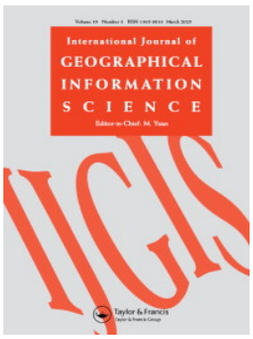

International Journal of Geographical Information Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tgis20

Understanding of the predictability and uncertainty in population distributions empowered by visual analytics

Peng Luo, Chuan Chen, Song Gao, Xianfeng Zhang, Deng Majok Chol, Zhuo Yang & Liqiu Meng

To cite this article: Peng Luo, Chuan Chen, Song Gao, Xianfeng Zhang, Deng Majok Chol, Zhuo Yang & Liqiu Meng (2025) Understanding of the predictability and uncertainty in population distributions empowered by visual analytics, International Journal of Geographical Information Science, 39:3, 675-705, DOI: 10.1080/13658816.2024.2427870

To link to this article: https://doi.org/10.1080/13658816.2024.2427870

Published online: 19 Nov 2024.

INTERNATIONAL JOURNAL OF GEOGRAPHICAL INFORMATION SCIENCE
2025, VOL. 39, NO. 3, 675–705
https://doi.org/10.1080/13658816.2024.2427870

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Understanding of the predictability and uncertainty in population distributions empowered by visual analytics

Peng Luo[a,b], Chuan Chen[b], Song Gao[c], Xianfeng Zhang[d], Deng Majok Chol[e], Zhuo Yang[b] and Liqiu Meng[b]

[a]Senseable City Lab, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA; [b]Chair of Cartography and Visual Analytics, Technical University of Munich, Munich, Germany; [c]Department of Geography, University of Wisconsin-Madison, Madison, Wisconsin, USA; [d]Institute of Remote Sensing and Geographical Information Systems, Peking University, Beijing, China; [e]Environmental Change Institute, University of Oxford, Oxford, UK

ABSTRACT

Understanding the intricacies of fine-grained population distribution, including both predictability and uncertainty, is crucial for urban planning, social equity, and environmental sustainability. The spatial processes associated with the distribution of populations are complex, and enhancing their predictability involves revealing nonlinear interactions among various explanatory variables. Additionally, population distribution is influenced by various factors that are often challenging to quantify, thereby introducing uncertainty into predictive models. Although the development of explainable artificial intelligence (XAI) helps identify underlying factors, the complex geographical processes and the special nature of spatial data present challenges for purely statistical-based explanation methods, leading to incomplete or incorrect explanations. To address these challenges, we introduce GeoVisX, a geospatial visual analytics framework integrated with XAI. GeoVisX integrates XAI with visual analytics to dissect the spatial processes. Through a case study of Munich, GeoVisX demonstrates its utility in analyzing spatial distribution and identifying key factors impacting population distribution at the 100 m grid level. Our findings highlight the GeoVisX's capability to enhance understanding of geographical phenomena, contributing to more informed urban policy and planning strategies. This study not only validates the effectiveness of GeoVisX but also emphasizes the importance of incorporating visual analytics and explainable methodologies for addressing complex geographical issues.

ARTICLE HISTORY
Received 5 July 2024
Accepted 5 November 2024

KEYWORDS
Population distribution; explainable artificial intelligence; visual analytics

CONTACT Chuan Chen chuan.chen@tum.de

1. Introduction

Revealing the mechanism of population distribution plays a critical role in urban planning and development, socio-economic analysis, and environmental management (Maantay *et al*. 2007, Luo *et al*. 2019, Li *et al*. 2021). It significantly contributes to optimizing urban infrastructure, promoting economic growth, achieving social equity, and ensuring environmental sustainability (Langford *et al*. 2008). The distribution mechanism of populations in space is highly complex, influenced by a multitude of interacting factors, such as economic opportunities (Mason 2001) and living environmental conditions (Abdul Salam *et al*. 2014). This mechanism can be explained by a series of explanatory variables, such as indicators of economic development and environment (Wesolowski *et al*. 2012), as well as the distributions of urban sub-centers (Huang *et al*. 2017) and residential areas (De Jong and Sell 1977). Given the different contributions and impacts of various explanatory variables, population distribution can broadly be predicted, leading to the development of numerous methods for estimating spatial population distributions (Deville *et al*. 2014).

On the other hand, the choice of residential areas often involves characteristics that are difficult to explain with available explanatory variables (Larson *et al*. 2009), thus making predictions challenging and leading to uncertainty in population mapping results. From an ethical perspective, uncertainty can result in overestimations or underestimations in population mapping tasks, leading to suboptimal spatial decisions. For example, incorrect estimates of population distribution when setting up new transportation hubs or public services can lead to spatial injustice for residents in certain areas (Currie 2004).

Exploring the predictability and uncertainty of population distribution is crucial for two reasons. Firstly, it helps deepen our understanding of human behavior, urban structure, and socio-economic development, revealing universal laws of population distribution, thereby expanding the geographic generalizability of models and enhancing future prediction capabilities. Secondly, despite the availability of advanced algorithms and models for estimating population distribution, such as dasymetric mapping (Mennis 2003, Liu *et al*. 2020, Song *et al*. 2024) and the weighting method (Doxsey-Whitfield *et al*. 2015), and corresponding datasets (e.g. GPW,[1] LandScan,[2] WorldPop[3]), the uncertainties in prediction models and results have not been thoroughly discussed. Investigating the causes of inaccuracies in population forecasts and the extent to which these results can be trusted is essential, as these inaccuracies could significantly impact urban planning, economic policies, and environmental protection. A detailed discussion and analysis of these errors can improve the accuracy of prediction models and enhance the reliability of decision-making.

Explainable machine learning helps in understanding the spatial processes of population distribution, including their predictability and uncertainty (Cheng *et al*. 2021, Hsu and Li 2023, Xing and Sieber 2023, Liu *et al*. 2024). For instance, Shapley values have been proposed to effectively decompose the complex influences of individual variables (Lundberg *et al*. 2020, Li *et al*. 2023b). Specifically, by building a predictive machine learning model using a series of explanatory variables to predict a target variable, explainable machine learning methods can decompose the contributions of different explanatory variables, helping us understand the spatial processes of the target

variable and its relationship with other variables (Deb and Smith 2021, Ji *et al*. 2022, Li *et al*. 2023). It is important to clarify that the identification of the spatial processes for a geographic variable mentioned in this work originates from quantitative human geography (Fotheringham and Sachdeva 2022), referring to the discovery of other geographic factors influencing the spatial distribution of that variable. From a statistical perspective, these factors are those that, based on a certain model, can replicate the geographic distribution with a high probability (Fischer and Wang 2011). These variables may either have a causal relationship with the subject of study or simply exhibit correlation (Hay and Johnston 1983). They can be identified by different statistical methods, such as regression coefficients in traditional regression models, variable importance in machine learning models (e.g., variance importance in random forest models), or metrics in XAI models (e.g., SHAP values).

To comprehend the spatial processes of fine-grained population distribution, it is necessary to construct population prediction models (Patel *et al*. 2017). Fine-grained population estimation studies typically explore various population-related explanatory variables with high spatial detail, combined with administrative units containing census data to establish the relationship between population numbers and explanatory variables (Bakillah *et al*. 2014). This relationship is then used to predict fine-grained population distribution. Commonly used data include remote sensing data and social sensing data, such as Nighttime Light (NTL) images (Wang *et al*. 2020), Normalized Difference Vegetation Index (NDVI) images, Land Surface Temperature (LST) images (Luo *et al*. 2019), Points of Interest (POI) data (Ye *et al*. 2019), and social media check-in data (Yao *et al*. 2017). In terms of model selection, early studies mainly used traditional statistical regression models, such as multiple linear regression. However, due to the often complex non-linear relationships between population distribution and explanatory variables, machine learning algorithms have become more popular in recent years (Stevens *et al*. 2015, Xing *et al*. 2020, Huang *et al*. 2021).

By incorporating an explainable machine learning framework into fine-grained population prediction methods, it is possible to dissect the elements influencing the predictability and uncertainty of population distribution. However, two issues still need to be addressed, which are important not only for explaining the spatial processes of population distribution but also for improving the explainability of most geographical phenomena:

First, for geographical issues, it is insufficient to understand spatial processes in numerical terms. The spatial patterns of geographical variables are crucial for understanding geographical issues. For example, individuals choosing to live in suburban areas may have higher expectations regarding the surrounding natural environment, such as air quality and accessibility to scenic areas, compared to those residing in urban areas. Therefore, understanding the spatial processes of population distribution requires more than just global explanations; it is essential to analyze the spatial variability of these spatial processes.

Second, when choosing residential locations, people consider not only the characteristics of the location but also the surrounding natural environment and facilities, that is, the spatial context of attributes. For example, for people with certain

environmental demands, the green space within a residential area may not be the most crucial factor provided that there are parks within a certain range of the home.

In light of the above background and challenges, we propose GeoVisX: geospatial visual analytics framework integrated with XAI. In the identification of spatial processes, statistical analysis (black lines in Figure 1) involves both data modeling and model interpretation (e.g., regression coefficients or XAI). However, such methods often struggle to reveal the spatially varying results and are prone to misinterpretation (as we will discuss in Section 3.1). Visual analysis is a process that leverages visual perception and cognition to analyze and interpret data. In the geographical context, visual analysis is an important method for analyzing the spatial patterns of geographic variables and can serve as a valuable complement to explainable machine learning (Andrienko *et al*. 2010). Visual analysis enables people to gain an intuitive understanding of the spatial patterns of geographic variables (blue lines at Figure 1). In the exploration of spatial processes, visualization is primarily employed to analyze the spatial distribution of geographic variables to identify their valuable patterns. However, fewer studies focus on visualizing the results of statistical models, such as regression coefficients or outputs from explainable machine learning models. This limited approach makes it challenging to generate new insights into the interpretation of models. Explainable machine learning, on the other hand, is used to uncover the reasoning and mechanisms behind model decisions that lead to these spatial patterns. While the former focuses on intuitive comprehension of the geographical data through human perception, the latter emphasizes the logical interpretation of the model. Therefore, combining these two approaches can enhance our understanding of the decision-making process in geographic models and deepen our insights into the spatial distribution of geographic variables (red lines in Figure 1). Such a visualization framework allows people to infer correlations not only based on the spatial distribution of geographic variables but also to avoid erroneous interpretations derived from

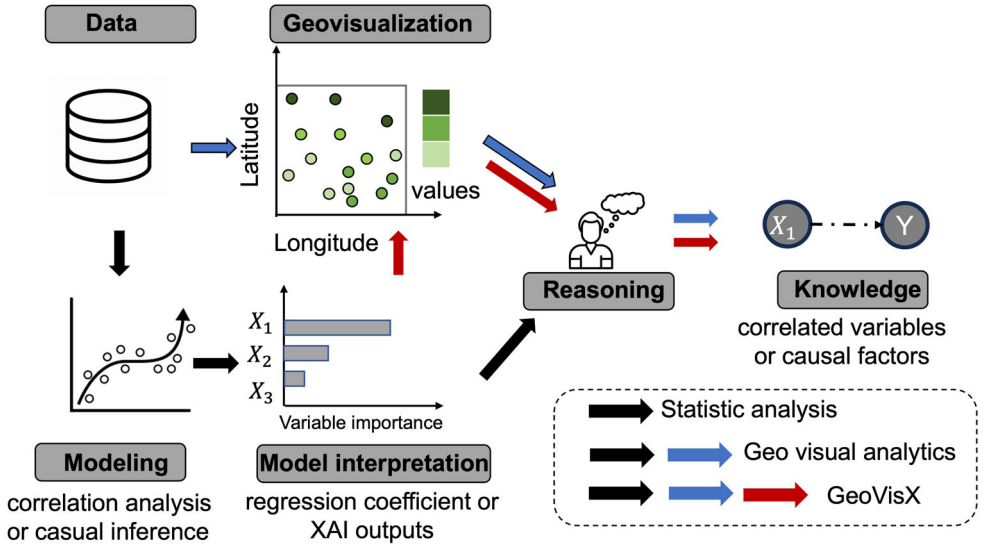

Figure 1. An illustration of the combination of model interpretation(e.g. XAI) and visual analytics.

statistical model outputs (often aggregated results), thereby uncovering accurate spatial knowledge. However, the potential for integrating these two approaches has not yet been fully explored.

GeoVisX combines visual analysis with explainable machine learning with techniques to achieve a deeper understanding of the spatial processes of geographical spatial distribution. We use GeoVisX to explain the predictability and uncertainty of population distribution, addressing two questions. Firstly, it explores the factors influencing population distribution and their spatial heterogeneity. Secondly, it analyzes and improves the accuracy and explainability of population distribution estimates. The first step is the spatial information augmentation of explanatory variables. To fully understand the spatial processes of spatial population distribution, this study first considers the spatial context of explanatory variables when modeling population distribution. The second step involves constructing machine learning models to predict population distribution and introducing explainable machine learning models to elucidate the factors influencing population distribution. The third step focuses on predicting and explaining the errors in previous population forecasts using explainable machine learning models. This research aims to deepen our understanding of the spatial processes underlying population distribution and to assess the uncertainties and causes of population prediction inaccuracies, thus providing more reliable guidance and data support for urban planning and policy-making.

In this study, we use the population distribution of Munich, Germany as a case study to validate the effectiveness and necessity of GeoVisX. By applying GeoVisX, we conducted a detailed analysis of the spatial variability in Munich's population distribution and identified key factors influencing this variability. Moreover, our study explored the uncertainties in population prediction models, providing a deeper understanding of the factors contributing to these uncertainties. These results not only confirm the practicality of GeoVisX in geospatial data analysis but also emphasize the importance of employing explainable machine learning and visual analytics techniques in complex urban planning and policy-making.

2. Study area and data

Munich, Germany, was chosen as the study area for this research. Munich is one of the largest cities in Germany, covering 310.71 km^2 and having approximately 1.56 million inhabitants in 2021. We selected Munich as the study area due to the city's progressive efforts in releasing high-quality geographic datasets. These include 100-meter grid-based census data and various land use datasets, making it particularly well-suited for population distribution studies. Additionally, the availability of such open-source data offers the potential for incorporating supplementary datasets in future experiments, further enhancing the scope of our research.

Additionally, the explanatory variables we collected for predicting and explaining population distribution do not perfectly align in terms of time with the 2011 census data. This decision is intentional, as we prioritize using explanatory variables that are both reliable and temporally consistent with each other. For instance, we selected the latest OSM data due to its high quality, while ensuring that the land use data and

remote sensing images were collected in the same year. The impact of these explanatory variables on population distribution remains relatively stable over short periods, which keeps the uncertainty introduced by using data from a different year manageable. Furthermore, any uncertainty in the predictive task can be assessed through accuracy validation of the prediction model.

2.1. Census data

For our research, we collected the 2011 population census data for Munich at a 100 m resolution. These data were obtained from publications by the Federal Statistical Office and state statistical offices.[4] The 2011 population grid is the most recent census data available at a 100 m resolution. To ensure spatial accuracy and consistency, we used the Lambert Azimuthal Equal Area Projection for this population dataset (Luo *et al*. 2019, Yang 2022).

2.2. Remote sensing data

The remote sensing products include nighttime light (NTL) imagery and normalized difference vegetation index (NDVI). NTL data are from the National Polar-orbiting Partnership - Visible Infrared Imaging Radiometer Suite (NPP-VIIRS) Day-Night Band (DNB) with a spatial resolution of 15 arcseconds (approximately 500 m), provided by the National Geophysical Data Center, the National Oceanic and Atmospheric Administration (NOAA), USA. NDVI data are from the MODerate Resolution Imaging Spectroradiometer (MODIS) on the Terra satellite with a spatial resolution of 250 m. We collected the NTL and NDVI data from the year 2012, as it is the earliest time the NPP-VIIRS DNB data became available.

2.3. Land use data

The land use data used in this study were collected from Urban Atlas datasets. For Functional Urban Areas (FUAs) with 50,000 or more residents, Urban Atlas provides consistent pan-European land use and land cover (LULC) data obtained from Very High Resolution (VHR) satellite imagery. We reclassified the Urban Atlas 2012 dataset, and the implementation metrics for this process are explained in the Appendix of the thesis by Yang (2022).

2.4. POI data and OSM data

The semantic data used in this study consist of Points of Interest (POIs) and OpenStreetMap (OSM) buildings, which are extracted from OpenStreetMap and represented with spatial coordinates. OSM POIs include various features represented as points, lines, or polygons, each described by tags with specific attributes. Examples of such attributes include mailboxes, stores, and schools. In addition, OSM POIs are derived from Geofabrik, an open source platform that provides free OpenStreetMap data (Gao *et al*. 2017). Building attributes are derived from OSM Buildings, a free 3D

building web viewer that supports the OpenStreetMap tagging scheme. The data uses EPSG:4326 (WGS84) as the geometry projection. The dataset includes various building attributes, such as name, type, height, and number of floors.

3. GeoVisX: geospatial visual analytics framework integrated with XAI

3.1. Visual analytics enhance spatial explanation

Although XAI is extensively applied to understand complex relationships among various variables, its application to geographic issues requires special attention. This is primarily due to the special nature of the spatial data. First, the presence of spatial heterogeneity implies that the data generation process varies across different regions (Goodchild and Li 2021). Thus, when exploring the spatial processes of geographic variables, it is crucial to always consider that the conclusions might vary across space, influenced by the localized spatial patterns that emerge. Second, geographic variables are often not independent but exhibit complex interactions with each other. In actual geographic research, it is usually difficult to collect all relevant variables. As a result, it is often challenging to identify the true factors that affect the distribution of geographic variables based on the available data. These natures of the spatial data pose a challenge for XAI applied to geographic variables, but some of them can be solved by combining with the visual analytics. For example, in analyzing the relationship between housing prices and the age of houses in Los Angeles, opposite trends across different communities were revealed through the visualization of regression coefficients from the GWR model (Sachdeva and Fotheringham 2023). Additionally, visual analytics has been employed to identify confounders and hidden variables in public health and medical research (Hauschild *et al*. 2015, Yu *et al*. 2020, Denz and Timmesfeld 2023), although its application remains underexplored in the field of geography.

Figure 2a illustrates how spatial visual analysis helps avoid incorrect interpretations of geographic phenomena. The inherent spatial heterogeneity of geographic variables means that trends and relationships observed at the global level may be inconsistent or even reversed at the local level, leading to phenomena similar to Simpson's paradox (Sachdeva and Fotheringham 2023). As shown in Figure 2a, although the relationship between X and Y is negative across the overall study area (as shown by the grey dots and grey area in the Figure 2a), it turns out to be positive within each subgroup (as shown by the colored areas in the Figure 2a) based on the value of another variable (Z in Figure 1a). The variable Z may represent natural environmental factors or administrative zoning regulations.

Figure 2b demonstrates the application of visual analysis in revealing hidden geographic factors to deepen the understanding of spatial processes. There is a significant nonlinear relationship between X and Y in the study area (as shown by the grey dots and grey area in Figure 2b). However, when mapping this relationship spatially and introducing other variables Z_1 and Z_2 for visualization analysis, we found that the relationship between X and Y may actually be influenced by another variable, Z_1. After grouping based on the value of Z_1 (green area and orange area), we observed that X and Y exhibit a significant linear relationship within each group.In this process,

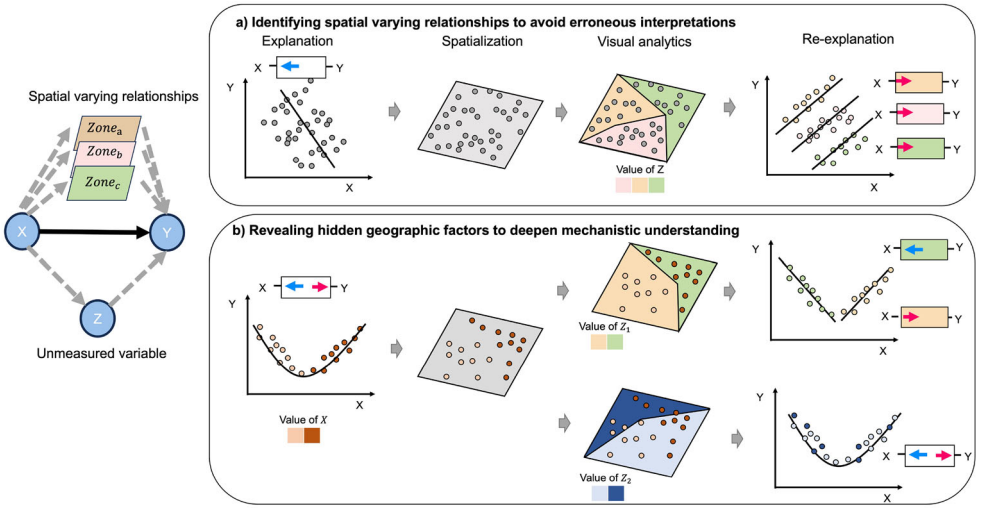

Figure 2. The explanation of how visual analytics enhance the understanding of spatial process.

visualization techniques not only help us understand the correct relationship between *X* and *Y* but also assist us in identifying potential confounders, Z_1.

The spatial process of geographic variables is extremely complex, typically involving interactions among multiple variables. In geographic analysis, the multitude of interrelated variables makes it difficult to identify and collect all factors influencing spatial distribution, thus limiting our depth of understanding of spatial processes. For example, the spatial processes underlying population distribution might differ significantly across urban-rural boundaries. If such differences are not considered as explanatory variables in the analysis using explainable machine learning, it becomes difficult to ascertain the causes of these differences from statistical data alone.

The unique characteristic of geographic variables is that each data point has a unique identifier represented by its latitude and longitude, therefore the value distribution of each geographic variable over the space exhibits a spatial pattern. Visually comparing the spatial patterns of different geographic variables may help reveal the relative impacts of explanatory variables. If the pattern of an outcome aligns with the pattern of a particular variable, it is highly likely that this variable is a critical factor in the mechanism of that geographic element (Luo *et al.* 2022). Thus, spatial patterns can provide insights beyond statistical features, and the potential of visual analysis can be tapped to understand the spatial processes of geographic variable distribution.

As shown in Figure 2b, combining new variables (Z_1 and Z_2) with original data (*X* and *Y*) for visualization can reveal patterns closely aligned with the spatial distribution of the variables. The discovery of such patterns is vital for multivariate spatial analysis, not only helping us identify key variables that may have previously gone unnoticed but also deepening our understanding of the spatial processes behind geographic data. Through this multivariate visual perspective, we can observe the complex relationships between variables, thereby unlocking new spatial processes of geographic phenomena, which are invaluable for accurate geographic analysis and decision-making.

In the field of geographic research, we believe that the explanatory analysis of geographic variables—i.e., revealing underlying spatial processes requires integration with visual analysis. This assertion is deeply rooted in the complex and multifaceted nature of spatial data, including the spatial heterogeneity of spatial processes and the complex interaction of multiple geographic variables. Visual analysis can significantly deepen our understanding of spatial processes and avoid erroneous mechanistic interpretations.

3.2. Framework of GeoVisX

We introduce GeoVisX, a framework that combines explainable machine learning with visual analytics to enhance the understanding of the spatial processes underlying the spatial distribution of geographic variables, encompassing both their predictability and uncertainty.

As shown in Figure 3, GeoVisX conducts a dual exploration into the predictability and uncertainty of geographic variables, therefore we employ a two-step training strategy (Figure 3a). Specifically, for the task of explaining predictability, we partition the complete *Dataset_A*, which records the values of geographic variables into a

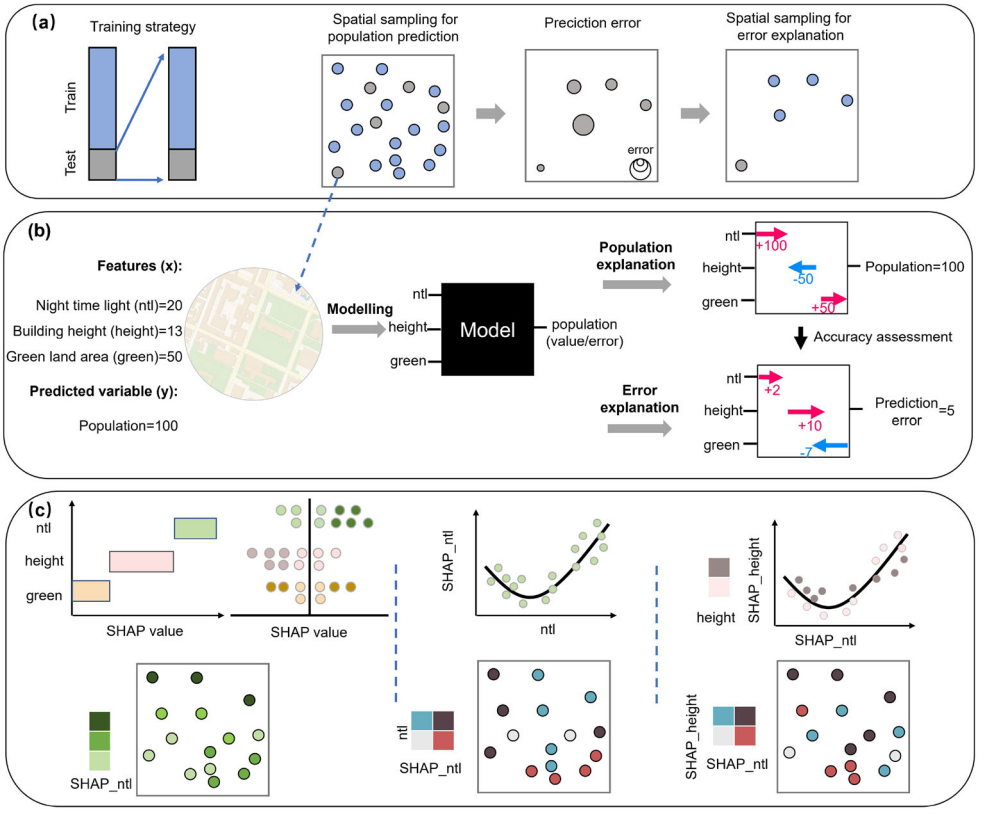

Figure 3. The workflow of GeoVisX.

training set (*Train_A*) and a test set (*Test_A*). Explanatory analysis is conducted within the training set. For the second step, addressing uncertainty, we assess the model prediction accuracy on the test set *Test_A*, using this as the new complete *Dataset_B*, which is then divided into a new training set (*Train_B*) and test set (*Test_B*). The explanation of uncertainty is performed on *Train_B*. It is noteworthy that *Dataset_A* contains recorded values of the geographic variables, whereas *Dataset_B* comprises the values of predictive model errors.

In the second step (Figure 3b), GeoVisX integrates explainable machine learning with a predictive machine learning model to understand the spatial distribution of geographic variables. Taking the analysis of population distribution as an instance, consider three explanatory variables: night-time light, building height, and green land area. For a given location in space, we know the values of these explanatory variables and the population count. To analyze the impact of the three variables on population distribution, an explainable machine learning model is constructed to decompose the contribution of each variable to the population. Subsequently, we calculate the error in the machine learning model's population predictions and apply the same process using the three explanatory variables to model and predict this error. This process reveals the contribution of explanatory variables to the model's prediction error.

In the third step (Figure 3c), after obtaining the statistical outcomes of spatial process through the explainable machine learning model, GeoVisX, we proceed to visualize each data point based on the spatial coordinates, enabling visual analytics. Within GeoVisX, visual analytics for a given explanatory variable mainly encompasses three aspects: spatial mapping of the variable's impact on the variable being explained, spatial mapping of the variable's impact in conjunction with its interactions, and spatial mapping of the interplay between two different explanatory variables' impacts.

By employing this structured approach, GeoVisX not only dissects the contributions of individual variables to the observed geographic outcomes but also addresses the challenges posed by spatial heterogeneity. The resultant visualizations serve as a powerful tool for deciphering complex spatial patterns and contributing factors, thus offering researchers and practitioners a deeper, more intuitive understanding of spatial phenomena.

3.3. Spatial context as the explanatory variable

In analyzing the spatial distribution of geographic variables, it is essential to consider the widespread dependencies and spatial spillover phenomena that exist within the spatial realm. Spatial spillover refers to the phenomenon where certain attributes or variables of an area can spill over into neighboring regions, causing their effects to diffuse spatially (Luo *et al*. 2022). This typically results in geographic phenomena not only being pronounced within the original area but also manifesting similar characteristics or trends in adjacent areas. For example, the prosperity of a commercial district might not only increase property prices within that area but could also enhance the attractiveness and housing prices in neighboring regions. Furthermore, when analyzing the spatial processes behind population distribution, it is crucial to consider not just the environmental and infrastructural characteristics of the residential area, but also

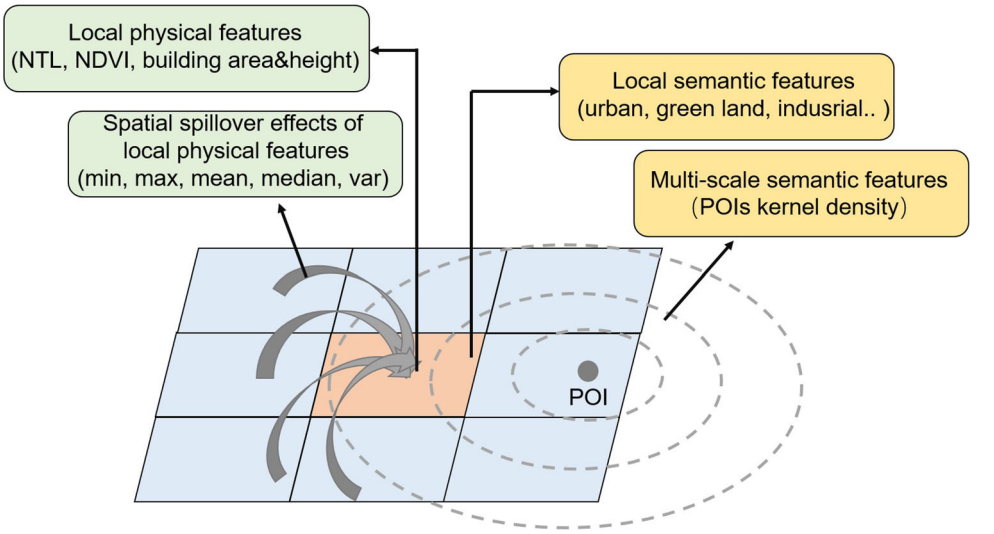

Figure 4. Impact of spatial context on the spatial distribution of geographical variable with population as an example.

the accessibility and amenities of the surrounding areas, which are vitally important to where people choose to reside.

The information about the surrounding neighborhood, or spatial context, is critically important for understanding the processes of spatial distribution of geographic variables. Therefore, within the GeoVisX framework, we have developed a method that considers the spatial context and captures these effects by concurrently considering both local and contextual spatial information.

As illustrated in Figure 4, with population distribution as an example, features such as NTL, NDVI, and building characteristics (area and height) fundamentally represent the immediate built and natural environment of a region. These features directly affect the attractiveness and suitability of a location for residential purposes. However, the spatial spillover effects of these features—quantified through statistical measures such as minimum, maximum, mean, median, and variance—extend the analysis to neighboring cells, providing insights into a broader regional context that could affect population distribution.

Additionally, local semantic features—categorized into urban, green spaces, and industrial areas—offer a detailed understanding of the area's functional landscape. When combined with multiscale semantic features derived from POIs through kernel density estimation, the analysis captures the intensity and scale of urban activities. In this study, the kernel density estimation was performed using ArcGIS Pro, where the defaulted quartic kernel function was employed (Silverman 2018) due to its high computational efficiently and smoothness properties (Chang *et al*. 2012). The scales of analysis for POI kernel density, ranging from 400 meters to 3000 meters, reflect the varying degrees to which amenities and services influence residential choices, as people may consider different amenities at various distances when choosing their residences.

3.4. XAI for spatial data

In this framework, we have integrated SHapley Additive exPlanations (SHAP) to clarify the role of features in spatial prediction and model uncertainty. SHAP is based on the Shapley values from game theory (Shapley 1953), which allocate the contributions of players towards a collective goal (Lundberg and Lee 2017, Li *et al*. 2023a). This approach is commonly used to highlight the significance and influence of a variable in predictive models, particularly in machine learning contexts. For instance, in spatial prediction involving m features (X_1, X_2, …, X_m) for a target geographic variable, our goal is to determine the contribution of each feature. SHAP helps us compute how each feature X_i contributes to the predictive model's output, whether it's forecasting geographic variables or elucidating model uncertainty.

$$\text{Shapley}\ (X_i) = \sum\nolimits_{S \subseteq N\ \{X_i\}} \frac{k!(m-k-1)!}{m!}(f(S \cup \{i\}) - f(S)) \tag{1}$$

$S \subseteq N\ \{X_i\}$ is a set of all possible combinations of features excluding X_i, m is the total number of fetures in the predictive task, and k represents the number of features in the set S. f(S) is the model prediction with features in S, and $f(S \cup \{i\})$ is the model prediction with all features in set S as well as feature X_i.

One significant feature of Shapley values is their ability to decompose the output of a predictive model into distinct portions based on the explanatory variables involved in the prediction. Consequently, Shapley values possess physical significance and additivity. Leveraging this characteristic, we can calculate the overall contribution of spatial context.

For the physical attributes, their contributions of spatial context are calculated as follows:

$$\text{shap}_{\text{context}(X_i)} = \text{shap}(X_{i\ \text{min}}) + \text{shap}(X_{i_{\text{max}}}) + \text{shap}(X_{i_{\text{mean}}}) + \text{shap}(X_{i_{\text{median}}}) + \text{shap}(X_{i_{\text{var}}}) \tag{2}$$

where $X_{i_{\text{min}}}$, $X_{i_{\text{max}}}$, $X_{i_{\text{mean}}}$, $X_{i_{\text{median}}}$, and $X_{i_{\text{var}}}$ represent the corresponding statistical indices for the first-order neighbors of the feature X_i.

For the semantic attributes, assuming that n different scales are considered, their contributions to the spatial context are calculated as follows:

$$\text{shap}_{\text{context}(X_i)} = \text{shap}(X_{\text{scale\_1}}) + \text{shap}(X_{\text{scale\_2}}) + \cdots + \text{shap}(X_{\text{scale\_n}}) \tag{3}$$

where $X_{\text{scale\_1}}$ to $X_{\text{scale\_}n}$ represent the values of the feature X_i at different spatial scales.

4. Methods

4.1. Data pre-processing

To minimize sensitivity to seasonal variations, we applied the annual mean composition to generate mean NDVI and NTL image. Google Earth Engine (GEE), a cloud-based geospatial processing tool that makes use of Google's cloud architecture, was used to acquire and preprocess both datasets (Tamiminia *et al*. 2020). GEE provides free global geospatial data at the petabyte level and uses automatic parallel processing to significantly increase the efficiency of data processing (Luo *et al*. 2021). The

NDVI and NTL data were resampled to 100 m grid level using an interpolation algorithm (Luo *et al*. 2019).

With the wealth of data offered by the Urban Atlas 2012 and Points of Interest (POI) databases, the multitude of categories can be overwhelming and may not align directly with the analytical goals of a study (Yang 2022). Therefore, a strategic categorization is necessary to streamline the data for efficient processing and meaningful analysis.

The initial step involves category merging to refine the dataset into a more manageable and analytically useful form. The Urban Atlas 2012 categories for land use are reclassified into 12 types: industrial, urban, greenland, forest, farmland, land without current use, sport, road, water, construction, mineral, and isolated areas. (Yang 2022).

For the POI data, the 108 original categories provide a detailed breakdown of urban features but require consolidation to enhance their interpretability and relevance to our research objectives. By applying a reclassification algorithm, these categories are systematically consolidated into 17 distinct classes, including retail, life services, and education. The details of the reclassified landuse and POI categories can be found in Yang (2022).

4.2. Generating spatial context- explanatory variables

We generated spatial context explanatory variables to explore their impact on population distribution and model performance, as shown in Figure 5.

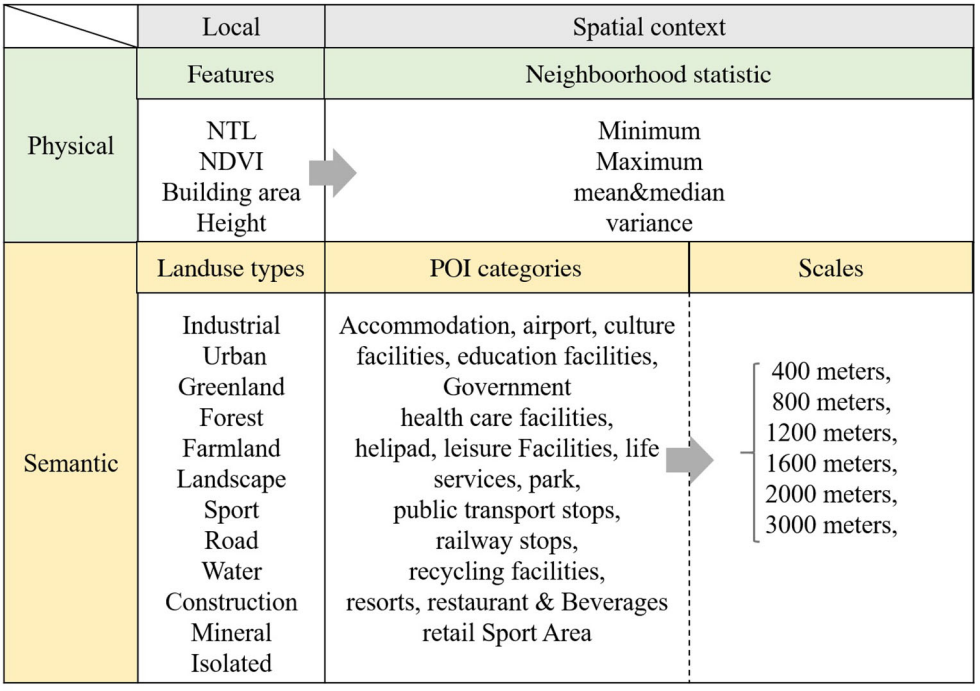

Figure 5. The description of explanatory variables for population.

For four physical variables—NTL, NDVI, Building Area, and Building Height—we computed the statistical values within the first-order neighborhood of a 100 m grid, focusing on eight neighboring cells. This analysis involved calculating the minimum, maximum, mean, median, and variance for each physical variable. As a result, this thorough analysis produced a total of 20 spatial context features for physical variables, providing a detailed depiction of the urban physical environment.

In terms of semantic variables, we leveraged POIs data to capture spatial context information. Kernel density estimation was applied to characterize the intensity of urban activities associated with POIs, using the 17 reclassified POI categories. We performed kernel density analysis at various scales to detect the concentration variations of urban activities. These scales, deliberately chosen, included 400 m, 800 m, 1200 m, 1600 m, 2000 m, and 3000 m, to ensure a multiscale representation of urban semantic density. We computed the kernel density value for each cell within our grid at these scales, resulting in a total of 102 semantic spatial context features that comprehensively embody the semantic structure of the urban landscape.

4.3. Population estimation modeling and explaining

In this study, we selected four different models to predict the population distribution: XGBoost (Chen and Guestrin 2016), Random Forest (Breiman 2001), LightGBM (Ke *et al*. 2017), and Linear Regression. Each model brings unique capabilities and strengths to the task of prediction, and has been widely used in geography studies (Yao *et al*. 2023). Our approach is to train the four models uniformly with the same strategy to ensure comparability. We then choose the model with the highest accuracy to serve as our primary algorithm for population prediction and interpretation.

The first three models are tree-based machine learning algorithms. XGBoost uses gradient boosting framework known for their effectiveness in classification and regression problems. Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees at training time and returns the mode of the classes (classification) or the mean prediction (regression) of each tree. LightGBM is a gradient boosting framework that uses tree-based learning algorithms and is designed for distributed and efficient training. On the other hand, linear regression is a basic statistical method that models the relationship between a dependent variable and one or more independent variables using a linear approach.

For the purpose of our analysis, the entire population dataset comprises 36,900 samples, with 20 percent of the data being allocated to form the test set. To select the optimal parameters for the three machine learning models, we used Bayesian optimization to perform 5-fold cross-validation on the training set (Li *et al*. 2023b).

After validating the accuracy on the test set, we select the best performing model. In order to explain the predictions of the selected model, we use the Shapley score and rely on the interpretative power of the SHAP (SHapley Additive exPlanations) framework. SHAP is a model-agnostic explanation method that provides a unified approach to interpreting the output of any machine learning model.

SHAP provides importance scores for each characteristic in a prediction, providing insight into the contribution of each explanatory variable. Its additive nature allows

for a comprehensive analysis of multivariate interactions that affect the population distribution, facilitating a deeper understanding of the factors involved.

4.4. Population estimation error modeling and explaining

One of the goals of our research is to analyze the uncertainty inherent in population prediction algorithms. For the population prediction module described above, we computed the absolute prediction errors for each location in the test set.

We then used these computed errors to create an error interpretation module. Here, the test set, now annotated with prediction errors, serves as the complete sample space for this module, which includes 7380 samples. We divided this dataset into a new training set and a test set, following an 80/20 split. We then used the same prediction algorithm used in the population prediction module to predict the magnitude of the errors. This recursive approach allows us not only predict population numbers, but also estimate the potential error of our predictions.

Finally, we interpret the prediction errors using the SHAP values. In doing so, we aim to reveal the factors contributing to the prediction errors and to understand their impact. This step is crucial to elucidate the underlying sources of uncertainty in our population predictions, thereby providing a more comprehensive insight into the spatial patterns of prediction variability.

4.5. Model evaluation

To analyze the predictability and uncertainty of population predictions, we employed the same model evaluation methods. We split the dataset into 80% training and 20% testing sets and evaluated performance on the testing set. We chose the mean absolute error (MAE), root mean square error (RMSE), and R^2 as evaluation metrics.

5. Results

5.1. Model accuracy assessment of population estimation

Table 1 presents the accuracy results of the population estimation predictions for the four selected models on the test set. To analyze the importance of multi-scale

Table 1. Performance metrics for population prediction models.

Model	Configuration	MAE (people)	RMSE (people)	R^2
XGBoost	Fixed Scale	15.64	33.72	0.7616
	Optimal Scale	15.53	32.78	0.7748
	Multiscale	14.98	31.62	0.7904
Random Forest	Fixed Scale	15.53	34.08	0.757
	Optimal Scale	15.55	33.97	0.7582
	Multiscale	15.24	33.10	0.7704
LightGBM	Fixed Scale	15.82	33.74	0.7613
	Optimal Scale	15.69	33.40	0.7662
	Multiscale	15.35	32.34	0.7807
Linear Regression	Fixed Scale	28.49	45.76	0.5612
	Optimal Scale	27.42	45.23	0.5712
	Multiscale	26.19	43.25	0.6079

semantic variables of spatial context, we added two sets of ablation experiments. The first set is the fixed scale experiment. For semantic variables, we only chose a 3 km POI kernel density as the explanatory variable. We believe that 3 km is the maximum range that most POIs, representing urban functional facilities, can cover. The second set is the optimal scale experiment. For each type of POI, we calculated multi-scale POI kernel densities, then trained an XGBoost model. Finally, using the models' variable importance, we selected the most important scale of kernel density for each type of POI as the optimal scale. We then only included the optimal scale kernel density of each POI type in the model for training and prediction.

It is evident that all four models achieved the highest accuracy when incorporating neighborhood and multi-scale information. Specifically, for XGBoost, the results considering both neighborhood and multi-scale information outperformed those considering only the optimal scale and a fixed scale by 3.67% and 6.64% in R^2, respectively.

In three different scenarios, XGBoost consistently outperformed in terms of accuracy. Considering neighborhood and multi-scale information, the XGBoost model could explain 79.04% of the population distribution, and its predicted RMSE was 31.62, which was lower than that of Random Forest, Lightgbm, and Linear Regression by 4.68%, 2.28%, and 36.78%, respectively. Therefore, our framework selects XGBoost as the basic model for population prediction and error interpretation.

Our findings also reveal a high nonlinearity between the population distribution and the explanatory variables, indicating that the results of linear regression are significantly inferior to those of the three machine learning methods.

5.2. Explaining population distribution

Figure 6a displays the average absolute SHAP values, indicating the impact of the explanatory variables on the population distribution estimates. The variables *p_urban_fa* (urban fabric area), *build area* (building area), and *height* (building height) have the strongest explanatory power for population distribution. These three factors represent the built-up area of a region and its capacity to accommodate population. Their average impact on the population of a 100 m grid cell is 22.40, 14.45, and 11.13, respectively. The fourth most significant variable is *height_neighbor* (surrounding building

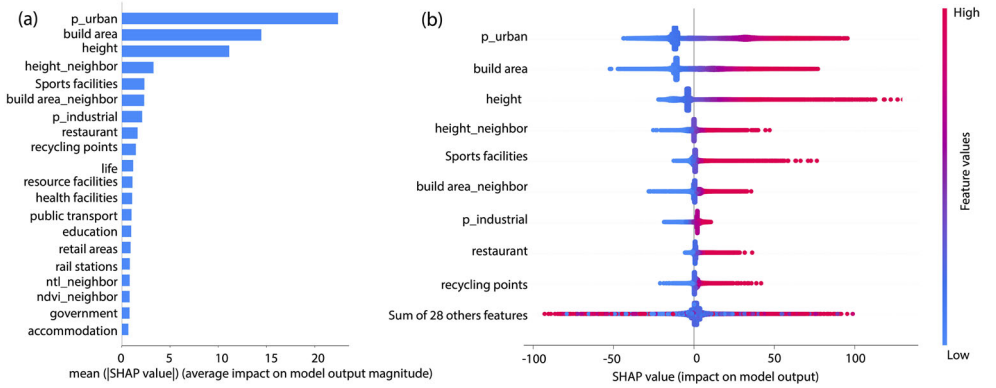

Figure 6. Feature importance for population value based on SHAP model.

height), with an average impact of 3.30 people, indicating that building height in the surrounding area can significantly explain the population size of a region, proving the effectiveness of neighborhood information in explaining population distribution. The fifth most important variable is the multi-scale kernel density of sports facilities, a type of POI, which indicates a strong correlation between the population distribution of Munich and the distribution of sports facilities.

Figure 6b is a summary plot of the SHAP values for the top 9 variables, where the color of the dots represents the high or low feature value of the variable. It is evident that for these variables, such as *p_urban_fa*, *build area*, and *height*, their values positively influence the population count, which is consistent with common sense.

We selected several typical variables to demonstrate the marginal relationship between features and population (Figure 7). In areas with lower NTL levels, there is a positive effect on population numbers. However, when NTL exceeds a certain value, it tends to reduce the population. Previous studies based on NTL modeling have commonly found a positive correlation with population size, primarily because these studies were conducted on larger spatial scales. In contrast, our findings are based on an intra-urban scale of 100-meter grid cells, highlighting the significant impact of scale on the relationship between geographical variables.

The variation in NDVI values does not significantly alter its impact on population numbers. NDVI mainly reflects the level of vegetation in 100-meter grid areas, indicating that the effect of greenery on population is relatively random.

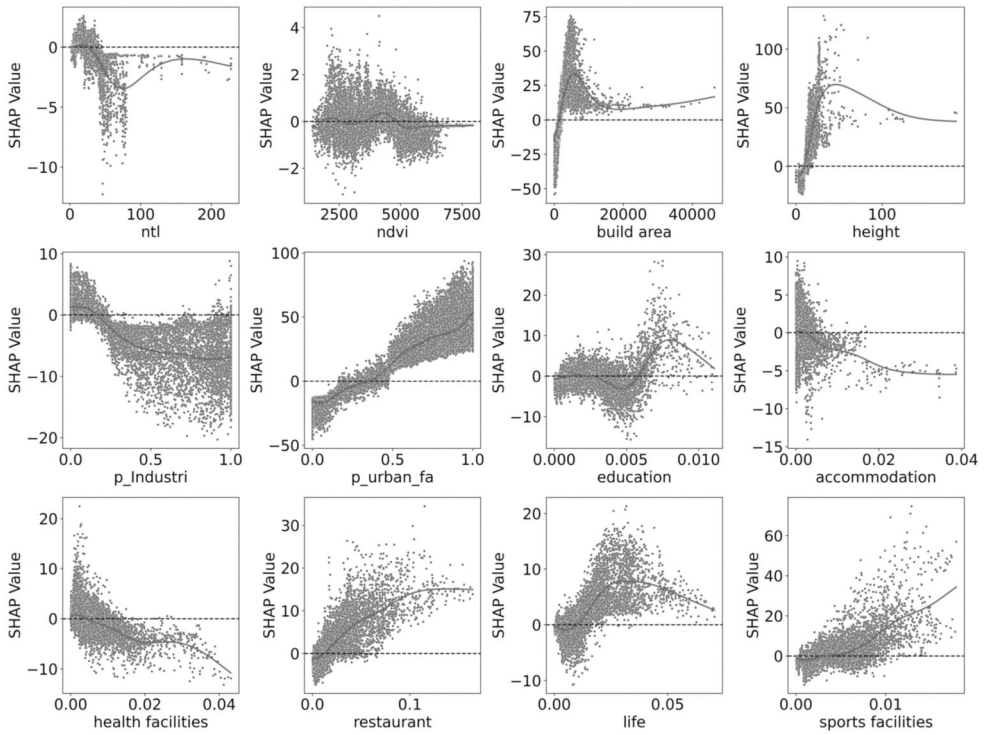

Figure 7. Marginal relationship between features and their contribution to population at 100 m grid level.

The influence of built area and height on population follows a consistent pattern. Initially, at lower values (smaller area, shorter buildings), they have a negative effect on population. However, as these values increase, they encourage residential settlement in the area, with the effect first increasing, then decreasing, and finally stabilizing. Regions characterized by medium levels of built area and height contribute significantly to population, while higher levels may indicate non-residential areas such as factories or sports facilities.

The patterns of impact on population by industrial percentage (*p_industri*) and urban fabric area (*p_urban_fa*) are quite distinct. Areas with lower *p_industri* experience an increase in population, whereas areas with higher *p_industri* see a decrease. This pattern is opposite to that of *p_urban_fa*.

Regarding the six types of semantic information selected (e.g., POI kernel density), we found that higher densities of health facilities lead to a stronger negative impact on population. This suggests that in Munich, the main residential areas tend to be further away from medical facilities. Conversely, higher densities of restaurants and sports facilities correlate with a stronger positive influence on population, reflecting the positive effect of well-developed commercial and recreational facilities on population aggregation.

In Figure 8, we selected four variables to visualize the spatial distribution of their contributions to population. An interesting observation is that nighttime lights generally boost population numbers in most densely populated areas, as indicated by the

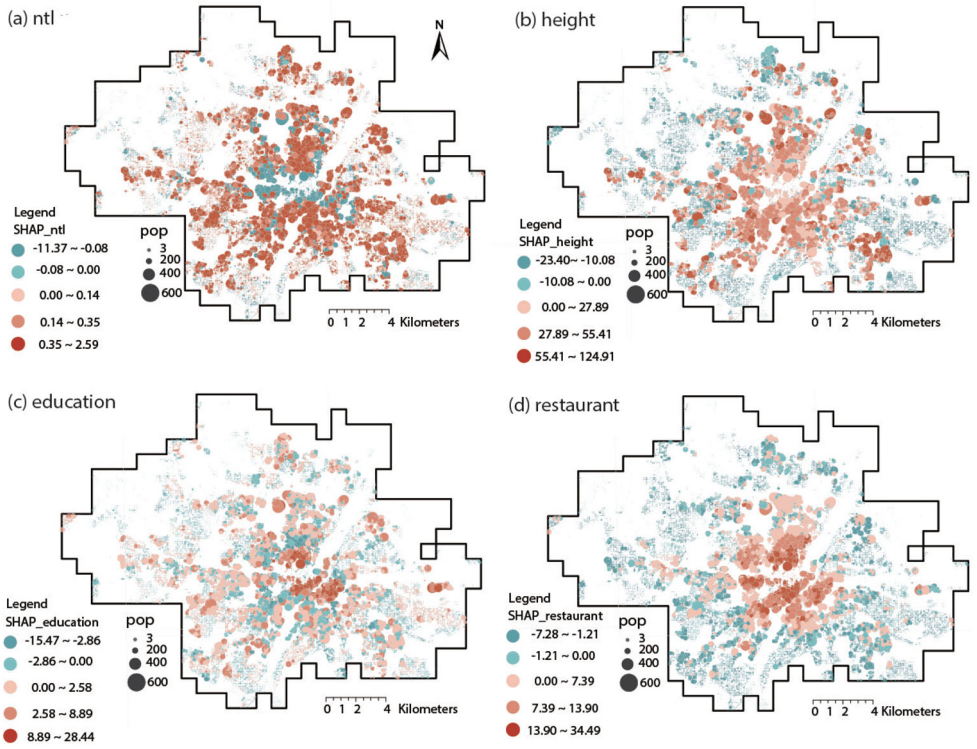

Figure 8. The spatial distribution of variable contributions to population.

dot sizes. However, in the central area of Munich, the effect is negative. This region primarily consists of business districts and major transportation hubs. Consequently, the excessive nighttime lighting may stem from commercial activities rather than residential living. Furthermore, our study found that building height promotes population concentration in city centers. In contrast, in suburban areas, population density is limited by the low height of buildings, leading to fewer inhabitants. Figure 8c and d also reveal that in Munich's city center, the residential population is positively influenced by entertainment activities, represented by education and restaurants.

In Figure 9, we selected the SHAP plots of two variables, *p_urban_fa* and *NTL*, and color-coded the scatter plots based on the other variable to discuss the interaction between these two variables and their impact on population. From Figure 9a, it is evident that with the increase of *p_urban_fa*, its impact on population shifts from negative to positive. Specifically, in areas with low *p_urban_fa*, higher heights result in greater negative impacts; conversely, in areas with high *p_urban_fa*, greater heights lead to increased positive effects.

Figure 9b reveals that NTL has a negative effect on population distribution when it exceeds a certain threshold. These areas are characterized by a high density of restaurants, suggesting that the high NTL values are due to the night lights of bustling commercial districts rather than residential lights.

Figure 9c displays the interactive effects between the variables *height* and *p_urban_fa*. It shows that in most densely populated areas, both variables contribute positively

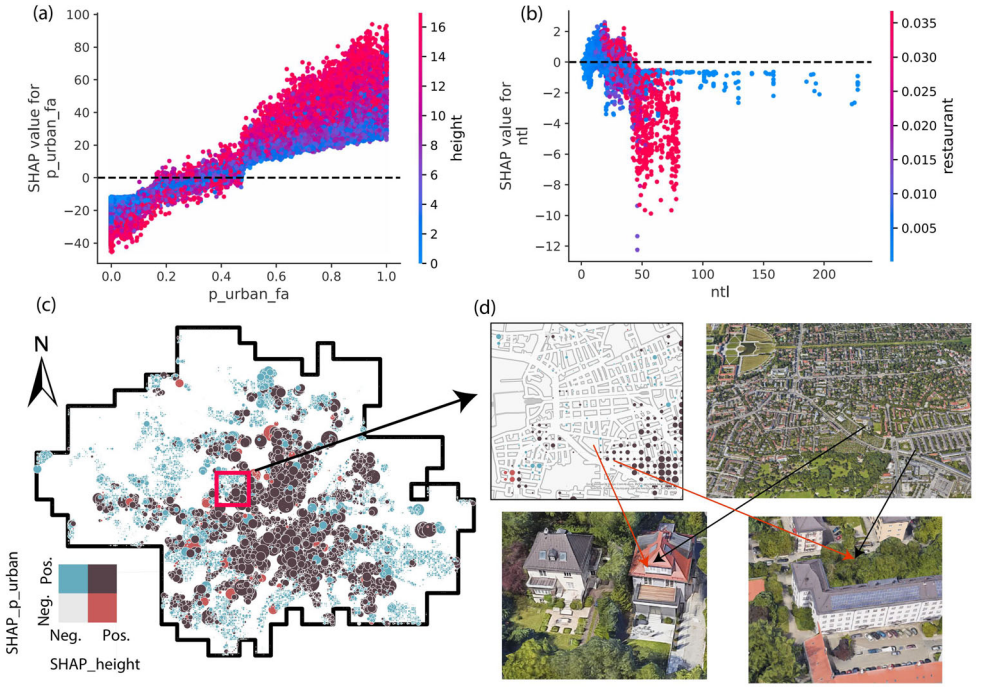

Figure 9. SHAP-based interaction effect: the SHAP-based partial dependence plots for a) p_urban_fa and b) ntl; c) and d) show the spatial distribution of the interaction effect between *p_urban_fa* and *height*.

to population growth. However, in some areas, we observed a scenario where the *urban_fa* had a positive effect, but *height* had a negative impact. We zoomed into a specific area for a detailed analysis, as shown in Figure 9d. This area is the Nymphenburg district of Munich. A distinct boundary line is evident, revealing two different spatial processes of population distribution. The community on the left experiences a negative impact of 'height' on population, whereas the community on the right sees a positive effect. Through a visual check on Google Maps, it is apparent that the two communities have significantly different housing structures. The left side primarily consists of villas with lower building heights, while the right side features taller residential buildings. For high-rise residential buildings, higher floors are likely to house more residents. However, in villa areas, an increase in building height does not lead to a positive boost in the internal population.

We further investigated the potential of visual analysis in understanding the spatial processes behind population distribution. In Figure 10, we selected two features, *education* (education resource) and *ntl*, to display their values and their impact on population distribution (SHAP) in a bivariate overlay. For the feature values, we used natural breaks to divide them into high and low groups. For SHAP values, we categorized them into positive and negative groups. In Figure 10a, it is evident that in the

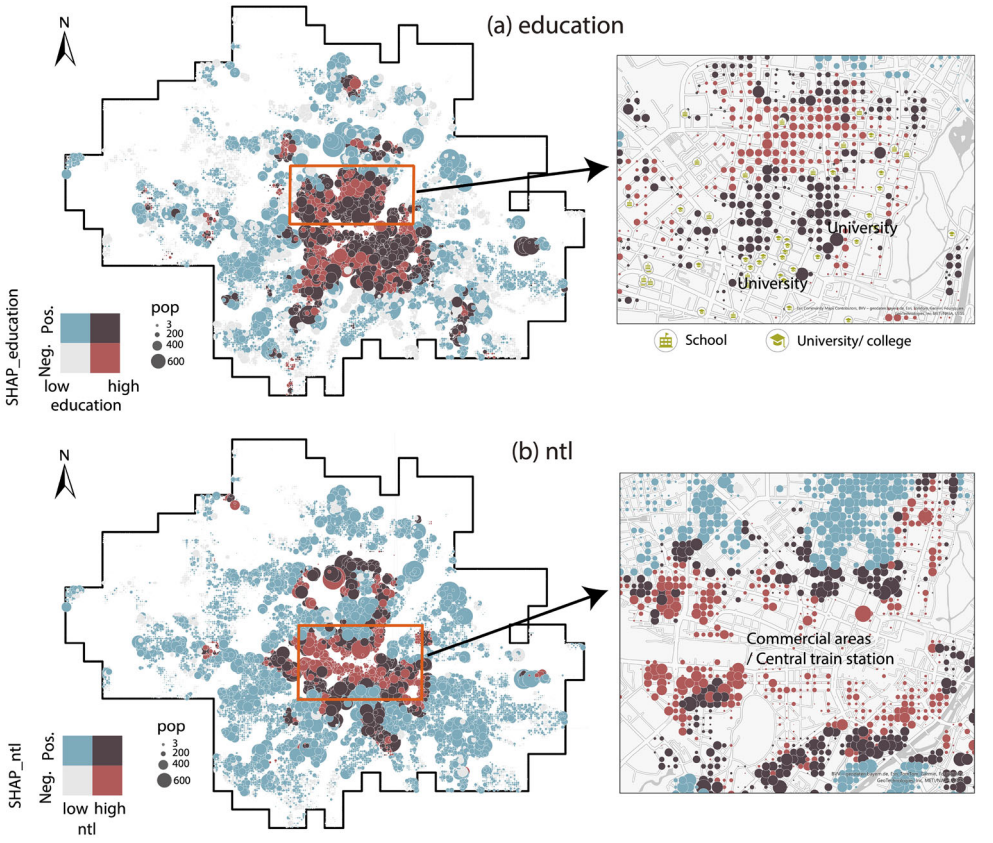

Figure 10. The spatial distribution of interaction effect between the features and their contributions to population: a) education, b)ntl.

suburban areas of Munich, *education* positively affects population distribution, despite its low numerical values. This indicates the significant role of educational facilities in potential population growth in suburban areas. In the urban areas of Munich, while the values for *education* are high, its impact on population distribution is not always significant. As indicated by selected local areas, deep red dots represent areas where *education* promotes population distribution, mostly around the two universities in Munich, suggesting that universities attract people to live nearby for educational purposes. However, in areas far from universities, despite the presence of many schools, *education*'s impact on population growth is not strong. This suggests that in urban areas, the factors influencing people's residential choices are diverse, including job opportunities, surrounding natural environment, and recreational activities, with the role of schools being relatively minor.

Figure 10b shows the spatial distribution of *ntl* and its impact on population. In suburban areas, *ntl* also potentially promotes population growth, while in the central business district and around the central train station, despite high *ntl* levels, its value correlates negatively with population. Our findings also indicate that in the surrounding areas of the central business district, the value of *ntl* positively correlates with population concentration, likely because these areas are primarily residential, and the nighttime lighting originates mainly from the living needs of the resident population.

5.3. Explaining the uncertainty of population prediction

Figure 11a shows the ranking of SHAP values for variables in explaining model errors. The most important features are *build area*, *p_urban_fa*, and *height*. Areas with more construction and taller buildings exhibit higher uncertainty in population estimates, while areas with less construction and lower buildings show lower uncertainty. Moreover, *height_neighbor* is the fourth-ranked variable, indicating the neighborhood effect of building height on the uncertainty of population estimation. Specifically, the height of buildings in an area can also affect the model's accuracy in predicting the population of neighboring areas.

In the task of explaining model accuracy, a negative SHAP value for a variable indicates that it helps reduce error and improve accuracy. Figure 11b quantifies the positive and negative impacts of explanatory variables on prediction accuracy across

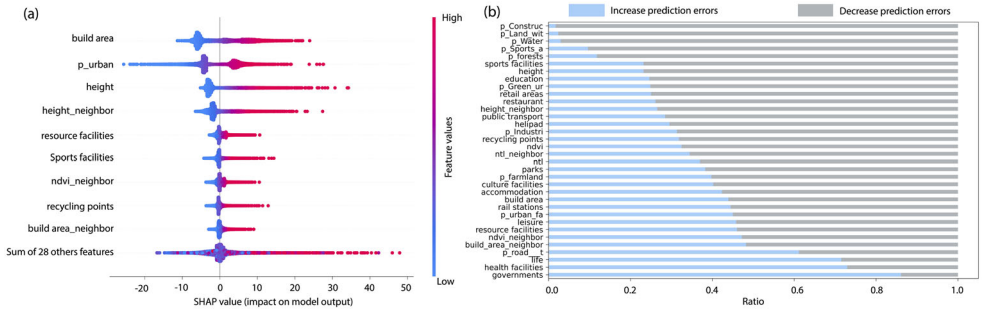

Figure 11. (a) The variable contributions to model uncertainty and (b) the the proportion of samples in which the feature reduces uncertainty.

Munich's 100 m population grid. The results reveal that *p_construc* (Construction sites), *p_land_wit* (Land without current use), *p_water* (Water area), and *p_sports_a* (Sports and leisure facilities) have the largest proportions of improving accuracy. These areas occupy a smaller proportion of the study area but have a more definite impact on population estimation. For example, areas where *p_water* is proportionally high can be confidently estimated to have zero population. Additionally, variables such as *sports facilities*, *height*, and *education* positively influence the accuracy of population predictions. However, government and health facilities tend to decrease the accuracy of population predictions in most areas, possibly because these types of locations share many features (like height and build area) with densely populated residential areas, yet they inherently have lower resident populations, leading to significant confusion in the model's predictions.

Figure 12 demonstrates the impact of twelve main variables on the accuracy of population estimation, with most relationships exhibiting strong non-linearity across feature values. For instance, as the NDVI increases, its impact on model accuracy initially rises, then decreases, and eventually increases again. This indicates that population estimation uncertainty is higher at moderate levels of NDVI. With the increase in build area and height, their influence on the model's prediction error significantly increases at first, then stabilizes.

Figure 13 illustrates the impact of interactions between two variables, *p_urban_fa* (a) and *p_industri* (b), and another variable on the accuracy of population prediction. It

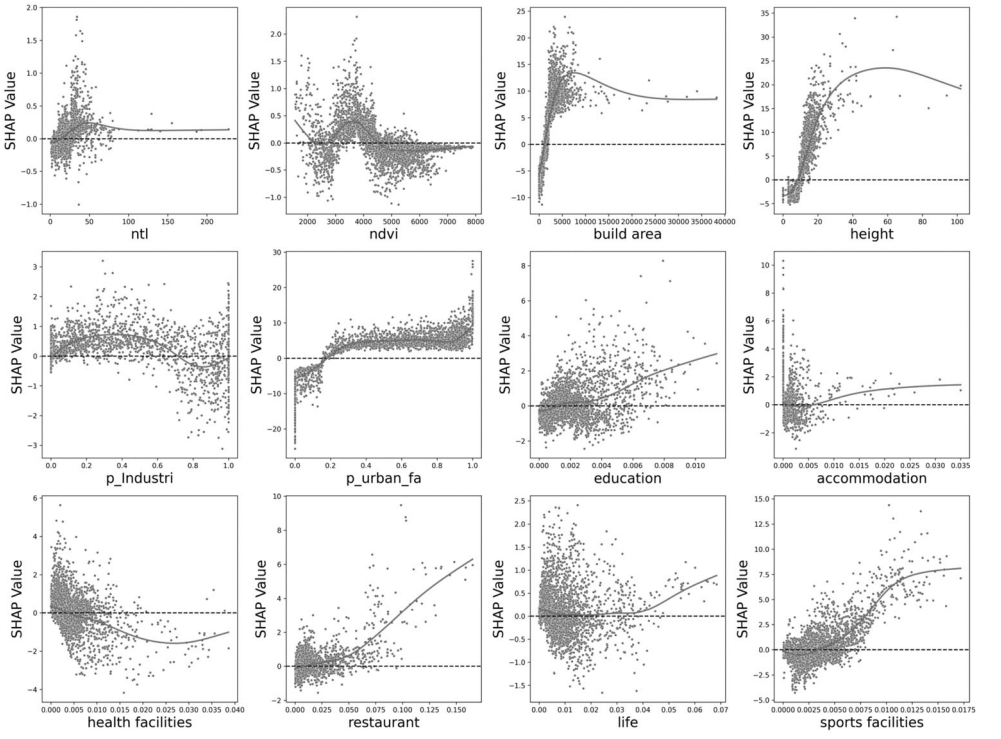

Figure 12. The marginal relationship between features and their contribution to the model uncertainty.

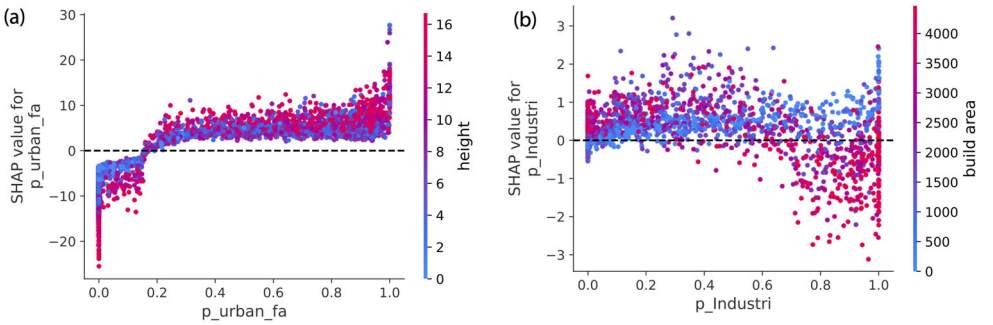

Figure 13. The partial dependence plots for a) p_urban_fa and b) p_industri; The dots are colored by the values of the feature.

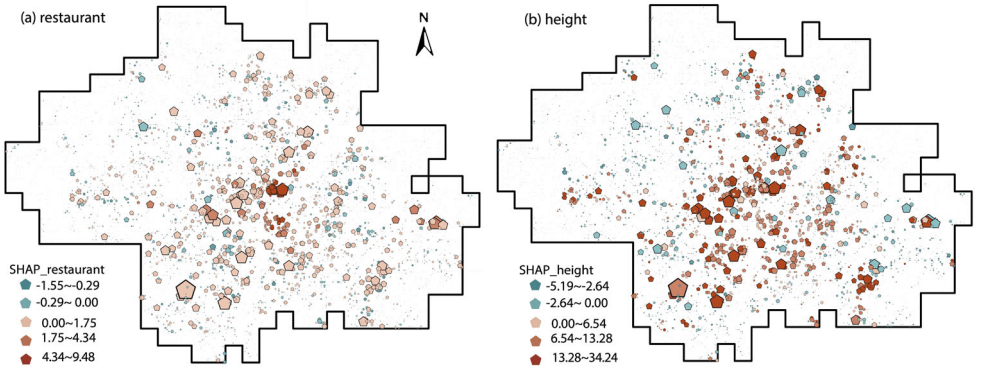

Figure 14. The spatial distribution of feature contributions to model uncertainty.

is observed that in areas with low *p_urban_fa*, there is an improvement in population accuracy, especially in areas with greater height. Conversely, in regions with high *p_urban_fa*, the higher the height, the more pronounced the negative impact on population accuracy.

In scenarios with low *p_industri*, areas with a high build area exhibit a significant negative effect on population prediction. Conversely, in areas with high *p_industri*, the higher the build area of the grid, the more noticeable the improvement in the accuracy of population prediction by *p_industri*.

Figure 14 displays the spatial distribution of the explanatory strength of two variables on predictive error. Pentagons in different shades indicate the extent to which each feature contributes to the error: red signifies an increase, while blue signifies a decrease in error, with deeper colors indicating a greater impact. In areas with higher population density, particularly in central urban zones, both the presence of restaurants and building height tend to increase the uncertainty in population predictions. The height can augment the error by up to approximately 34, whereas restaurants have a maximum impact of about 10. The restaurant industry might contribute to this increased uncertainty due to its attraction for population movement and aggregation. In city centers, where buildings are often taller, the high density of commercial and

residential structures significantly influences the predictive error, possibly due to the more complex distribution of people in high-rise areas.

Conversely, in suburban regions, these features may exhibit a decreasing effect on uncertainty. Less developed areas typically have sparser restaurant distribution and lower buildings, which may help the model to estimate population more accurately. The uniformity in population distribution facilitated by low-density commercial establishments and shorter buildings might have a negative impact on predictive error, effectively reducing uncertainty. This visualization not only allows us to observe the impact of "restaurants" and "height" on predictive error but also to understand the variation in their effects across different regions.

6. Discussion

We recognize the potential of visual analytics for analyzing spatial relationships; however, there is a notable lack of research in this area. Our study is a preliminary and straightforward exploration, but we believe this direction warrants deeper investigation. First, we discuss the importance of visual analytics and numerical analysis methods (e.g., XAI methods) for spatial correlation analysis and spatial process identification, focusing on fundamental characteristics of spatial data analysis, such as the prevalent issue of confounding variables. This importance includes helping to avoid erroneous interpretations of spatial processes and aiding in the discovery of accurate spatial processes. Second, using population distribution as an example, we explore the feasibility of combining visual analytics with numerical models.

6.1. Visualization as a high-dimensional approach to understanding geographical variables

GeoVisX is a model-agnostic visualization framework applicable to any spatial analysis aimed at uncovering spatial processes, whether through causal inference or correlation analysis. GeoVisX is not designed for enhancing the precision of correlation or causality estimation within existing statistical models. Instead, it addresses key challenges inherent in geographic data to prevent misinterpretation of results. First, spatial heterogeneity causes relationships—whether causal or correlational—between variables to vary across space, with aggregated statistical analyses often masking these underlying patterns and trends. Second, the complex interactions among geographic variables make high-dimensional data analysis susceptible to hidden confounding effects. We argue that these issues can be alleviated through the geovisualization of model interpretations. In GeoVisX, the chosen approach to interpretation leverages outputs from XAI models, providing deeper insights into spatial processes.

As demonstrated through two case studies in Figure 2, we illustrate how XAI can uncover the underlying spatial processes of geographic data, even with incomplete datasets. Despite the difficulty in encompassing all explanatory variables, geographic variables exhibit significant spatial patterns. These patterns can be visualized and analyzed through the human visual system. From this viewpoint, visual analysis extends beyond simple data representation, serving as an effective means of exploring spatial

processes in high-dimensional spaces. The intuitive nature of visual analysis facilitates the identification of subtle geographic patterns that may remain hidden from traditional analytical models, thereby revealing deeper and more nuanced insights. In particular, visual analysis enables researchers to observe local variations, which are often obscured in global or aggregate analyses. By comparing local patterns and identifying their similarities with other geographic variables, it becomes possible to uncover new latent variables or confounders. This method is particularly effective for handling complex geographic datasets with intricate interactions, which can challenge the capabilities of conventional models, thus providing richer insights and new directions for research and practice in geographic science.

In our study on the spatial distribution of population in Munich, we employed visual analytics to demonstrate how numerical analysis results can be spatially mapped to reveal hidden spatial processes. For example, the impact of NTL on population distribution shows a concentrated negative effect in the city center. This area, a major commercial hub of Munich featuring the central train station, suggests that high NTL values, predominantly from commercial activities, do not accurately represent residential populations. Through such methods, visualization serves not merely as a tool for data presentation but as a key approach to explaining and understanding complex geographical phenomena. It enables us to view data from new perspectives, identifying patterns and trends that may not be evident through raw data analysis alone.

Our findings highlight the significant role and immense potential of visual analysis in elucidating the spatial processes of geographic distribution and enhancing spatial understanding. Visual analytics is essential in interpreting geographical data, particularly when explaining the complex, multi-dimensional interactions underlying geographical phenomena.

However, in this study, we have conducted only a preliminary exploration and validated its feasibility, employing relatively basic visualization techniques, such as bivariate interaction mapping. Future research should investigate the use of more advanced visual analytics techniques to further harness its potential for spatial understanding. Furthermore, the visual analytics method presented in this paper separates the modeling component from the visual analytics component. The results are first obtained through the machine learning model, then visualized, and finally interpreted to draw new conclusions. Future research should consider integrating visualization and analysis into a cohesive process, allowing readers to discover conclusions based on their own interests and domain knowledge.

6.2. The spatial process of population distribution

In this study, we aim to elucidate the underlying spatial processes of population distribution at a fine-grained spatial level, including aspects of predictability and the causes of uncertainty in predictive models. To understand the complex spatial processes of spatial distribution, we developed the GeoVisX framework by integrating interpretable machine learning with visualization techniques. Additionally, GeoVisX accounts for the spatial spillover of geographic variables by incorporating neighborhood information into the mechanistic explanations. We applied GeoVisX to explain the spatial processes

behind the population distribution of Munich at a 100-meter grid scale to validate the effectiveness of this module.

Regarding the predictability of spatial population distribution, we identified several fundamental physical features, such as urban area proportion, building area, and building height, which have the most significant impact on population distribution. Furthermore, we discovered the importance of neighborhood information for explaining population distribution. For instance, the average building height in a region can influence the population by approximately 12 people and affect the neighboring areas by about 4 people. This underscores the importance of incorporating spatial context into the mechanistic explanations of geographical phenomena.

Through visualization analysis, we also uncovered some local-scale spatial processes of population distribution. Contrary to the patterns observed at larger scales, a positive correlation between nighttime light and population count shows significant spatial heterogeneity. In central urban areas with developed commercial activities, nighttime light tends to be negatively correlated with population. Additionally, we revealed complex interactions among multiple variables, such as the interaction between the proportion of urban areas and building height, which clearly delineates the boundaries between villas and apartment buildings, reflecting the residential preferences of various groups.

In terms of uncertainty in population distribution, physical features still have the most significant impact, including building area and height. However, we found that certain unique physical features, such as construction sites, water bodies, and forests, greatly reduce the uncertainty in population predictions. We also explored how interactions between different variables contribute to model uncertainty. We observed that a lower proportion of built-up areas reduces uncertainty in population predictions, especially in regions with taller buildings, while the opposite is true in areas with a higher proportion of built-up areas.

Finally, the spatial processes of the population distribution revealed by this study mainly refer to variables that contribute to the generation of population distribution. However, this contribution does not necessarily imply causality. GeoVisX is a visual analytics framework that is independent of statistical models, designed to support both causal inference and correlation analysis. In geographic research, correlation helps identify patterns and associations between variables, which is valuable for prediction and exploratory analysis. However, causality goes a step further by uncovering the underlying mechanisms driving geographical patterns, enabling more targeted interventions and policy-making. Correlation does not imply causation, and distinguishing between the two is essential to avoid drawing misleading conclusions (Pearl 2009). Causal inference is widely applied in the time series analysis, but in many geographical studies, only cross-sectional data is available. Therefore, spatial causal inference is still underexplored. Future research can integrate more causal inference tools into the current framework, either from the field of visual analytics or XAI (Carloni *et al*. 2023). Some XAI methods have been developed for causal inference (Feuerriegel *et al*. 2024), and they have been applied to causal inference in medical diagnosis (Ji *et al*. 2022). Recent studies explore how visualization contributes to identifying causality in AI models (Zimmermann *et al*.

2021). The GeoVisX framework can be integrated with spatial causal inference methods (Gao *et al*. 2023).

7. Conclusion

To enhance the understanding of the spatial processes underlying spatial population distribution, revealing both its predictability and the uncertainties in prediction models, we have developed the GeoVisX framework by integrating XAI and visual analytics techniques. GeoVisX initially addresses the issue of spatial spillover in population distribution processes, generating spatial context information for different explanatory variables. It constructs predictive machine learning models and employs the SHAP algorithm for explainable machine learning to interpret the spatial processes of population distribution. Finally, GeoVisX spatially maps the mechanistic interpretation results, using multiple visual analytics techniques to unearth the spatial patterns of these spatial processes. The case study in Munich has proven its effectiveness in deciphering the spatial processes of population distribution, thereby demonstrating the significant potential of combining visual analytics with explainable machine learning in explaining the spatial processes of geographical distribution and enhancing spatial understanding. In future research, we plan to explore a broader range of visual analytics techniques and explainable machine learning methods to more accurately understand the spatial distribution processes of geographic variables. In addition, causal inference tools are expected to be integrated into the GeoVisX framework in the future.

Notes

1. http://sedac.ciesin.columbia.edu/data/collection/gpw-v4
2. https://landscan.ornl.gov/
3. www.worldpop.org/
4. https://www.zensus2011.de/DE/Home/Aktuelles/DemografischeGrunddaten.html

Acknowledgment

We would like to thank the editors and anonymous reviewers for their constructive suggestions and comments for improving this manuscript.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Data available statement

The data and codes that support the findings of the present study are available on Figshare at https://doi.org/10.6084/m9.figshare.25908820.

Notes on contributors

Peng Luo is a Postdoctoral Research Fellow at MIT Senseable City Lab and was a visiting scholar at University of Oxford. He holds a Ph.D. from the Chair of Cartography and Visual Analytics at the Technical University of Munich. His research centers on GeoAI, spatially explicit modeling, urban analytics, and explainable and uncertainty-aware spatial modeling.

Chuan Chen is a PhD candidate in Chair of Cartography and Visual Analytics at Technical University of Munich. He holds a Master of Science in Geodesy and Geoinformation at TUM and a Bachelor of Engineering from Wuhan University. He currently focuses on research in Knowledge Graph, Responsible GIS, and Visual Explainable GeoAI.

Song Gao is an associate professor in GIScience at the Department of Geography, University of Wisconsin-Madison. He holds a PhD in Geography at the University of California, Santa Barbara. His main research interests include place-based GIS, geospatial data science and GeoAI approaches to human mobility and social sensing.

Xianfeng Zhang received his Ph.D. from the University of Western Ontario, London, Canada in 2005. He is currently a full professor and deputy director with the Institute of Remote Sensing and GIS, Peking University. His research interests include remote sensing for natural hazard management, deep learning, remote sensing big data, and remote sensing for transportation.

Deng Majok Chol is a DPhil candidate at Oxford University. He holds a B.S. in Political Science and Economics from Arizona State University, an MBA from George Washington University, and an MPA from Harvard University. His work focuses on social hydrology, particularly the impact of climate and socio-economic changes on population exposure to floods.

Zhuo Yang hold the master degree in Cartography at Technical University of Munich. She holds the bachelor degree at Wuhan Univerisity. Her resarech is focues on developing population mapping methods utilized geospatial data.

Liqiu Meng is a professor of Cartography at the Technical University of Munich, and a member of German National Academy of Sciences. She is serving as Vice President of the International Cartographic Association. Her research interests include geodata integration, mobile map services, multimodal navigation algorithms, geovisual analytics, and ethical concerns in social sensing.

ORCID

Peng Luo http://orcid.org/0000-0002-3680-8509
Song Gao http://orcid.org/0000-0003-4359-6302

References

Abdul Salam, A., *et al.*, 2014. Population distribution and household conditions in Saudi Arabia: reflections from the 2010 census. *SpringerPlus*, 3 (1), 530.

Andrienko, G., *et al.*, 2010. Space, time and visual analytics. *International Journal of Geographical Information Science*, 24 (10), 1577–1600.

Bakillah, M., *et al.*, 2014. Fine-resolution population mapping using openstreetmap points-of-interest. *International Journal of Geographical Information Science*, 28 (9), 1940–1963.

Breiman, L., 2001. Random forests. *Machine Learning*, 45 (1), 5–32.

Carloni, G., Berti, A., and Colantonio, S., 2023. The role of causality in explainable artificial intelligence. arXiv preprint arXiv:2309.09901.

Chang, K.H., Kao, H.M., and Chang, T.J., 2012. Lagrangian modeling of particle concentration distribution in indoor environment with different kernel functions and particle search algorithms. *Building and Environment*, 57, 81–87.

Chen, T., and Guestrin, C., 2016. Xgboost: A scalable tree boosting system. *In*: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, San Francisco, California, USA. New York, NY: Association for Computing Machinery, 785–794.

Cheng, X., *et al.*, 2021. A method to evaluate task-specific importance of spatio-temporal units based on explainable artificial intelligence. *International Journal of Geographical Information Science*, 35 (10), 2002–2025.

Currie, G., 2004. Gap analysis of public transport needs: measuring spatial distribution of public transport needs and identifying gaps in the quality of public transport provision. *Transportation Research Record: Journal of the Transportation Research Board*, 1895 (1), 137–146.

De Jong, G.F., and Sell, R.R., 1977. Population redistribution, migration, and residential preferences. *The ANNALS of the American Academy of Political and Social Science*, 429 (1), 130–144.

Deb, D., and Smith, R.M., 2021. Application of random forest and shap tree explainer in exploring spatial (in) justice to aid urban planning. *ISPRS International Journal of Geo-Information*, 10 (9), 629.

Denz, R., and Timmesfeld, N., 2023. Visualizing the (causal) effect of a continuous variable on a time-to-event outcome. *Epidemiology (Cambridge, Mass.)*, 34 (5), 652–660.

Deville, P., *et al.*, 2014. Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences of the United States of America*, 111 (45), 15888–15893.

Doxsey-Whitfield, E., *et al.*, 2015. Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4. *Papers in Applied Geography*, 1 (3), 226–234.

Feuerriegel, S., *et al.*, 2024. Causal machine learning for predicting treatment outcomes. *Nature Medicine*, 30 (4), 958–968.

Fischer, M.M., and Wang, J., 2011. *Spatial data analysis: models, methods and techniques*. Berlin, Heidelberg: Springer Science & Business Media.

Fotheringham, A.S., and Sachdeva, M., 2022. Modelling spatial processes in quantitative human geography. *Annals of GIS*, 28 (1), 5–14.

Gao, B., *et al.*, 2023. Causal inference from cross-sectional earth system data with geographical convergent cross mapping. *Nature Communications*, 14 (1), 5875.

Gao, S., Janowicz, K., and Couclelis, H., 2017. Extracting urban functional regions from points of interest and human activities on location-based social networks. *Transactions in GIS*, 21 (3), 446–467.

Goodchild, M.F., and Li, W., 2021. Replication across space and time must be weak in the social and environmental sciences. *Proceedings of the National Academy of Sciences*, 118 (35), e2015759118.

Hauschild, A.C., *et al.*, 2015. Carotta: revealing hidden confounder markers in metabolic breath profiles. *Metabolites*, 5 (2), 344–363.

Hay, A.M., and Johnston, R., 1983. The study of process in quantitative human geography. *L'Espace Géographique*, 12 (1), 69–76.

Hsu, C.Y., and Li, W., 2023. Explainable geoai: can saliency maps help interpret artificial intelligence's learning process? an empirical study on natural feature detection. *International Journal of Geographical Information Science*, 37 (5), 963–987.

Huang, D., *et al.*, 2017. Emerging polycentric megacity in china: An examination of employment subcenters and their influence on population distribution in beijing. *Cities*, 69, 36–45.

Huang, X., *et al.*, 2021. Sensing population distribution from satellite imagery via deep learning: Model selection, neighboring effects, and systematic biases. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 5137–5151.

Ji, S., *et al.*, 2022. Understanding cycling distance according to the prediction of the xgboost and the interpretation of shap: A non-linear and interaction effect analysis. *Journal of Transport Geography*, 103, 103414.

Ke, G., *et al.*, 2017. LightGBM: A highly efficient gradient boosting decision tree. *In*: *Proceedings of the 31st international conference on neural information processing systems (NIPS'17)*. Red Hook, NY: Curran Associates Inc., 3149–3157.

Langford, M., *et al.*, 2008. Urban population distribution models and service accessibility estimation. *Computers, Environment and Urban Systems*, 32 (1), 66–80.

Larson, K.L., *et al.*, 2009. Residents' yard choices and rationales in a desert city: social priorities, ecological impacts, and decision tradeoffs. *Environmental Management*, 44 (5), 921–937.

Li, M., *et al.*, 2021. Prediction of human activity intensity using the interactions in physical and social spaces through graph convolutional networks. *International Journal of Geographical Information Science*, 35 (12), 2489–2516.

Li, Y., *et al.*, 2023. A locally explained heterogeneity model for examining wetland disparity. *International Journal of Digital Earth*, 16 (2), 4533–4552.

Li, Z., 2023a. Geoshapley: A game theory approach to measuring spatial effects in machine learning models. arXiv preprint arXiv:2312.03675.

Li, Z., 2023b. Leveraging explainable artificial intelligence and big trip data to understand factors influencing willingness to ridesharing. *Travel Behaviour and Society*, 31, 284–294.

Liu, X., *et al.*, 2020. High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. *Nature Sustainability*, 3 (7), 564–570.

Liu, P., Zhang, Y., and Biljecki, F., 2024. Explainable spatially explicit geospatial artificial intelligence in urban analytics. *Environment and Planning B: Urban Analytics and City Science*, 51 (5), 1104–1123.

Lundberg, S.M., *et al.*, 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence*, 2 (1), 56–67.

Lundberg, S.M., and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4768–4777.

Luo, P., *et al.*, 2019. Modeling population density using a new index derived from multi-sensor image data. *Remote Sensing*, 11 (22), 2620.

Luo, P., *et al.*, 2022. Identifying determinants of spatio-temporal disparities in soil moisture of the northern hemisphere using a geographically optimal zones-based heterogeneity model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 185, 111–128.

Luo, P., Song, Y., and Wu, P., 2021. Spatial disparities in trade-offs: economic and environmental impacts of road infrastructure on continental level. *GIScience & Remote Sensing*, 58 (5), 756–775.

Maantay, J.A., Maroko, A.R., and Herrmann, C., 2007. Mapping population distribution in the urban environment: The cadastral-based expert dasymetric system (ceds). *Cartography and Geographic Information Science*, 34 (2), 77–102.

Mason, A., 2001. *Population change and economic development in East Asia: Challenges met, opportunities seized*. Redwood City: Stanford University Press.

Mennis, J., 2003. Generating surface models of population using dasymetric mapping. *The Professional Geographer*, 55 (1), 31–42.

Patel, N.N., *et al.*, 2017. Improving large area population mapping using geotweet densities. *Transactions in GIS: TG*, 21 (2), 317–331.

Pearl, J., 2009. *Causality*. Cambridge, England: Cambridge University Press.

Sachdeva, M., and Fotheringham, A.S., 2023. A geographical perspective on simpson's paradox. *Journal of Spatial Information Science*, 26, 1–25.

Shapley, L.S., 1953. *A value for n-person games*. Santa Monica, CA: RAND Corporation.

Silverman, B.W., 2018. *Density estimation for statistics and data analysis*. Abingdon, OX: Routledge.

Song, Y., *et al.*, 2024. Unraveling near real-time spatial dynamics of population using geographical ensemble learning. *International Journal of Applied Earth Observation and Geoinformation*, 130, 103882.

Stevens, F.R., *et al.*, 2015. Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PloS One*, 10 (2), e0107042.

Tamiminia, H., *et al.*, 2020. Google earth engine for geo-big data applications: A meta-analysis and systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 164, 152–170.
Wang, L., Fan, H., and Wang, Y., 2020. Improving population mapping using luojia 1-01 nighttime light image and location-based social media data. *The Science of the Total Environment*, 730, 139148.
Wesolowski, A., *et al.*, 2012. Quantifying the impact of human mobility on malaria. *Science (New York, NY)*, 338 (6104), 267–270.
Xing, X., *et al.*, 2020. Mapping human activity volumes through remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 5652–5668.
Xing, J., and Sieber, R., 2023. The challenges of integrating explainable artificial intelligence into geoai. *Transactions in GIS*, 27 (3), 626–645.
Yang, Z., 2022. *Fine-scale machine learning based population mapping*. Master's thesis. Technical University of Munich.
Yao, Y., *et al.*, 2017. Mapping fine-scale population distributions at the building level by integrating multisource geospatial big data. *International Journal of Geographical Information Science*, 31, 1–25.
Yao, Y., *et al.*, 2023. A site selection framework for urban power substation at micro-scale using spatial optimization strategy and geospatial big data. *Transactions in GIS*, 27 (6), 1662–1679.
Ye, T., *et al.*, 2019. Improved population mapping for china using remotely sensed and points-of-interest data within a random forests model. *The Science of the Total Environment*, 658, 936–946.
Yu, Y.H., *et al.*, 2020. Visualization tool of variable selection in bias–variance tradeoff for inverse probability weights. *Annals of Epidemiology*, 41, 56–59.
Zimmermann, R.S., *et al.*, 2021. How well do feature visualizations support causal understanding of cnn activations? *Advances in Neural Information Processing Systems*, 34, 11730–11744.

Appendix

Table A1. Variable descriptions used in the analysis.

Variable Name	Description
p_urban_fa	urban fabric area
build area	buiding area
height	buiding height
height_neighbor	surrounding building height
p_industri	industiral area
education	education resource
p_construc	construction sites
p_land_wit	land without current use
p_water	water area
p_sports_a	sports and leisure facilities

This table provides a description of the variables used in this paper. The order of the variables is based on the order in which they appear in the text.