions (red dots).

Across the seven pandemic stages, several variables consistently appeared among the key influential factors. FAR ranked among the top four factors in four stages—SIR1, SIR4, SIR6, and SIR7—with its

SHAP value distribution indicating an intensification from moderate effects in early stages (SIR1) to significantly positive influences in later stages (SIR6 and SIR7). The percentage of foreigners (PCT_FOR) was prominent in three stages—SIR2, SIR3, and SIR7—with its SHAP scatterplots consistently exhibiting distinct polarization and positive effects in high-value regions. ROD maintained significance across three stages—SIR1, SIR4, and SIR6—with its SHAP distributions indicating persistent positive effects in high-density areas. END also ranked among the top four in three stages—SIR2, SIR3, and SIR5—with its SHAP scatterplots suggesting sustained positive effects in high-density regions. This reflects the potential disease transmission risk associated with recreational venues.

By contrast, several built-environment and service-related variables consistently exhibited low SHAP contributions and did not enter the top-ranked predictors across most stages, suggesting limited independent explanatory power once dominant density, mobility, and occupation-related factors were accounted for.

3.3.3. Nonlinear relationship analysis

PDPs were employed to more accurately interpret the relationships between key urban morphology factors (e.g. FAR and ROD) and COVID-19 infection rates. These plots, extracted from the random forest model, revealed the marginal effects and thresholds of key variables at each infection stage while controlling for other factors, particularly in cases of nonlinear relationships. The horizontal axis represents the values of key variables while the vertical axis depicts their partial dependence on infection rates. Thus, the plots (figure 8) provide insights into the nonlinear dynamics between socioeconomic and urban morphology variables and COVID-19 infection rates.

The following descriptions explicitly guide the interpretation of PDP curves by detailing how the direction, slope, and threshold intensity of key variables evolve across pandemic stages.

During SIR1, ROD exhibited a distinct threshold effect at approximately 25 km km^{-2}, beyond which its impact stabilized. This significant threshold characteristic evolved into a gradual impact pattern within the range of 20–40 km km^{-2} in subsequent stages (e.g. SIR4 and SIR6). FAR exhibited a threshold at 0.25 in this stage with a gradual increase, which evolved into a more pronounced threshold effect at 0.3 in SIR6. Across pandemic stages, PDP results indicate a progressive strengthening of FAR-related effects. During early restriction phases (e.g. SIR1), FAR exhibited only modest increases in predicted infection risk with relatively flat response curves. In contrast, later stages (notably SIR6 and SIR7) displayed pronounced threshold-like increases, with sharper upward shifts emerging once FAR exceeded approximately 0.3–0.5 in later stages, suggesting that FAR-related exposure risks may become more pronounced under conditions of increased population mobility. Information and communication worker density (IC_WD) demonstrated a stable positive correlation within the range of 2.5–5.0 people/km^2.

During SIR2, the percentage of foreigners (PCT_FOR) displayed a significant threshold effect at approximately 1.5%, with the impact increasing substantially and then stabilizing beyond this value. This pattern maintained similar characteristics in SIR3 and reappeared in SIR7. Household crowding (HHC) began to exhibit a negative correlation at 0.5 people/household; the percentage of the population aged 15–64 (PCT_15-64) started to gradually increase at 50%; and END exhibited a stable positive correlation beyond 1.0 counts/km^2.

During SIR3, POPD demonstrated unique dual threshold characteristics at 400 and 600 people/km^2, revealing complex relationships between transmission dynamics and density. The SR began to exhibit a gradual increase at 0.15 rooms/household.

During SIR4, living-related and personal service worker density (LPAS_WD) exhibited a clear threshold at approximately 5 people/km^2. During SIR5, housing crowding (HIC) and HHC both began to significantly decline at 0.5 people/household, while END maintained a stable increase beyond 1.0 counts/km^2.

During SIR6, FAR was predominant with a significant threshold effect at 0.3, while CUD exhibited a gradual increase within the range of 1.0–2.0 counts/km^2, and IND demonstrated a slow increase beyond 2.0 intersections/km^2. Finally, during SIR7, construction worker density (CST_WD) exhibited a distinct threshold at 7.5 people/km^2, and housing with six stories or more (HT6SM) exhibited a stable increase beyond 10%.

Overall, these nonlinear effects reveal that the influence of built-environment and socioeconomic factors on infection rates evolved dynamically across pandemic stages.

To further confirm the robustness of these findings, additional statistical validation was conducted as follows.

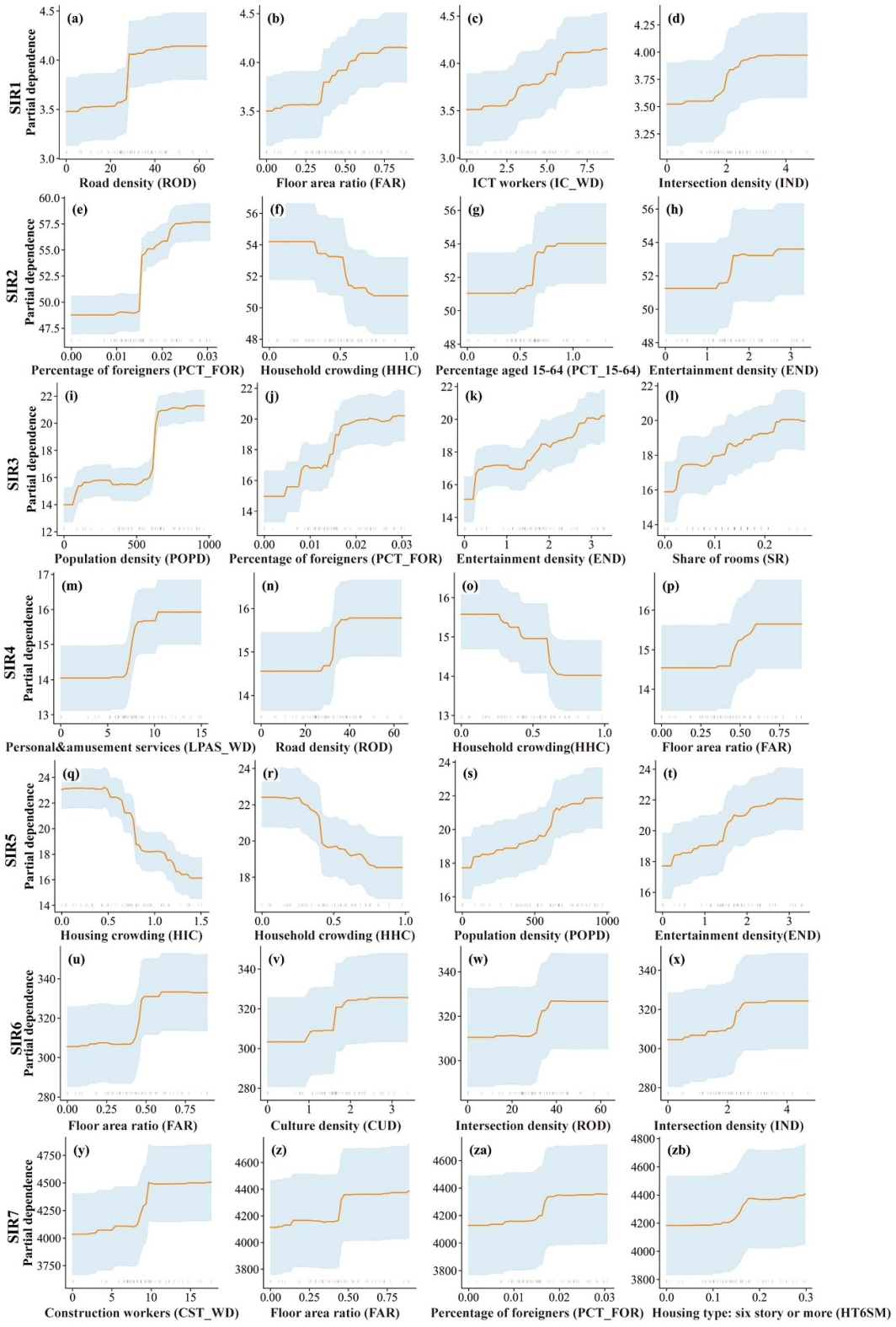

Figure 8. Partial dependence plots of four key explanatory variables on COVID-19 infection rates in SIR1-7.

3.2.4 Robustness validation of results

OLS and GWR analyses were conducted to assess the robustness and spatial validity of the machine learning results. As shown in table E.1, the GWR model achieved consistently higher adjusted R^2 values ($\Delta R^2 = +\ 0.05$–0.08) and lower AICc across all seven infection stages, confirming enhanced explanatory power and local model fit. The optimal neighborhood sizes (24–41) indicate that spatial interactions influencing infection rates mainly occurred within localized urban clusters. Spatial coefficient maps

(figure E.2) further revealed clear core–periphery gradients: positive effects of FAR and ROD were concentrated in central wards such as Shinjuku, Chuo, and Minato, while peripheral municipalities exhibited weaker or negative associations. Socioeconomic heterogeneity was also evident, as the proportion of foreign residents (PCT_FOR) showed stronger localized effects during SIR2, SIR5, and SIR7 (table E.2; figure E.2). These spatial variations highlight the localized nature of COVID-19 transmission dynamics across the Tokyo metropolitan area.

4. Discussion

4.1. Dynamic patterns and spatial distribution of COVID-19 infection rates

The findings indicate consistent spatial clustering of infection rates during each pandemic phase, particularly in the central areas of Tokyo's 23 special wards, such as Shinjuku, Minato, and Shibuya. This spatial clustering phenomenon is consistent with findings from studies in Melbourne and Berlin that have indicated that urban centers with a high POPD and frequent economic activities are more prone to infection hotspots (Gaisie *et al* 2022, Xu *et al* 2022).

During the various infection stages, the government implemented multiple prevention measures to control the spread of the virus, including states of emergency and priority measures. This study found these measures to be temporally associated with changes in the spatial distribution of infection rates. For example, during the first state of emergency, infection rates were primarily concentrated in central urban areas. As the measures were enforced, people's mobility significantly decreased, especially in densely populated residential and commercial districts. This finding aligns with those of Alidadi *et al* (2023), who noted that lockdown measures in urban core areas effectively slowed the spread of the pandemic. However, with the relaxation of policies, such as during the SIR2, infection rates rebounded in central districts and some suburban areas. Several factors may jointly contribute to this pattern, including the restoration of public transportation and commercial activities and potential declines in risk perception (Liu *et al* 2021). During this stage, this study observed a more pronounced spatial diffusion of infection rates, particularly in areas with convenient transportation and high POPD, such as Taito and Toshima.

These findings demonstrate that infection risk was not spatially static but evolved in response to policy interventions and behavioral adaptations. The persistence of hotspots in central Tokyo, despite multiple control phases, highlights the contextual association of disease transmission on the city's built form. Therefore, dividing the pandemic into distinct stages was methodologically necessary, as it allowed the identification of shifts in transmission mechanisms—from mobility-driven spread during strict control periods to socially driven diffusion during reopening. This stage-sensitive framework captures temporal variation in exposure contexts rather than causal transitions in transmission mechanisms.

4.2. Influencing socioeconomic and urban morphology factors

Through a machine learning model analysis, this study identified key socioeconomic and urban morphology factors that were associated with COVID-19 infection rates. Nonlinear relationships were revealed between these factors and infection rates, which have not previously been widely reported.

Urban morphology factors were found to exhibit strong associations with infection rate variation during the pandemic. ROD exhibited a consistent positive correlation with infection rates, with a notable threshold at 25 km km^{-2}. This finding aligns with those of studies from Florida and Hong Kong that have examined the role of transportation infrastructure in disease transmission (Hamidi *et al* 2020, Tepe 2023). However, unlike those studies' linear relationship observations, the present study identified distinct thresholds and temporal evolution patterns. A spatial visualization of Tokyo's road network density (figure G.1) indicates that areas that exceed this threshold are primarily concentrated in commercial hubs such as the Chiyoda, Chuo, Minato, and Shinjuku districts. This suggests that ROD may act as a proxy for activity intensity and mobility concentration rather than exerting a direct infrastructural effect (Guan *et al* 2023). Recent studies have further emphasized the association between ROD, economic activity, and mobility in Asian cities (Ma *et al* 2021, Meng *et al* 2024 Guo *et al* 2025).

The impact of FAR revealed dynamic threshold changes, with 0.25 observed during early stages and 0.3 in later stages. These thresholds being exceeded corresponded to Tokyo's mixed-use development zones, particularly in areas such as Shinjuku and Shibuya (figure G.2), which integrate commercial, residential, and transportation functions. These threshold values closely correspond to the 200%–300% FARs permitted by Tokyo's planning regulations, implying that these critical density levels are embedded in existing urban design standards (Tokyo Metropolitan Government 2023). Rather than implying a causal effect of FAR itself, these patterns suggest that built-form intensity conditions exposure opportunities under different mobility.

The combined interpretation of SHAP and PDP results also indicates a potential synergistic relationship between FAR and ROD—where compactness and connectivity jointly amplify exposure opportunities. Although formal interaction effects were not explicitly modeled, the identified nonlinearities imply interdependent mechanisms of density, accessibility, and contact intensity, which should be further examined in future SHAP interaction analyses.

During SIR6, the Tokyo Olympics coincided with increased associations between FAR, ROD, and infection rates near event venues. These patterns should be interpreted as contextual associations shaped by temporary functional intensification rather than direct effects of the built environment.

Among population-related factors, POPD exhibited dual thresholds at 400 and 600 people/km^2, which provides novel insights compared with a previous study (Kaufman *et al* 2021). The proportion of foreigners (PCT_FOR) exhibited a significant threshold at 1.5%, which reflects vulnerabilities associated with language barriers and limited access to medical services (Fukuchi *et al* 2022).

During SIR5, housing crowding (HIC) and HHC demonstrated negative correlations with infection rates. This is consistent with the findings of Alidadi *et al* (2023) but contradicts traditional transmission theories (Consolazio *et al* 2021, Nguyen *et al* 2021). This counterintuitive pattern may indicate behavioral adaptations such as voluntary isolation or reduced external mobility in crowded households.

The density of construction workers (CST_WD) only became significant during the Omicron-dominated SIR7 phase. Infection rates notably increased when the worker density exceeded 7.5 people/km^2 (Alidadi *et al* 2023), which is potentially due to high-density gatherings and enclosed work environments increasing contact frequency.

Notably, no significant association was found between green park space density and infection rates, which contrasts with the findings of Lin *et al* (2023) and Klompmaker *et al* (2021). Tokyo's unique spatial characteristics, such as its scattered, small-scale park distribution pattern, may weaken the practical impact of green space density. Additionally, Tokyo's extensive indoor public space network, such as transit-oriented development areas, may serve as partial substitutes for parks' social functions. This null finding underscores the importance of local urban context and cautions against generalizing green-space effects across cities.

Lastly, the SHAP and PDP analyses revealed these nonlinear dynamics, particularly in the interaction between FAR and ROD, which both exhibited significant threshold effects. Information and communication worker density (IC_WD) displayed stable positive correlations within the range of 2.5–5.0 people/km^2, illustrating selective temporal significance.

In summary, the observed results suggest that the dominant influencing factors shifted over time from physical density and mobility variables in early stages to occupational and social composition factors in later stages. This temporal evolution confirms the necessity of a stage-based analytical framework and provides evidence that urban morphology exerts heterogeneous effects under different epidemiological and behavioral contexts.

5.3. Policy implications and limitations

Based on this study's spatial analysis results and the PDP threshold characteristics, the findings highlight urban contexts where differentiated public health attention may be warranted. The analysis revealed that areas exceeding the initial risk thresholds (ROD $>$ 25 km km^{-2} and FAR $>$ 0.3) exhibit significant transmission risk, with these relationships becoming notably stronger at higher threshold levels. These thresholds should be interpreted as risk-sensitive reference points rather than operational planning targets. According to the observed spatial patterns and nonlinear associations, several policy-relevant implications can be discussed in a qualitative and context-dependent manner.

1. For areas characterized by very high density and connectivity (e.g. ROD $>$ 40 km km^{-2} or FAR $>$ 1.0), primarily concentrated in central Tokyo (e.g. Chiyoda, Chuo): These areas may benefit from enhanced situational monitoring, ventilation guidance, and public risk communication during epidemic surges, particularly in mixed-use and transit-oriented settings. Such measures should be considered in coordination with existing public health surveillance systems, rather than as standalone urban control instruments.
2. For areas with moderately high exposure contexts (e.g. 30 $<$ ROD $\leqslant$ 40 km km^{-2} or 0.5 $<$ FAR $\leqslant$ 1.0): Adaptive behavioral guidance (e.g. staggered work hours or voluntary teleworking) may help reduce peak exposure without imposing uniform restrictions. Targeted risk communication, including multilingual outreach, may improve information accessibility in socially diverse neighborhoods.

3. For moderate- exposure areas (e.g. 25 $<$ ROD $\leqslant$ 30 km km^{-2} or 0.3 $<$ FAR $\leqslant$ 0.5): Routine public health monitoring and community-level awareness may be sufficient, given the comparatively lower exposure intensity.

By distinguishing stage-specific exposure contexts, this study suggests that the relevance of built-environment and socioeconomic factors varies across epidemic phases, with mobility and density-related associations more pronounced during early restriction periods and occupational or social composition factors becoming more salient during later stages. These insights are intended to inform adaptive risk awareness rather than prescribe fixed planning thresholds.

While this study provides valuable insights into COVID-19 transmission, it has several limitations. First, the analysis relied on data from 53 municipalities within a single metropolitan region. This limited sample size may constrain statistical power and limits the generalizability of the findings to cities with different urban structures, governance regimes, or mobility systems. Second, due to data availability constraints, certain micro-level behavioral factors (e.g. individual mobility trajectories, compliance intensity) were not incorporated. Third, although quantitative sensitivity analyses (appendix A tables A.3–A.4) demonstrate that the main findings are robust to alternative model specifications and variable exclusion strategies, the identified nonlinear thresholds should be interpreted as context-dependent patterns rather than universally transferable cutoffs. Moreover, these sensitivity analyses primarily assess predictive robustness and ranking stability, and do not fully resolve potential endogeneity or reverse causality between urban form, socioeconomic structure, and infection risk. Finally, given the relatively small sample size, more computationally intensive uncertainty estimation approaches—such as extensive bootstrapping of SHAP values or formal interaction testing—were not feasible.

Future research should combine built-environment indicators with policy stringency indices, healthcare capacity, and human mobility data to strengthen causal inference. Specifically, future studies could test the hypothesis that the marginal effects of built density indicators (e.g. FAR and ROD) interact positively with mobility levels and negatively with policy stringency—such that density-related transmission risks are amplified under high-mobility, low-restriction conditions but attenuated under strict mobility controls. This interaction framework would help disentangle contextual built-environment effects from policy-modulated behavioral responses.

5. Conclusions

This study examined the spatiotemporal patterns of COVID-19 infection rates across the Tokyo Metropolitan Area using an interpretable, stage-based analytical framework. Consistent spatial clustering was observed throughout all pandemic stages, with recurrent hotspots in central wards and persistently lower rates in peripheral municipalities, indicating stable spatial inequalities in infection risk.

Using random forest models with SHAP and partial dependence diagnostics, we identified stage-specific, nonlinear associations between area-level urban form, socioeconomic indicators, and infection rates. Several built-environment variables, including ROD, FAR, and POPD, exhibited threshold-like patterns, while occupational composition indicators (e.g. construction worker density) became more relevant during later, Omicron-dominated stages. These results highlight that the relevance and marginal patterns of risk correlates are not temporally stable, but evolve with intervention phases and behavioral context.

Importantly, the findings should be interpreted as associational rather than causal, reflecting contextual exposure conditions at the municipal level. For environmental epidemiology and public health practice, the results suggest that spatially explicit surveillance and stage-sensitive interpretation of area-level indicators may improve situational awareness during epidemic waves, compared with assuming uniform or linear effects over time.

Several limitations warrant attention. The analysis is restricted to a single metropolitan area and an aggregated spatial scale, and residual confounding related to mobility, policy stringency, and healthcare capacity cannot be excluded. Future studies should integrate mobility indicators and policy response metrics to test whether built-environment associations are amplified under high-mobility conditions and attenuated under stricter controls, thereby supporting post-pandemic monitoring of urban infectious disease risks.

Acknowledgment

This study was supported by the Japan Society for the Promotion of Science (Project Number: JPJS00124016566) and JST SPRING (Grant Number: JPMJSP2136).

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary information files).

Supplementary material available at http://doi.org/10.1088/2752-5309/ae44c1/data1.

Author contributions

Yangguang Xiao
Conceptualization (equal), Data curation (equal), Formal analysis (equal), Funding acquisition (equal), Investigation (equal), Methodology (equal), Software (equal), Visualization (equal), Writing – original draft (equal)

Kojiro Sho 0000-0002-5059-4332
Conceptualization (equal), Funding acquisition (equal), Methodology (equal), Project administration (equal), Resources (equal), Supervision (equal), Validation (equal), Writing – original draft (equal)

Yuya Shibuya
Methodology (equal), Validation (equal), Writing – review & editing (equal)

Kimihiro Hino 0000-0003-1243-1329
Methodology (equal), Validation (equal), Writing – review & editing (equal)

Shichen Zhao
Resources (equal), Supervision (equal)

Ronita Bardhan 0000-0001-5336-4084
Methodology (equal), Validation (equal), Writing – review & editing (equal)

References

Ali T, Mortula M and Sadiq R 2021 GIS-based vulnerability analysis of the United States to COVID-19 occurrence *J. Risk Res.* **24** 416–31
Alidadi M and Sharifi A 2022 Effects of the built environment and human factors on the spread of COVID-19: a systematic literature review *Sci. Total Environ.* **850** 158056
Alidadi M, Sharifi A and Murakami D 2023 Tokyo's COVID-19: an urban perspective on factors influencing infection rates in a global city *Sustain. Cities Soc.* **97** 104743
Alita D, Setiawansyah S and Putra A D 2021 C45 algorithm for motorcycle sales prediction on CV Mokas Rawajitu *J. Infotek Glob.* **11** 127–34
Anselin L 1995 Local indicators of spatial association—LISA *Geogr. Anal.* **27** 93–115
Azzopardi-Muscat N, Brambilla A, Caracci F and Capolongo S 2020 Synergies in design and health: the role of architects and urban health planners in tackling key contemporary public health challenges *Acta Biomed.* **91** 9–20
Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32
Chandrashekar G and Sahin F 2014 A survey on feature selection methods *Comput. Electr. Eng.* **40** 16–28
Chen E, Ye Z and Wu H 2021 Nonlinear effects of built environment on intermodal transit trips considering spatial heterogeneity *Transp. Res. D* **90** 102677
Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system *Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* (ACM) pp 785–94
Chicco D, Warrens M J and Jurman G 2021 The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation *PeerJ Comput. Sci.* **7** e623
Consolazio D, Murtas R, Tunesi S, Gervasi F, Benassi D and Russo A G 2021 Assessing the impact of individual characteristics and neighborhood socioeconomic status during the COVID-19 pandemic in the provinces of Milan and Lodi *Int. J. Health Serv.* **51** 311–24
Cruz J, Mamani W, Romero C and Pineda F 2021 Selection of characteristics by hybrid method: RFE, ridge, lasso, and Bayesian for the power forecast for a photovoltaic system *SN Comput. Sci.* **2** 202
Dadashpoor H and Alidadi M 2017 Towards decentralization: spatial changes of employment and population in Tehran Metropolitan Region, Iran *Appl. Geogr.* **85** 51–61
Dall'erba S 2005 Distribution of regional income and regional funds in Europe 1989–1999: an exploratory spatial data analysis *Ann. Reg. Sci.* **39** 121–48
Didavi A B K, Agbokpanzo R G and Agbomahena M 2021 Comparative study of decision tree, random forest and XGBoost performance in forecasting the power output of a photovoltaic system *4th Int. Conf. on Bio-Engineering for Smart Technologies (Biosmart)* pp 1–5
Ewing R and Cervero R 2010 Travel and the built environment *J. Am. Plan. Assoc.* **76** 265–94
Friedman J H 2001 Greedy function approximation: a gradient boosting machine *Ann. Stat.* **29** 1189–536
Fukuchi H, Uehara W, Kamata H and He G 2022 COVID-19 policies and hoteliers' responses in Japan *Ann. Tour. Res. Empir. Insights* **3** 100067
Gaisie E, Oppong-Yeboah N Y and Cobbinah P B 2022 Geographies of infections: built environment and COVID-19 pandemic in metropolitan Melbourne *Sustain. Cities Soc.* **81** 103838

Greenwell B 2017 pdp: an R package for constructing partial dependence plots *The R Journal* **9** 421–36
Guan C, Tan J, Li Y, Cheng T, Yang J, Liu C and Keith M 2023 How do density, employment and transit affect the prevalence of COVID-19 pandemic? *Health Place* **84** 103117
Guida C and Carpentieri G 2021 Quality of life in the urban environment and primary health services for the elderly during the COVID-19 pandemic: an application to the city of Milan (Italy) *Cities* **110** 103038
Guo W, Wang J, Liu X, Pan Z, Zhuang R, Li C and Tang H 2025 Towards resilient communities: evaluating the nonlinear impact of the built environment on COVID-19 transmission risk in residential areas *Build. Environ.* **267** 112289
Hamidi S and Hamidi I 2021 Subway ridership, crowding, or population density: determinants of COVID-19 infection rates in New York City *Am. J. Prev. Med.* **60** 614–20
Hamidi S, Sabouri S and Ewing R 2020 Does density aggravate the COVID-19 pandemic? Early findings and lessons for planners *J. Am. Plan. Assoc.* **86** 495–509
Hu M, Roberts J D, Azevedo G P and Milner D 2021 The role of built and social environmental factors in COVID-19 transmission: a look at America's capital city *Sustain. Cities Soc.* **65** 102580
Huang X, Lu G, Yin J and Tan W 2021a Non-linear associations between the built environment and the physical activity of children *Transp. Res. D* **98** 102968
Huang X, Yang Q and Yang J 2021b Importance of community containment measures in combating the COVID-19 epidemic: from the perspective of urban planning *Geo-Spat. Inf. Sci.* **24** 363–71
Karako K, Song P, Chen Y and Karako T 2022 COVID-19 in Japan during 2020-2022: Characteristics, responses, and implications for the health care system *J. Glob. Health.* **12**
Karaye I M and Horney J A 2020 The impact of social vulnerability on COVID-19 in the U.S.: an analysis of spatially varying relationships *Am. J. Prev. Med.* **59** 317–25
Kashem S B, Baker D M, González S R and Lee C A 2021 Exploring the nexus between social vulnerability, built environment, and the prevalence of COVID-19: a case study of Chicago *Sustain. Cities Soc.* **75** 103261
Kaufman H W, Niles J K and Nash D B 2021 Disparities in SARS-CoV-2 positivity rates: associations with race and ethnicity *Popul. Health Manage.* **24** 20–26
Kim S and Lee S 2023 Nonlinear relationships and interaction effects of an urban environment on crime incidence: application of urban big data and an interpretable machine learning method *Sustain. Cities Soc.* **91** 104419
Klompmaker J O, Hart J E, Holland I, Sabath M B, Wu X, Laden F, Dominici F and James P 2021 County-level exposures to greenness and associations with COVID-19 incidence and mortality in the United States *Environ. Res.* **199** 111331
Lai K Y, Webster C, Kumari S and Sarkar C 2020 The nature of cities and the COVID-19 pandemic *Curr. Opin. Environ. Sustain.* **46** 27–31
Lak A, Sharifi A, Badr S, Zali A, Maher A, Mostafavi E and Khalili D 2021 Spatio-temporal patterns of the COVID-19 pandemic and place-based influential factors at the neighborhood scale in Tehran *Sustain. Cities Soc.* **72** 103034
Li X, Zhou L, Jia T, Peng R, Fu X and Zou Y 2020 Associating COVID-19 severity with urban factors: a case study of Wuhan *Int. J. Environ. Res. Public Health* **17** 6712
Liao Q, Dong M, Yuan J, Fielding R, Cowling B J, Wong I O L and Lam W W T 2021 Assessing community vulnerability over 3 waves of COVID-19 pandemic, Hong Kong, China *Emerg. Infect. Dis.* **27** 1935–9
Lin J, Huang B, Kwan M-P, Chen M and Wang Q 2023 COVID-19 infection rate but not severity is associated with availability of greenness in the United States *Landsc. Urban Plan.* **233** 104704
Liu C, Liu Z and Guan C 2021 The impacts of the built environment on the incidence rate of COVID-19: a case study of King County, Washington *Sustain. Cities Soc.* **74** 103144
Liu L 2020 Emerging study on the transmission of the novel coronavirus (COVID-19) from urban perspective: evidence from China *Cities* **103** 102759
Lundberg S M and Lee S-I 2017 A unified approach to interpreting model predictions *Adv. Neural Inf. Process. Syst.* **30** 4765–74 (available at: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html)
Ma S, Li S and Zhang J 2021 Diverse and nonlinear influences of built environment factors on COVID-19 spread across townships in China at its initial stage *Sci. Rep.* **11** 12415
Marselle M R, Lindley S J, Cook P A and Bonn A 2021 Biodiversity and health in the urban environment *Curr. Environ. Health Rep.* **8** 146–56
Meng Y, Ho H C and Wong M S 2024 Changing associations of built environment with usage of urban space due to the COVID-19 pandemic in the United States *Cities* **152** 105205
Nagata S, Nakaya T, Adachi Y, Inamori T, Nakamura K, Arima D and Nishiura H 2021 Mobility change and COVID-19 in Japan: mobile data analysis of locations of infection *J. Epidemiol.* **31** 387–91
Nasiri R, Akbarpour S, Zali A R, Khodakarami N, Boochani M H, Noory A R and Soori H 2022 Spatio-temporal analysis of COVID-19 incidence rate using GIS: a case study—Tehran metropolitan, Iran *GeoJournal* **87** 3291–305
Nguyen T H, Shah G H, Schwind J S and Richmond H L 2021 Community characteristics and COVID-19 outcomes: a study of 159 counties in Georgia, United States *J. Public Health Manage. Pract.* **27** 251–7
Pan Y, Zhang L, Yan Z, Lwin M O and Skibniewski M J 2021 Discovering optimal strategies for mitigating COVID-19 spread using machine learning: experience from Asia *Sustain. Cities Soc.* **75** 103254
Prajwala T R 2015 A comparative study on decision tree and random forest using R tool *Int. J. Adv. Res. Comput. Commun. Eng.* **4** 196–9
Ren H, Zhao L, Zhang A, Song L, Liao Y, Lu W and Cui C 2020 Early forecasting of the potential risk zones of COVID-19 in China's megacities *Sci. Total Environ.* **729** 138995
Sharifi A 2022 An overview and thematic analysis of research on cities and the COVID-19 pandemic: toward just, resilient, and sustainable urban planning and design *iScience* **25** 105297
Sharifi A and Khavarian-Garmsir A R 2020 The COVID-19 pandemic: impacts on cities and major lessons for urban planning, design, and management *Sci. Total Environ.* **749** 142391
Shoari N, Ezzati M, Baumgartner J, Malacarne D, Fecht D and Ramagopalan S V 2020 Accessibility and allocation of public parks and gardens in England and Wales: a COVID-19 social distancing perspective *PLoS One* **15** e0241102
Si Z, Yang M, Yu Y and Ding T 2021 Photovoltaic power forecast based on satellite images considering effects of solar position *Appl. Energy* **302** 117514
Tepe E 2023 The impact of built and socio-economic environment factors on COVID-19 transmission at the ZIP-code level in Florida *J. Environ. Manage.* **326** 116806
Tokyo COVID-19 Task Force website n.d. (available at: stopcovid19.metro.tokyo.lg.jp/)

Tokyo Metropolitan Government 2022 COVID-19 information in Tokyo. Tokyo Metropolitan Government (available at: www.metro.tokyo.lg.jp/english/about/index.html)
Tokyo Metropolitan Government 2023 Chapter 3: city planning—building use districts, FAR and BCR restrictions (Tokyo Metropolitan Government) (available at: www.toshiseibi.metro.tokyo.lg.jp/documents/d/toshiseibi/pdf_keikaku_chousa_singikai_pdf_keikaku_en_03)
Twohig K A *et al* 2022 Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study *Lancet Infect. Dis.* **22** 35–42
Ulubaş Hamurcu A and Yılmaz M 2023 Geostatistical assessment of the built environment and spatio-temporal distribution patterns of COVID-19 cases in Istanbul, Türkiye *Build. Environ.* **243** 110666
Venter Z S, Barton D N, Gundersen V, Figari H and Nowell M 2020 Urban nature in a time of crisis: recreational use of green space increases during the COVID-19 outbreak in Oslo, Norway *Environ. Res. Lett.* **15** 104075
Wang J and Zha Y 2024 Do urban form characteristics perpetuate disparities of pandemic-induced mobility changes? Evidence from Fulton County, Georgia *Travel Behav. Soc.* **36** 100803
Wang L, Zhang S, Yang Z, Zhao Z, Moudon A V, Feng H, Liang J, Sun W and Cao B 2021 What county-level factors influence COVID-19 incidence in the United States? Findings from the first wave of the pandemic *Cities* **118** 103396
Williamson E J *et al* 2020 Factors associated with COVID-19-related death using OpenSAFELY *Nature* **584** 430–6
Xiao L and Liu J 2023 Exploring non-linear built environment effects on urban vibrancy under COVID-19: the case of Hong Kong *Appl. Geogr.* **155** 102960
Xiao L, Lo S, Liu J, Zhou J and Li Q 2021 Nonlinear and synergistic effects of TOD on urban vibrancy: applying local explanations for gradient boosting decision tree *Sustain. Cities Soc.* **72** 103063
Xu G, Jiang Y, Wang S, Qin K, Ding J, Liu Y and Lu B 2022 Spatial disparities of self-reported COVID-19 cases and influencing factors in Wuhan, China *Sustain. Cities Soc.* **76** 103485
Yabe T, Tsubouchi K, Fujiwara N, Wada T, Sekimoto Y and Ukkusuri S V 2020 Non-compulsory measures sufficiently reduced human mobility in Tokyo during the COVID-19 epidemic *Sci. Rep.* **10** 18053
Yang W, Li Y, Liu Y, Fan P and Yue W 2024 Environmental factors for outdoor jogging in Beijing: insights from using explainable spatial machine learning and massive trajectory data *Landsc. Urban Plan.* **243** 104969
Yip T L, Huang Y and Liang C 2021 Built environment and the metropolitan pandemic: analysis of the COVID-19 spread in Hong Kong *Build. Environ.* **188** 107471
You H, Wu X and Guo X 2020 Distribution of COVID-19 morbidity rate in association with social and economic factors in Wuhan, China: implications for urban development *Int. J. Environ. Res. Public Health* **17** 3417
Zheng R, Xu Y, Wang W, Ning G and Bi Y 2020 Spatial transmission of COVID-19 via public and private transportation in China *Travel Med. Infect. Dis.* **34** 101626