remote sensing

MDPI

Article

A Multi-Source Data-Driven Analysis of Building Functional Classification and Its Relationship with Population Distribution

Dongfeng Ren [1], Xin Qiu [1,2,*] and Zehua An [3]

[1] School of Geomatics, Liaoning Technical University, Fuxin 123000, China; rendongfeng@intu.edu.cn
[2] Chinese Academy of Surveying and Mapping, Beijing 100039, China
[3] Nari Group Corporation (State Grid Electric Power Research Institute), Nanjing 211106, China; anzehua@sgepri.sgcc.com.cn
* Correspondence: 472220788@stu.lntu.edu.cn

Abstract: Buildings, as key factors influencing population distribution, have various functional attributes. Existing research mainly focuses on the relationship between land functions and population distribution at the macro scale, while neglecting the finer-grained, micro-scale impact of building functionality on population distribution. To address this issue, this study integrates multi-source geospatial and spatio-temporal big data and employs the XGBoost algorithm to classify buildings into five functional categories: residential, commercial, industrial, public service, and landscape. The proposed model innovatively incorporates texture, geometric, and temporal features of building images, as well as socio-economic characteristics extracted using the distance decay algorithm. The results yield the following conclusions: (1) The proposed method achieves an overall classification accuracy of 0.77, which is 0.12 higher than that of the random forest-based approach. (2) The introduction of time features and the distance decay method further improved the model performance, increasing the accuracy by 0.04 and 0.03, respectively. (3) The correlation between the building functions and population distribution varies significantly across different scales. At the district and county levels, residential, commercial, and industrial buildings show a strong correlation with population distribution, whereas this correlation is relatively weak at the street scale. This study advances the understanding of building functions and their role in shaping population distribution, providing a robust framework for urban planning and population modeling.

Keywords: functional classification of buildings; XGBoost model; multi-source geospatial and spatio-temporal big data; Pearson coefficient; distance decay function

Citation: Ren, D.; Qiu, X.; An, Z. A Multi-Source Data-Driven Analysis of Building Functional Classification and Its Relationship with Population Distribution. *Remote Sens.* **2024**, *16*, 4492. https://doi.org/10.3390/rs16234492

Academic Editor: Salah Bourennane

Received: 2 November 2024
Revised: 23 November 2024
Accepted: 28 November 2024
Published: 29 November 2024

1. Introduction

As the fundamental building blocks of urban life, buildings serve a multitude of functional purposes, including providing spaces for residence, work, education, healthcare, and entertainment. The functional attributes of buildings not only directly reflect the spatial organization and utilization of urban areas, but also exert a profound influence on population distribution patterns [1,2]. In the field of demography, building data serve as a foundation for the spatialization of population data, which is frequently utilized as a pivotal element in understanding population distribution [3]. It has been demonstrated that the multidimensional characteristics of buildings, including the patch size, area weight, and number of floors, can provide an effective reflection of population distribution patterns [4]. A comprehensive investigation of the interrelationship between building functions and population distribution is essential for achieving the more precise spatialization of population data.

At the macro scale, current studies have demonstrated a significant correlation between population distribution and land use types. For instance, Liao Shunbao and Li Zehui discovered through regression analysis that population density is most closely

correlated with arable land, settlements, and industrial and mining land [5]. With regard to individual buildings, there is a consensus among scholars that the population distribution is predominantly concentrated in residential buildings, which are subdivided into residential and non-residential categories. These are then combined with census data to achieve a fine-grained mapping of the population distribution [6]. Furthermore, some studies have analyzed the impact of the building type on population distribution by classifying residential building types (e.g., villa, ordinary residence, dense residence, etc.) [7]. The majority of existing studies have concentrated on the analysis of site types and residential buildings, with comparatively limited attention devoted to the nuanced impact of different building types on population distribution. Furthermore, the increasing trend of mixing urban functions makes it challenging to provide a comprehensive reflection of population distribution characteristics by focusing solely on the classification of traditional residential buildings and sites. It is therefore imperative to refine the classification of building functions and explore its deeper impact on population distribution [8].

The construction of a functional classification system is of paramount importance in the realms of urban planning, the distribution of resources, and the management of disasters. The correct functional classification of buildings can facilitate a more accurate delineation of urban functional areas, thereby supporting more effective decision-making in urban development, the spatialization of populations, and the optimization of resources. However, the majority of extant methodologies for building functional classification are predicated on traditional ground surveys or expert experience, which are ineffectual and exorbitant in the context of voluminous data, thereby impeding adaptation to the accelerated urbanization of modern cities [9].

The initial studies concentrated on the depiction of urban functional areas and land use patterns. This focus was primarily due to the small and fine-grained nature of building patches, which presented significant challenges in the extraction of relevant building elements. In the early stages of research, land use was classified by comparing the spectral, spatial, and radiometric features of remotely sensed images [10,11]. In recent years, research on the identification of urban functional areas has become a prominent area of interest within the academic community. For instance, Li et al. put forth a novel framework for classifying urban functional zones, integrating urban morphological characteristics and surface temperature features [12]. Liu et al. employed a multi-feature approach, utilizing building footprints, POI (points of interest) data, and remotely sensed images, to classify urban functional zones based on the random forest model [13].

The advent of remote sensing technology, GIS, and artificial intelligence has led to a surge of interest in the study of building function classification [14]. In recent years, researchers have increasingly attempted to integrate multiple data sources and employ machine learning techniques for the purpose of building function classification. For example, the automated classification of building functions based on a random forest model has been attempted using building contour data, POI data, and remote sensing image features [1,15]. Similarly, the use of large-scale social media image data to determine building functions has also been considered [9]. Nevertheless, although existing studies have provided effective classification ideas, these methods often suffer from semantic gap problems and remain limited in terms of feature extraction and model accuracy. For instance, conventional POI feature extraction techniques frequently prove inadequate for the nuanced characterization of building functionality. Moreover, the existing, incomplete set of building features remains insufficient for the comprehensive identification of building functions.

As a typical socio-perceptual dataset, POIs provide information on the specific use of a given building or area, effectively addressing the issue of a semantic gap in remote sensing imagery [16]. Nevertheless, POIs, as a category of abstract points, are not an effective means of characterizing the functional type of each building. The distance decay model is a tool used in geography to describe the impact of cultural and spatial interactions between places. It is commonly employed to analyze accessibility to pedestrian intensity and public service facilities [17,18]. The same types of POIs tend to show aggregation effects

in a region. Integrating the effects of distance attenuation of these POIs allows for a more comprehensive assessment of the overall service function of a building in a region. This approach effectively quantifies the characteristics of the impact of POI data on the functional classification of a building and provides more effective information for the functional classification of a building. Furthermore, temporal features can be employed to reflect the characteristics of different buildings at varying times of the day, thereby reflecting their respective functions. As a significant source of temporal feature data, nighttime lighting provides a foundation for investigating urban vitality and the spatialization of a population [19,20]. The incorporation of temporal features and distance decay models into the building function classification model serves to enhance its accuracy and robustness, thereby providing a more comprehensive basis for analysis.

XGBoost (eXtreme Gradient Boosting) is an efficient machine learning model based on the gradient boosting algorithm, which is widely used in classification and regression problems [21]. The model's parallel computation, regularization, and automatic handling of missing data enable it to perform well when dealing with complex data [22,23]. In contrast, random forest is relatively slow in processing large-scale data, while models such as support vector machines (SVM) and neural networks (NN) are more complex in terms of parameterization, and less explanatory [24,25]. In recent years, XGBoost has been successfully applied to a number of fields, including remote sensing data analysis and urban functional area classification. For instance, the deployment of XGBoost for urban functional identification not only enhances the classification accuracy but also markedly optimizes the time efficiency [26]. Furthermore, XGBoost incorporates a feature importance analysis function, which is capable of automatically calculating the influence of each feature in the decision-making process. This is of particular importance for the interpretability of the classification problem. In conclusion, XGBoost represents an optimal solution for large-scale building feature classification, offering superior computational efficiency, prediction accuracy, and interpretability.

This study proposes an innovative building function classification method based on the XGBoost model, integrating multi-source geospatial and spatio-temporal big data. It employs a feature extraction method that thoroughly mines the semantic information of points of interest (POIs) and incorporates temporal features into the building function classification framework. The aim of this approach is to achieve higher accuracy in building function recognition. This study introduces two significant innovations. Firstly, based on the XGBoost model, the integration of multi-dimensional features—including building profiles, POI characteristics, image textures, and temporal features—enhances the model's overall classification accuracy by 12 percentage points. Secondly, it is the first study to analyze the correlation between building functions and population distribution at both district and street scales, revealing the differentiated mechanisms through which building functions influence population distribution across various scales. By addressing the existing gap in understanding the relationship between building functions and population distribution, this research provides a robust scientific foundation and technical support for the automated identification of building functions, precise urban planning, and optimized resource allocation. These outcomes will have a positive impact on the future management of intelligent cities.

2. Research Methods

In this study, we utilize multiple datasets, including remotely sensed imagery, POI data, building contours, and nighttime lighting data, to extract features relevant to the functional classification of buildings. The process began with preprocessing each dataset to extract key features. Building texture information was derived from remote sensing imagery, while POI data was analyzed using a distance decay function to capture socioeconomic characteristics. Building footprints were used to extract the morphology of each building, and nighttime lighting data was downscaled to reflect the nighttime vigor of the buildings in the area. These features were then combined into a composite feature vector

for each building, which was used as input to an XGBoost classifier to predict the functional type of the building. The classification accuracy was assessed through accuracy metrics and F1 scores. Subsequently, Pearson's correlation coefficient was used to investigate the relationship between the distribution of various functional buildings and the population at different spatial scales, particularly at the district, county, and street level. The technical methodology of this study is shown in Figure 1.

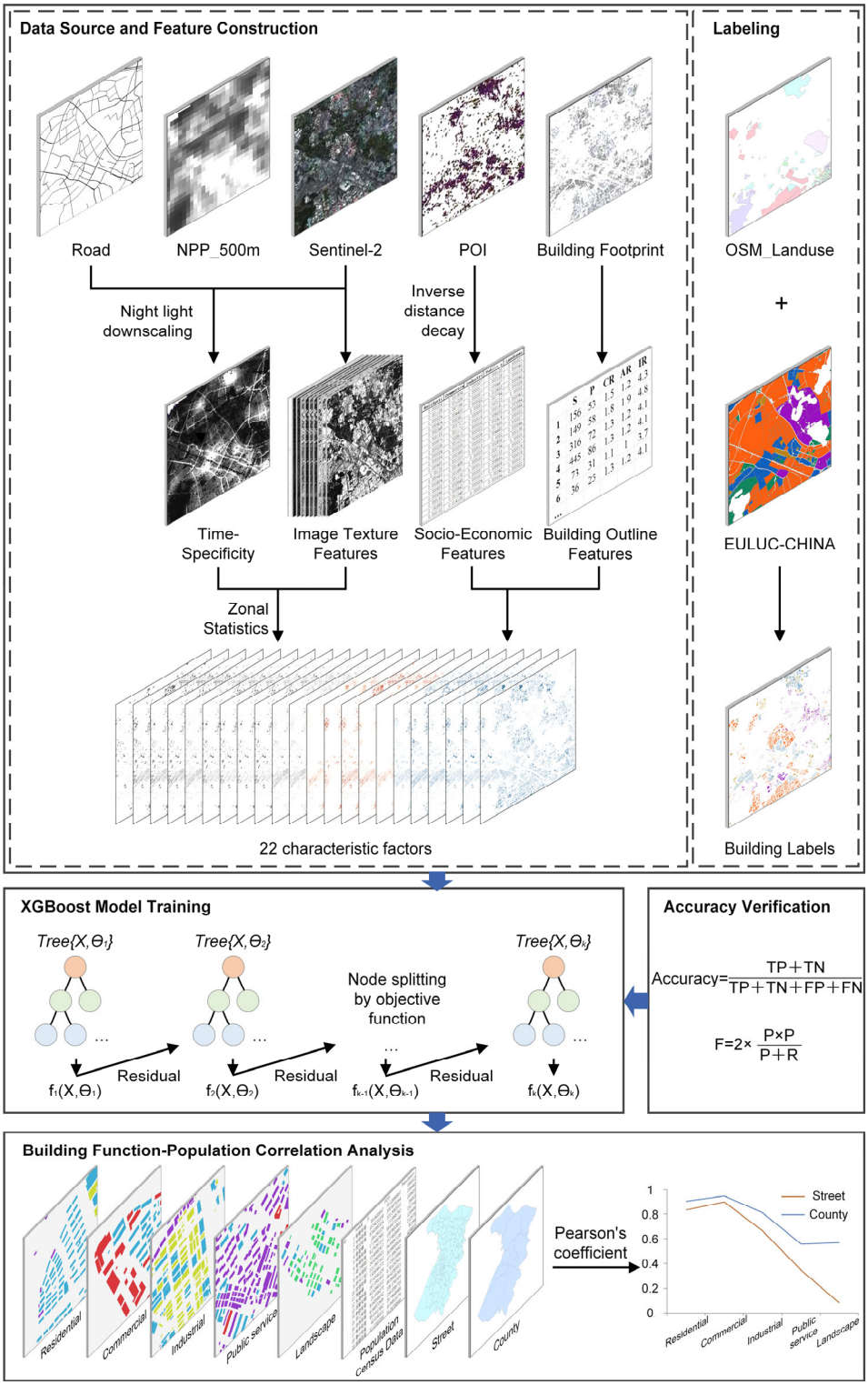

Figure 1. Technical roadmap for analyzing building function and population distribution.

2.1. Functional Definition

To ensure the precision and uniformity of the analytical outcomes, this study incorporated building footprint data and defined a singular building patch as the fundamental unit of analysis. The study area was be categorized into five types: residential, commercial, industrial, public service, and landscape buildings.

2.2. Feature Extraction

To fully explore the information contained within the data, this study integrated data from multiple sources. In addition to the commonly used socio-economic features and building profile features, image texture features and temporal features were included to comprehensively characterize the functional attributes of buildings and improve the classification accuracy and refinement. The image texture features reflect the spectral and structural information of the building surface, while the socio-economic features account for the economic activities and population distribution within the region. The building profile features describe the geometric shape and spatial layout of the buildings, and the temporal features capture the various states of the buildings at night. The integration of these multidimensional features provides a more comprehensive foundation for identifying building functions and enhances the model's classification efficacy.

2.2.1. Image Texture Feature Extraction

High-precision remote sensing images provide accurate representations of urban land use. As demonstrated by prior research [27], spectral, textural, and spatial features are extensively utilized in the fields of land use classification, building roof identification, and building extraction. In this study, image texture features were introduced to recognize building functions. This approach effectively captures the surface structure and material properties of buildings, which is crucial for distinguishing between different types of buildings (e.g., residential, industrial, commercial, etc.).

Texture features were derived from the gray-level co-occurrence matrix (GLCM) [28], a statistical model that represents the spatial relationships between pixel gray levels within an image. By calculating the joint distribution of pixel pairs over defined directions and distances, the GLCM captures essential spatial characteristics of the image's texture, allowing for a more nuanced analysis of structural patterns.

In this study, the advantages of various indices for characterizing building surface properties were leveraged to extract texture features from high-resolution remote sensing images. The selected texture features include the contrast, similarity, homogeneity, angular second moment, energy, maximum probability, entropy, GLCM mean, GLCM variance, and GLCM correlation—totaling ten indices, as shown in Table 1. To ensure consistency, the spatial resolution of all image bands was standardized to 10 m × 10 m through image band resampling. The first principal component (PC) band was chosen for the GLCM calculation, with the outputs for the ten texture indicators computed using SNAP software. Finally, the average value for each index was extracted across each image patch.

Table 1. Image texture feature parameters.

Parameter	Meaning
Contrast	Reflects the differences in surface brightness
Similarity	Evaluates the similarity of neighboring pixels
Homogeneity	Reflects the local homogeneity of the texture
Angular second moment	Describes how drastically the gray scale distribution of an image changes
Energy	Indicates the repeatability and stability of the texture
Maximum probability	Indicates the gray value that occurs most frequently in a texture image
Entropy	Reflects the texture complexity
GLCM mean	Provides the overall average of the gray values in the texture image
GLCM variance	Measures the degree of dispersion of gray values
GLCM correlation	Evaluates the linear relationship between gray values

2.2.2. Extraction of Socio-Economic Characteristics

The functions of buildings are often strongly linked to the economic activities and social behaviors characteristic of their surrounding areas. To comprehensively capture these dynamics, this study employed point of interest (POI) data as the primary source of socio-economic information. POI data represent geographic points tagged with categories generated by human economic activities, providing insights into the spatial distribution and density of different types of locations. As such, they serve as an effective means of representing functional zoning and economic activity levels in an area. In this study, the POI data are reclassified into five main categories: residential, commercial, industrial, public service, and landscape. Table 2 illustrates the classification method used.

Table 2. The POI reclassification details.

Major Category	Sub-Category
Residential	Residential communities
Commercial	Business office, catering services, shopping services, financial insurance services, health care services, accommodation services, life services, and leisure entertainment
Industrial	Companies engaged in production, factories, industrial parks, and agricultural base
Public service	Public facilities, transportation facilities, education and cultural services, sports and leisure services, medical care services, and government agencies and social organizations
Landscape	Parks and squares, and places of interest

Since points of interest (POIs) are abstract entities, they cannot directly characterize individual buildings. In this study, a distance decay model was applied to calculate the characteristic values of each POI type for each building. Typically depicted as a downward concave curve on the x-axis, the distance decay reflects a variable's decline with increasing distance, in line with Tobler's first law of geography: "everything is related, but near things are more related than distant things". Common decay functions include inverse distance, exponential, and Gaussian. Given the influence of POIs on building functions, inverse distance decay is optimal here, highlighting the effect of nearby POIs on building functions to enhance feature differentiation and improve model performance. The formula for inverse distance attenuation is as follows:

$$f(d) = \frac{1}{d^p} \quad (1)$$

where d is the distance from the building to the POI, $d \neq 0$; $f(d)$ is the attenuation value, indicating the influence at distance d. The parameter p, known as the attenuation index, typically takes a positive integer to control the rate of decay. For this study, p was set to 2, resulting in an inverse square attenuation that emphasized the impact of nearby POIs on building function. This choice highlights how proximity enhances influence, in line with the goals of the classification model.

2.2.3. Building Outline Feature Extraction

The geometric form and spatial layout of a building can provide valuable insights into its intended use and design intent. A substantial body of research has demonstrated a robust correlation between building profile features and urban functions. In accordance with existing studies [29], this study primarily extracted five key feature attributes: area (S), perimeter (P), circularity (CR), height (H), aspect ratio (AR), and irregularity (IR) of buildings. Among these attributes, area and perimeter serve as fundamental geometric features, reflecting the dimensions and extent of a building. Circularity, a measure of the complexity of a building's outline, is defined as the ratio of the area of the building to that of its smallest enclosing circle. This metric helps identify a building's compactness. Aspect

ratio quantifies the elongation of a building's shape, typically expressed as the ratio of its length to its width in the smallest enclosing rectangle. Irregularity, in contrast, assesses the complexity of the building's outline, focusing on the smoothness and intricacy of the building's boundary. The formulas for calculating circularity (*CR*), aspect ratio (*AR*), and irregularity (*IR*) are presented in Equations (2)–(4), respectively. The combination of these indicators facilitates the effective differentiation between buildings of varying functional types.

$$CR = \frac{S}{\pi \times R^2} \tag{2}$$

$$AR = \frac{L}{W} \tag{3}$$

$$IR = \frac{P}{\sqrt{S}} \tag{4}$$

where *CR* is the circularity, *S* is the area of the building, and *R* is the radius of the building's smallest external circle; *AR* is the aspect ratio, *L* is the long side of the building's smallest external rectangle, and *W* is the short side of the building's smallest external rectangle; and *IR* is the irregularity, and *P* is the perimeter of the building.

2.2.4. Temporal Feature Extraction

Nighttime lighting (NTL) data are a crucial indicator of regional socio-economic activities and population distribution [20]. Additionally, they play a pivotal role in identifying building functions. Given the significant variations in the intensity and configuration of nighttime illumination across different building types, this study utilized nighttime lighting data to aid in the classification of building functions. Many contemporary studies employ nighttime lighting data with spatial subdivisions of 500 m and 1 km; however, these intervals are often insufficient for micro-scale analyses. To enhance the resolution of nighttime lighting data from a broader scale to a building-specific level—while minimizing data distortion and noise amplification—this study adopted a multi-source data fusion approach. This method ultimately yielded a nighttime lighting dataset with a spatial resolution of 100 m.

In this study, the normalized vegetation index (NDVI) and normalized water body index (NDWI) were initially extracted from high-resolution remote sensing images. These indices were then synthesized using multi-temporal maxima and medians to emphasize the vegetation and water body features. Subsequently, point of interest (POI) data and road network data were transformed into density distributions through Gaussian kernel density estimation. The final density distributions were obtained by assigning road weights using the analytical hierarchy process (AHP). The processed data were integrated with nighttime lighting (NTL) data, which were downscaled using a regression model and refined to a resolution of 100 m through min–max normalization and Z-score normalization techniques. This approach allows for a more accurate representation of nighttime activity characteristics at the building level [30]. The overall process of nighttime lighting downscaling is illustrated in Figure 2.

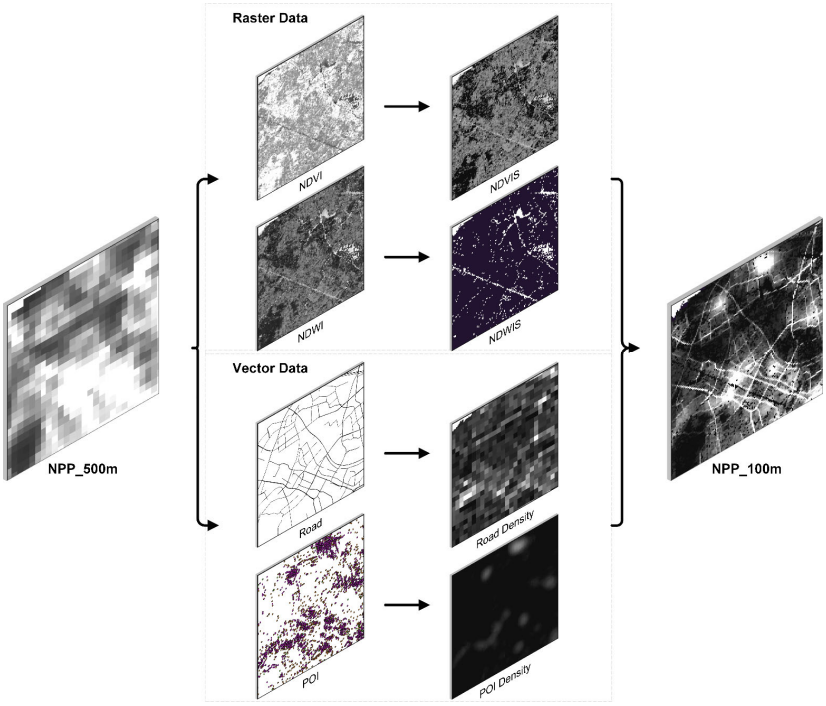

Figure 2. Flowchart of night light downscaling.

2.3. Zonal Statistics

The fundamental idea behind partitioning statistics lies in conducting raster analyses by designating one raster layer to define regions and another to provide values. When regions are initially defined by vector elements, the first step is to convert these vector elements into the raster format. This conversion often employs the image-center method, where rasterization is based on the size and values of the raster cells that correspond to the vector surface. As illustrated in Figure 3, this approach enables a seamless integration of raster data characteristics with vector features, allowing vector elements to effectively convey and represent the spatial details encapsulated within the raster data.

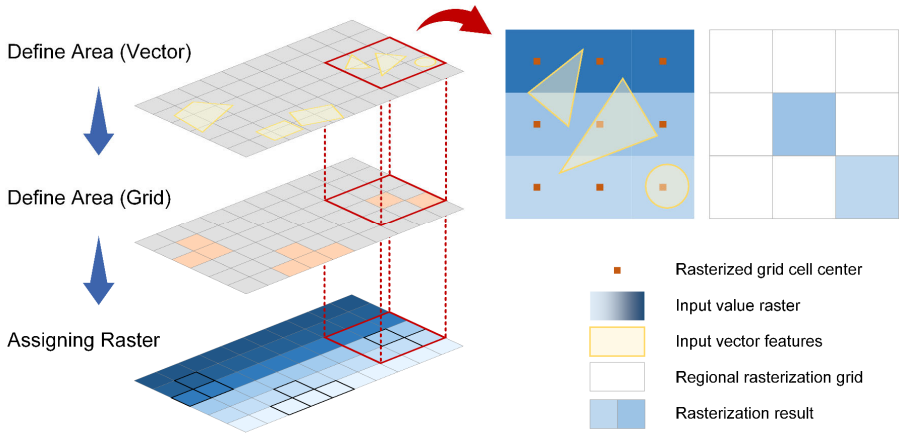

Figure 3. Schematic diagram of zoning statistics.

In this study, raster feature data within the study area were spatially aggregated using the partition statistics tool, with the statistical mean of each raster feature assigned as a new attribute to the corresponding vector unit. Since the buildings were located in relatively compact areas, it was essential to ensure the accuracy of the features extracted for

each building and to avoid excluding any relevant areas from the final output. To achieve this, the feature raster cell size was set to 5 × 5 m, which provided an appropriate level of detail for capturing building-specific characteristics while maintaining spatial precision.

2.4. XGBoost Classifier

XGBoost is a decision tree algorithm based on gradient boosting. It is a widely used tool in the analysis and modelling of large-scale data due to its high efficiency, powerful processing capability, and good classification accuracy [31]. XGBoost enhances model performance by incrementally constructing a series of weak classifiers (decision trees), which collectively improve the overall accuracy and reduce error through a weighting mechanism. Compared to traditional classification algorithms, XGBoost excels in handling missing data, mitigating overfitting, and enabling parallelized training. The computational principles underlying each component are as follows: Suppose the dataset contains n samples, each with m features. Then, the input feature matrix is denoted as $X = \{x_1, x_2, \ldots, x_n\}$ and the feature representation of each sample is $x_i = \{x_{i1}, x_{i2}, \ldots, x_{im}\}$. The classification label is denoted as $y = \{y_1, y_2, \ldots, y_n\}$, where y_i is the building function label corresponding to the i-th sample. The output is a set of probability distributions indicating the probability that each sample belongs to each type of building function, and the model makes classification decisions by maximizing this probability. XGBoost constructs a set of weak learners (decision trees) through multiple rounds of iterations, and in each round of iterations, the predictive function of the model is updated as follows:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{5}$$

In this context, $\hat{y}_i^{(t)}$ represents the predicted outcome after the t-th iteration, while $f_t(x_i)$ is the decision tree for the current iteration. Initially, $\hat{y}_i^{(0)} = 0$, and the model continuously iterates to update the predictions, gradually converging toward the true values.

The objective function of XGBoost contains a loss function and a regular term, denoted as follows:

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{t=1}^{T} \Omega(f_t) \tag{6}$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2 \tag{7}$$

In this context, $l(y_i, \hat{y}_i^{(t)})$ is the loss function used to measure the error between the predicted values and the true labels. In this study, log loss is used as the loss function for classification problems. $\Omega(f_t)$ is the regularization term that controls the complexity of the model to prevent overfitting. T represents the number of leaf nodes in the tree, while γ and λ are the regularization parameters.

In XGBoost, feature importance analysis is a vital tool for assessing the contribution of each feature within the decision tree. Common evaluation metrics include the split gain, frequency of occurrence, and size of covered samples, among others. The split gain is particularly significant as it quantifies the extent to which features contribute to reducing the loss function when nodes are split. The results of the feature importance analysis aid in identifying and filtering the most influential features in classification, providing valuable insights for model optimization. The following describes how this process works:

$$Importance(j) = \sum_{t=1}^{T} \sum_{s \in S_j^{(t)}} \Delta Loss_{j,s} \tag{8}$$

In this context, $S_j^{(t)}$ represents all the nodes where feature j appears in the t-th tree, and $\Delta Loss_{j,s}$ denotes the gain in the objective function after each split.

In this study, the XGBoost classifier integrated a variety of multidimensional features, including the image texture, socio-economic data, building contours, and temporal features, to generate five categories of labels for building footprints in Suzhou City. This was accomplished using the 2020 OpenStreetMap (OSM) land use data, supplemented by the land use category dataset (EULUC-China) and Google Maps image data. Cross-validation and hyperparameter optimization were employed to identify the optimal parameters, which include colsample_bytree = 0.9, learning_rate = 0.2, max_depth = 14, n_estimators = 300, and subsample = 0.8. Additionally, a synthetic minority oversampling technique (SMOTE) was utilized to oversample minority class samples, thereby balancing the data distribution. This study evaluated the contributions of different POI features (e.g., commercial, residential, industrial) to the classification of building functions, leveraging XGBoost's feature importance analysis to further identify the key features' roles in the model.

2.5. Accuracy Validation

The term "accuracy" is a metric for evaluating the performance of classification models. It measures the proportion of correctly classified samples among all samples in the model. Accuracy is defined as the ratio of the number of correctly classified samples to the total number of samples. The formula for calculating accuracy is given in (9).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

The terms *TP* (true positives) and *TN* (true negatives) refer to the number of samples that the model correctly identifies as belonging to the positive or negative categories, respectively. In contrast, *FP* (false positives) and *FN* (false negatives) indicate the number of samples that were incorrectly classified by the model as belonging to the positive or negative categories, respectively.

The concept of accuracy is intuitive and straightforward, serving as a measure of the overall correctness of the model's categorization. However, in the case of an unbalanced dataset with disparate categories, the accuracy metric may overestimate the model's performance, as it is less sensitive to the performance of minority categories. Therefore, this study employs a range of additional metrics (e.g., F1-score) for a more comprehensive evaluation, facilitating a more accurate assessment of the model's performance.

The F1-score is a crucial metric for assessing the efficacy of classification models, particularly in scenarios where categories are unevenly distributed. By integrating the misclassification and omission rates, it avoids the pitfalls of relying solely on the accuracy rate, providing a more accurate reflection of the model's performance. The F1-score is the harmonic mean of precision and recall, designed to account for both the proportion of correctly categorized items and the ability to capture all positive category samples. The formula is as follows:

$$F = 2 \times \frac{P \times R}{P + R} \tag{10}$$

where *F* is the F1-score coefficient, which takes values ranging from 0 to 1. The closer the value is to 1, the better the classification effect of the model; *P* is the precision rate; and *R* is the recall rate.

2.6. Building–Population Correlation Analysis

The Pearson correlation coefficient is commonly used to quantify the linear relationship between two continuous variables. In this study, the Python pandas library was employed to compute the Pearson correlation coefficient between the building functions and population distribution across various scales. This approach provides a quantitative measure of the impact of different building types on population distribution. The formula for calculating the Pearson correlation coefficient is as follows:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (11)$$

where x_i and y_i are the observed values of the building function-related and population density variables, respectively, and $\bar{x}$ and $\bar{y}$ are their means. Specifically, x is derived by aggregating the floor areas of different building types at the street and district scales, while y corresponds to the population data at the same administrative levels. The Pearson correlation coefficient ranges from −1 to 1: $r = 1$ indicates a perfect positive correlation, $r = -1$ signifies a perfect negative correlation, and $r = 0$ implies no linear correlation.

3. Study Area and Data Sources

3.1. Overview of the Study Area

Suzhou City is located in the central region of the Yangtze River Delta on the eastern coast of China. It encompasses five districts and four county-level cities, covering a total area of 8657.32 square kilometers. The city's strategic position near Shanghai, Jiaxing, Taihu Lake, and the Yangtze River creates a distinctive geographical configuration. Suzhou's convenient transportation, rich historical heritage, and thriving industries have facilitated the development of diverse building types, including commercial, industrial, landscape, and residential structures, each exhibiting various architectural styles. The distribution patterns of these buildings are influenced by multiple factors, such as the city's industrial layout, transportation planning, and population movement. As shown in Figure 4, this study selected Suzhou City as the focal point for investigation. By analyzing data from various sources, we aimed to gain a deeper understanding of the intricate relationship between the functional classification of buildings and the population distribution. This research will provide a scientific basis for urban planning and management, contributing to the sustainable development of the city.

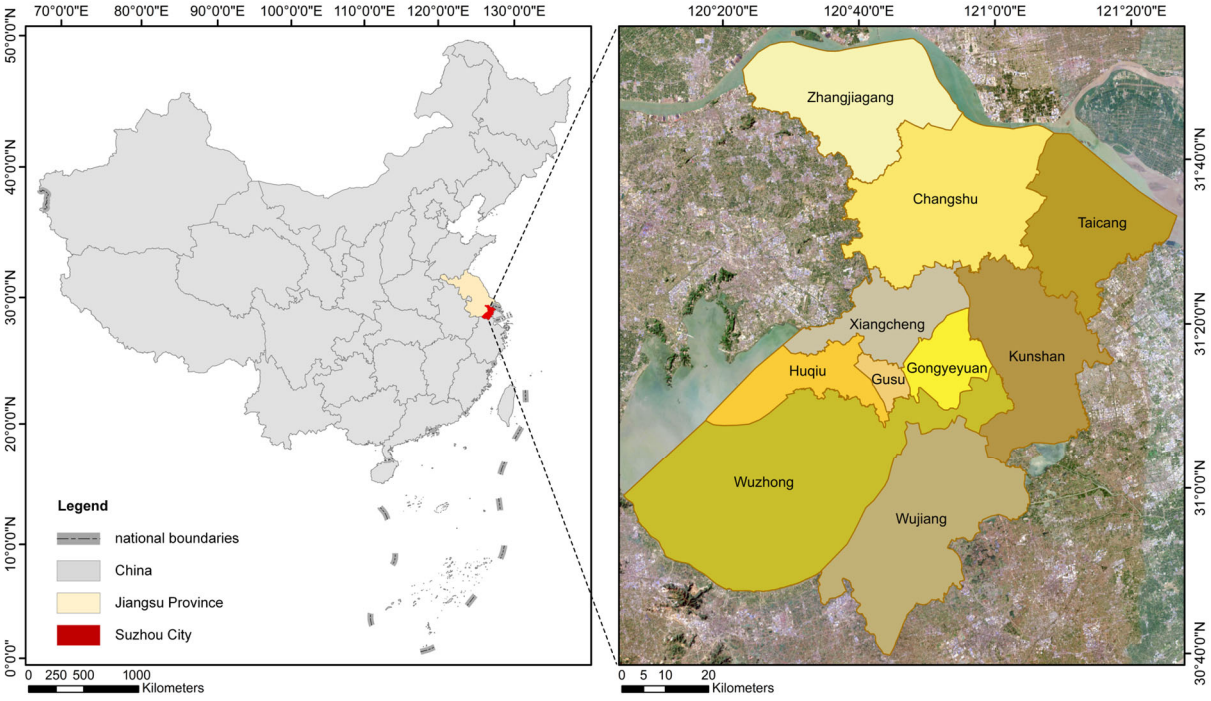

Figure 4. Administrative divisions of Suzhou City.

3.2. Data Sources

This study integrated multiple data sources, including Sentinel-2 images for texture feature extraction, building footprint data for contour feature extraction [32], and point of

interest (POI) data reflecting urban functions. Nighttime lighting data were used to capture the state of buildings at night, while various land use datasets facilitated label production [16]. Additionally, road data were employed for downscaling nighttime lighting, and Baidu map image data assisted in the validation process. Administrative boundary data and census data were also included to delineate the study area. The specific data sources are detailed in Table 3. In this study, these data were utilized comprehensively and subjected to rigorous analysis to clarify the functional classification of buildings and its relationship with population distribution.

Table 3. Data sources.

Data Name	Type	Resolution	Data Source
Sentinel-2	raster	10 m	European Space Agency (ESA)
Building footprint data	vector (surface)	/	3D-GloBFP
POI	vector (point)	/	Amap
Land use1	vector (surface)	/	OSM
Land use2	vector (surface)	/	EULUC-China2018
Road data	vector (line)	/	OSM
Night lights	raster	500 m	Resources and Environment Data Center (www.resdc.cn)
Google Map	/	/	Google Map
Administrative boundaries	vector (surface)	/	Jiangsu Provincial Department of Natural Resources
Census data	table	/	National Bureau of Statistics
Study area background image	raster	1 km	ArcGIS Online World Imagery

4. Experiments and Results

In this study, we used the global 3D-GloBFP building footprint dataset obtained by Sun Yat-sen University and other teams based on XGBoost training, which contains building vector patches and heights. In accordance with findings from previous studies and relevant regulations [33], buildings with a floor area of less than 30 square meters have been excluded, resulting in a total of 951,474 valid building patches. This study utilized a combination of labeled and unlabeled building data, ensuring that the selected samples represent a variety of spatial locations, building forms, and functional types. Additionally, unlabeled samples were included in the analysis, allowing the model to predict their functional attributes. The ratio of training samples to validation samples was set at 7:3.

4.1. Feature Extraction and Importance Analysis Results

4.1.1. Feature Extraction Results

In this study, based on the above method, a total of 22 feature parameters, including image texture, POI, building contour features and night lighting features, were selected and input into the XGBoost model, and the feature scheme is shown in Figure 5.

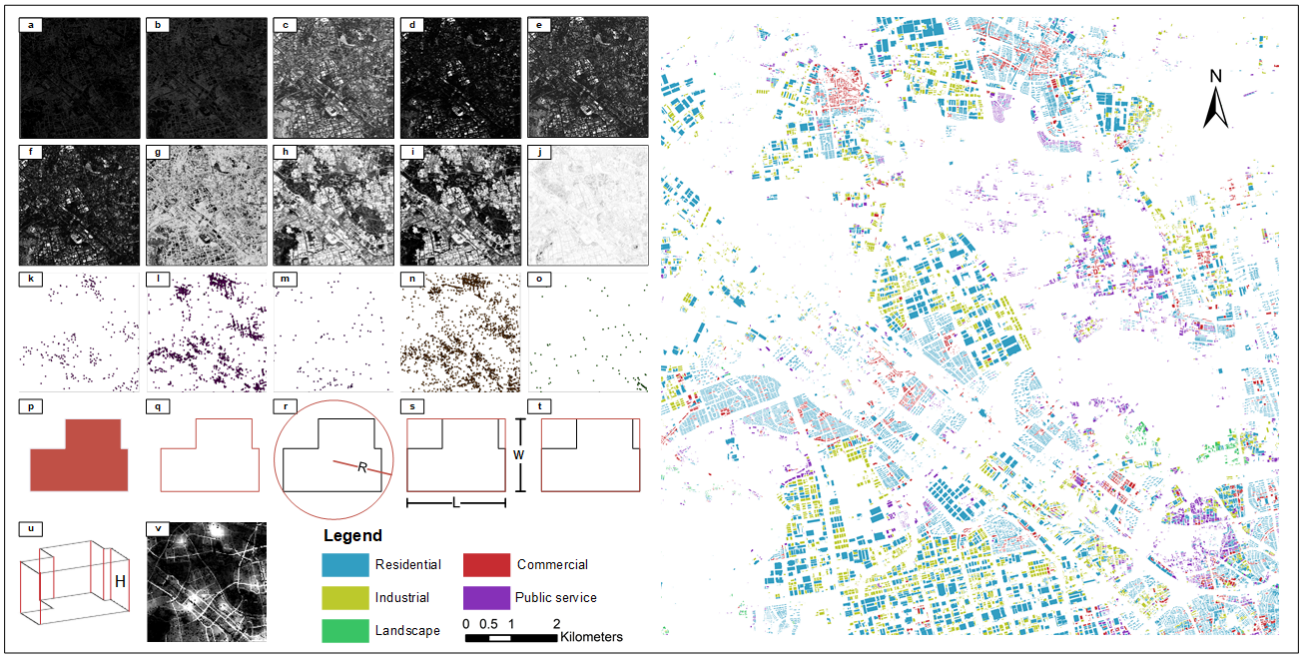

Figure 5. Feature extraction diagram. (Figure (**a**)—contrast; (**b**)—similarity; (**c**)—homogeneity; (**d**)—angular moment of two; (**e**)—energy; (**f**)—maximum probability; (**g**)—entropy; (**h**)—GLCM mean; (**i**)—GLCM variance; (**j**)—GLCM correlation; (**k**)—POI residential area; (**l**)—POI commercial area; (**m**)—POI industrial area; (**n**)—POI public service area; (**o**)—POI landscape area; (**p**)—floor area; (**q**)—building perimeter; (**r**)—building roundness; (**s**)—building minimum outer rectangle; (**t**)—building irregularity; (**u**)—building height; (**v**)—nighttime lighting).

4.1.2. Characteristic Importance Analysis

The result of the feature importance analysis based on the XGBoost model is shown in Figure 6. The result shows that there is a significant difference in the contribution of each feature to the classification of building functions.

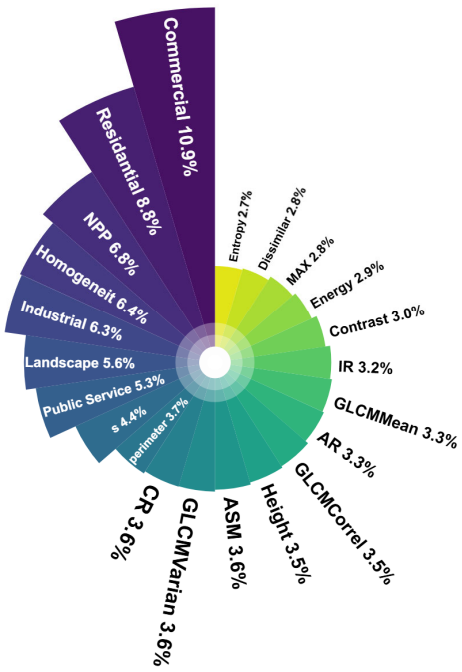

Figure 6. Feature importance analysis results.

By quantifying the percentage importance of each feature, we clarified their relative influence within the model. The commercial POIs had the highest contribution at 10.94%, demonstrating the most significant impact on the target variable. The residential POIs followed closely with 8.76%. The NPP index and homogeneity contributed 6.81% and 6.38%, respectively, indicating their critical roles in classifying building functions. Additionally, the industrial and landscape POI characteristics accounted for 6.32% and 5.63%, respectively, while the public service POIs contributed 5.31%, highlighting the strong relationship between the POI characteristics and building functions. Among the physical features, the floor area accounted for 4.42%, perimeter for 3.72%, and shape coefficient (CR) for 3.64%. The image texture features, such as the GLCM variance (3.63%) and energy (3.57%), also significantly influenced the model results. Overall, the POI features demonstrated a substantial contribution, and the innovative introduction of temporal features further enhanced the model's explanatory power. Moreover, the combination of building contours and image texture features collectively increased the model's ability to classify building functions.

4.2. Building Classification Results

Through XGBoost model training, this study classified the building functions in the study area into five categories: residential, commercial, industrial, public service, and landscape. Table 4 presents the classification accuracies, including the precision and F1-score values for each category, as well as the overall accuracy and F1-score.

Table 4. Accuracy of building classification results.

Models	Identification Accuracy					
	Residential	**Commercial**	**Industrial**	**Public Service**	**Landscape**	**Overall**
A	0.79	0.81	0.76	0.72	0.72	0.77
B	0.64	0.88	0.66	0.58	0.70	0.65
C	0.77	0.76	0.73	0.68	0.70	0.74
D	0.75	0.77	0.71	0.66	0.70	0.73
Models	**Identification F1-Score**					
	Residential	**Commercial**	**Industrial**	**Public Service**	**Landscape**	**Overall**
A	0.84	0.75	0.76	0.71	0.53	0.77
B	0.74	0.68	0.67	0.47	0.20	0.63
C	0.83	0.66	0.72	0.68	0.47	0.73
D	0.80	0.72	0.70	0.63	0.47	0.72

Notes: A represents the results from the XGBoost model using NPP features and the inverse distance decay algorithm for POI computation. B reflects the random forest model under the same conditions. C shows the results from the XGBoost model using NPP features and the kernel density algorithm. D indicates the results from the XGBoost model without NPP features but using the inverse distance decay algorithm for POI computation.

The experimental results for the building classification indicate that the model achieved an overall accuracy and F1-score of 0.77, reflecting a stable classification performance. Notably, the model was most accurate in classifying residential buildings, with a precision of 0.79 and an F1-score of 0.84, demonstrating a strong capacity to recognize this category. Commercial buildings also showed good classification performance, achieving an accuracy of 0.81 and an F1-score of 0.75; however, the slightly lower F1-score suggests some misclassification. The model exhibits the balanced recognition of industrial buildings, with both an accuracy and F1-score at 0.76. Public service buildings had a classification accuracy of 0.72 and an F1-score of 0.71, indicating consistent classification results. In contrast, landscape buildings achieved an accuracy of 0.72, but the F1-score was only 0.53, highlighting the model's relative weakness in identifying this category. This may be attributed to a lack of distinct data features or an imbalance in category samples. Overall,

the model's classification effectiveness varies across different building types, particularly excelling in residential and commercial categories.

In contrast, although Model B achieved the highest accuracy (0.88) for classifying commercial buildings, its performance was weaker in other categories, especially public service buildings (F1 score of 0.47) and landscape buildings (F1 score of 0.20). This phenomenon may be related to the limitations of the random forest model in feature selection, which tends to misclassify when feature correlations are high. Model C showed an overall stable performance (accuracy of 0.74, F1 score of 0.73) and performed well in the classification of industrial and public service buildings, suggesting that the kernel density algorithm effectively captures the spatial distribution patterns of these building types. However, its performance was still not as good as Model A. Model D, despite not using NPP features, demonstrated a similar overall performance to Model C, showing strong balance (accuracy of 0.73, F1 score of 0.72). Overall, the introduction of NPP features and the distance decay algorithm significantly improved the classification performance, especially in the classification of residential and commercial buildings (as seen in Model A). This indicates that combining multi-source features with optimization algorithms is an effective approach to improving building classification accuracy.

The results of the methodological classification of this study are shown in Figure 7:

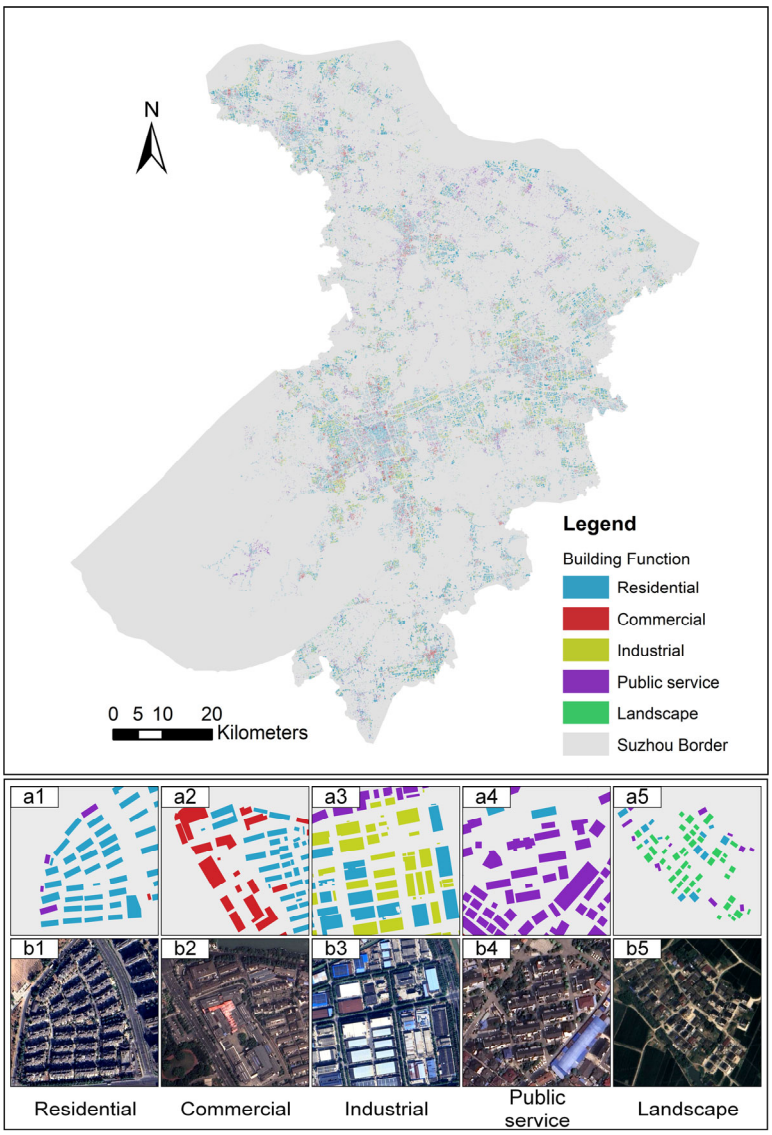

Figure 7. Comparison of (**a1–a5**) predicted results and (**b1–b5**) actual conditions (Google Map).

As illustrated in the figure, the proposed model demonstrated an optimal recognition efficacy for residential buildings with regular layouts, such as smaller districts, and showed superior performance in identifying commercial buildings. The model also achieved enhanced accuracy in recognizing structures with consistent designs, such as shops along both sides of a road. Similarly, it effectively identified factory buildings within large industrial parks, including specific dormitories associated with these factories, which corresponds with the observed higher accuracies and F1-scores across all building categories. Regarding public service buildings, the model showed partial recognition capabilities for schools and hospitals. However, it tends to confuse these with residential buildings, resulting in a slightly lower F1-score for the public service category. The identification of landscape buildings (green spaces) was relatively low, as the model often misclassifies buildings near farmland as green spaces, contributing to a diminished F1-score in this category.

4.3. The Relationship Between the Distribution of Buildings and the Population

In this study, the correlation result map (Figure 8) was generated by analyzing the building function data obtained from the predictions using the Pearson coefficient. This analysis focused on building types at two spatial scales—districts and streets—alongside the population distribution of Suzhou City in 2020.

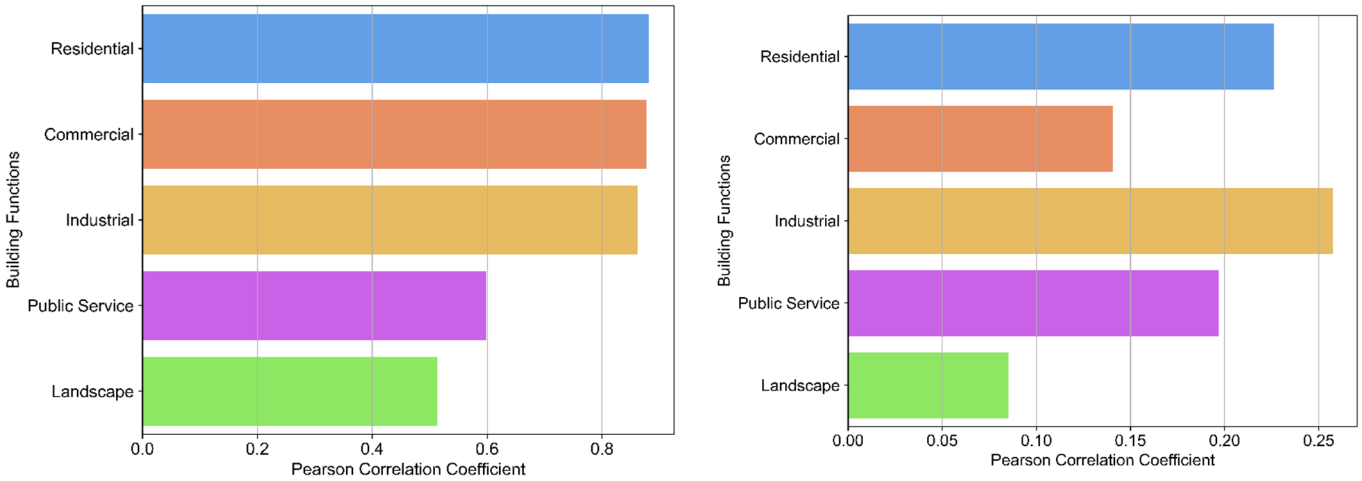

Figure 8. Analysis of the correlation between building functions and population distribution. (**a**) District and county scale; (**b**) street scale.

The results of the correlation analysis reveal a statistically significant positive correlation between population and building functions at the district and county levels. The correlation coefficient for residential buildings is 0.8816, indicating the strongest association between this building type and population. Commercial buildings follow closely with a correlation coefficient of 0.8772, highlighting their influence on population concentration. Industrial buildings also show a relatively strong association, with a correlation coefficient of 0.8620. In contrast, the correlation for public service and landscape buildings is comparatively weaker, with coefficients of 0.5977 and 0.5134, respectively. While there remains a positive correlation, the degree of influence for these building types is lower.

In contrast, the correlation between street population and building functions is weak, with overall correlation coefficients generally lower than those observed at the district and county levels. The correlation coefficient for residential buildings is 0.2261, indicating a relatively loose relationship between street population and residential structures. Similarly, the correlation coefficient for commercial buildings is 0.1405, suggesting that the influence of commercial buildings on street population is also limited. The coefficient for

industrial buildings is slightly higher at 0.2576, yet it still reflects a weak link between street population and building functions. For public service and landscape buildings, the correlation coefficients are 0.1970 and 0.0854, respectively, indicating an extremely weak connection between the street population and these building types.

5. Discussion

5.1. Importance Analysis of Features and Model Improvement

Feature importance analysis is a technique used to identify the factors that most influence model performance. Among the various features considered, POI data stands out as one of the most influential components, with commercial and residential POIs being particularly significant. These POI categories serve as strong indicators for distinguishing between commercial and residential areas, as they are directly correlated with specific building functions. However, certain POI types, such as landscape POIs and public service POIs, contribute less to the classification, likely due to their sparse geographical distribution and relatively limited impact on differentiating building functions. In addition to POI data, building profile characteristics, including the floor area, perimeter, and height, play a crucial role in distinguishing building types. Larger floor areas and perimeters are especially effective in identifying functional buildings, as they provide a direct measure of the scale and structure. Conversely, building height appears to be less influential in the classification process, possibly because of its considerable variability across different building types, which limits its ability to serve as a clear discriminating factor. Furthermore, image texture features, such as homogeneity and various GLCM (gray-level co-occurrence matrix) metrics, help capture subtle visual patterns indicative of building functions. In contrast, features like energy and entropy show lower importance, suggesting they are less effective at differentiating building types. Additionally, nighttime lighting data (NPP) prove to be a significant feature, providing valuable temporal insights into building usage patterns, particularly in commercial and residential contexts. The high importance of nighttime lighting data reflects the model's ability to capture dynamic changes in building functionality, especially in urban environments where activity levels fluctuate between day and night.

While the identified features are effective, certain limitations highlight areas for potential improvement in the model. The relatively low importance of sparsely distributed POI types (e.g., landscape POIs) underscores the need for advanced methodologies to better address such data. This could involve the adoption of more sophisticated interpolation techniques or the integration of dynamic datasets that reflect seasonality and temporal fluctuations in POI distribution. Additionally, while nighttime lighting data have demonstrated significant value, incorporating other temporal data sources—such as mobile phone signaling data or smart city sensor networks—could further enhance the model's responsiveness to real-time changes in building functions. To improve generalizability, the model should also be tested in a wider range of geographical contexts. Conducting experiments across regions with varying urban planning styles, population densities, and building types would provide a more comprehensive evaluation of the model's robustness. These enhancements, when combined, are expected to yield a more accurate and versatile framework for classifying building functions and understanding their dynamic relationship with urban population distribution.

5.2. The Relationship Between Building Functions and Population at Different Scales

The correlation coefficient between residential buildings and population at the district and county levels is 0.8816, indicating a strong influence of residential buildings on population aggregation. This result aligns with the theoretical proposition that the supply of residential buildings is directly related to the distribution of residents during the urbanization process. The correlation coefficient for commercial buildings is 0.8772, suggesting that the presence of commercial facilities not only attracts consumer populations but also promotes population growth in the surrounding areas. This finding underscores the

importance of commercial functions in enhancing urban vitality and drawing in residents. Furthermore, industrial buildings exhibit a high correlation of 0.8620, indicating that industrial zones significantly impact regional populations, especially in more industrialized areas where these facilities create substantial employment opportunities, thereby attracting an influx of people. Conversely, the correlation between public service and landscape buildings and population distribution is weaker. These buildings mainly offer services or recreational spaces that cater to a more transient user base, rather than being tied to the permanent population of a specific area. Public service facilities are often located in city centers or commercial districts, attracting a mix of local residents and external visitors or event participants, which diminishes their direct impact on the local population distribution. Landscape buildings, typically frequented by tourists or occasional users, have even less influence on residential populations.

In contrast, the correlation between the street population and the functions of nearby buildings is significantly lower. This phenomenon may be closely related to the demographic characteristics at the street level, the distribution patterns of buildings, and the socio-economic context. The street population typically comprises a diverse range of individuals who exhibit high mobility due to work and studying, which weakens the correlation between building functions and population stability. For example, while commercial buildings can provide employment opportunities, they may fail to attract the target population if the surrounding residential types and structures do not align. Moreover, the configuration of buildings on the street can lead to a mix of complementary and competing functions, facilitating population mobility between different building types. In many urban centers, there is considerable integration between commercial and residential functions, resulting in residents relying less on specific building functions. Additionally, factors such as the availability of public amenities and the accessibility of the street environment significantly influence residents' lifestyle choices, which may not be adequately captured in correlation analyses. Ultimately, the socio-economic context at the street level plays a crucial role in shaping the relationship between building functions and population. In areas characterized by uneven economic development, the availability of building functions may not meet residents' needs, creating a gap between the actual impact and theoretical expectations. In this context, the availability and affordability of buildings often exert a greater influence on residents' choices than the functional attributes of the buildings themselves.

6. Conclusions

The innovation of this study lies in its integration of multiple data sources, leveraging the advanced XGBoost classification model and sample balancing technology (SMOTE). This approach supplements nighttime lighting data as temporal features of buildings and introduces the inverse distance decay method to extract POI features. Such methods enable the co-optimization of features, parameters, and samples, resulting in the more accurate classification of building functions. Building on this foundation, this study systematically analyzed the relationship between building functions and population distribution across different scales (districts, counties, and streets), aiming to identify the reasons for the observed variations at these scales. This research enhances our understanding of the interplay between building functions and population demand, providing urban planners with data to optimize resource allocation and facility layouts. Furthermore, it offers a theoretical foundation and practical guidance for urban planning and management. The principal conclusions are as follows:

1. This study introduces a novel approach to classifying building functions in Suzhou. By integrating multi-source geospatial and spatio-temporal big data and incorporating temporal features, the model achieves an overall classification accuracy of 0.77. This represents a 0.04 improvement over the model that does not include temporal features and a 0.12 improvement compared to the results of the random forest model;

2. The POI features extracted using the inverse distance decay method more effectively represent the influence of different types of POIs on building functions. The results indicate an accuracy improvement of 0.03 compared to those obtained through kernel density analysis for calculating POI features. This evidence supports the effectiveness of the inverse distance attenuation method in capturing the impact of POIs on building functions;
3. The relationship between building functions and population distribution varies across spatial scales. At the district and county levels, the correlation is strong; however, at the street level, factors such as population mobility, diversity, competition among building functions, and socio-economic differences significantly weaken this correlation.

Author Contributions: All authors contributed to the study conception and design. D.R. was responsible for funding acquisition, methodology, project administration, resources, and supervision. X.Q. performed data curation, formal analysis, software development, and was the lead writer for the original draft and its revisions. Z.A. contributed to data curation, investigation, validation, and visualization, and assisted in writing the review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study are obtained from easily accessible public datasets, and the public can retrieve them independently. Data can also be made available upon reasonable request by contacting the corresponding author.

Acknowledgments: We would like to thank Dai for her valuable suggestions on the research, and thank the school and research institutes for their support and platform.

Conflicts of Interest: Zehua An is part of the Nari Group Corporation (State Grid Electric Power Research Institute). The authors declare no competing interests.

References

1. Du, S.; Zheng, M.; Guo, L.; Wu, Y.; Li, Z.; Liu, P. Urban Building Function Classification Based on Multisource Geospatial Data: A Two-Stage Method Combining Unsupervised and Supervised Algorithms. *Earth Sci. Inform.* **2024**, *17*, 1179–1201. https://doi.org/10.1007/s12145-024-01250-5.
2. Pan, L.; Yang, F.; Lu, F.; Qin, S.; Yan, H.; Peng, R. Multi-Agent Simulation of Safe Livability and Sustainable Development in Cities. *Sustainability* **2020**, *12*, 2070. https://doi.org/10.3390/su12052070.
3. Mei, Y.; Gui, Z.; Wu, J.; Peng, D.; Li, R.; Wu, H.; Wei, Z. Population Spatialization with Pixel-Level Attribute Grading by Considering Scale Mismatch Issue in Regression Modeling. *Geo-Spat. Inf. Sci.* **2022**, *25*, 365–382. https://doi.org/10.1080/10095020.2021.2021785.
4. Dong, N.; Yang, X.; Cai, H. A method for demographic data spatialization based on residential space attributes. *Prog. Progress. Geogr.* **2016**, *35*, 1317–1328. https://doi.org/10.18306/dlkxjz.2016.11.002.
5. Liao, S.; Li, Z. Study on Spatialization of Population Census Data Based on Relationship between Population Distribution and Land Use—Taking Tibet as an Example. *J. Nat. Resour.* **2003**, *18*, 659–665.
6. Ural, S.; Hussain, E.; Shan, J. Building Population Mapping with Aerial Imagery and GIS Data. *Int. J. Appl. Earth Obs. Geoinf.* **2011**, *13*, 841–852. https://doi.org/10.1016/j.jag.2011.06.004.
7. Shang, S.; Du, S.; Du, S.; Zhu, S. Estimating Building-Scale Population Using Multi-Source Spatial Data. *Cities* **2021**, *111*, 103002. https://doi.org/10.1016/j.cities.2020.103002.
8. Lin, A.; Sun, X.; Wu, H.; Luo, W.; Wang, D.; Zhong, D.; Wang, Z.; Zhao, L.; Zhu, J. Identifying Urban Building Function by Integrating Remote Sensing Imagery and POI Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 8864–8875. https://doi.org/10.1109/JSTARS.2021.3107543.
9. Hoffmann, E.J.; Abdulahhad, K.; Zhu, X.X. Using Social Media Images for Building Function Classification. *Cities* **2023**, *133*, 104107. https://doi.org/10.1016/j.cities.2022.104107.
10. Liu, X.; Ma, L.; Li, X.; Ai, B.; Li, S.; He, Z. Simulating Urban Growth by Integrating Landscape Expansion Index (LEI) and Cellular Automata. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 148–163. https://doi.org/10.1080/13658816.2013.831097.
11. Moran, M.S.; Inoue, Y.; Barnes, E.M. Opportunities and Limitations for Image-Based Remote Sensing in Precision Crop Management. *Remote Sens. Environ.* **1997**, *61*, 319–346. https://doi.org/10.1016/S0034-4257(97)00045-X.
12. Li, B.; Liu, Y.; Xing, H.; Meng, Y.; Yang, G.; Liu, X.; Zhao, Y. Integrating Urban Morphology and Land Surface Temperature Characteristics for Urban Functional Area Classification. *Geo-Spat. Inf. Sci.* **2022**, *25*, 337–352. https://doi.org/10.1080/10095020.2021.2021786.
13. Liu, B.; Deng, Y.; Li, X.; Li, M.; Jing, W.; Yang, J.; Chen, Z.; Liu, T. Sub-Block Urban Function Recognition with the Integration of Multi-Source Data. *Sensors* **2022**, *22*, 7862. https://doi.org/10.3390/s22207862.

14. Niu, N.; Liu, X.; Jin, H.; Ye, X.; Liu, Y.; Li, X.; Chen, Y.; Li, S. Integrating Multi-Source Big Data to Infer Building Functions. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 1871–1890. https://doi.org/10.1080/13658816.2017.1325489.
15. Xie, X.; Liu, Y.; Xu, Y.; He, Z.; Chen, X.; Zheng, X.; Xie, Z. Building Function Recognition Using the Semi-Supervised Classification. *Appl. Sci.* **2022**, *12*, 9900.
16. Gong, P.; Chen, B.; Li, X.; Liu, H.; Wang, J.; Bai, Y.; Chen, J.; Chen, X.; Fang, L.; Feng, S.; et al. Mapping Essential Urban Land Use Categories in China (EULUC-China): Preliminary Results for 2018. *Sci. Bull.* **2020**, *65*, 182–187. https://doi.org/10.1016/j.scib.2019.12.007.
17. Jing, Y.; Ma, D.; Liu, Y.; Cui, J.; Zhang, S.; Chen, Y. Decoding the Street-Based Spatiality of Urban Gyms: Implications for Healthy City Planning. *Land* **2021**, *10*, 175.
18. Wu, H. What Affects Commute Cycling in Sydney: Access, Infrastructure and Demographics. *Transp. Res. Interdiscip. Perspect.* **2024**, *24*, 101076.
19. Dong, C.; Zhang, Y.; Kang, F. *Renkou Kongjianhua Jishu*; China Population Publishing House: Beijing, China, 2024; ISBN 978-7-5101-8908-1.
20. Guo, W.; Zhang, J.; Zhao, X.; Li, Y.; Liu, J.; Sun, W.; Fan, D. Combining Luojia1-01 Nighttime Light and Points-of-Interest Data for Fine Mapping of Population Spatialization Based on the Zonal Classification Method. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 1589–1600. https://doi.org/10.1109/JSTARS.2023.3238188.
21. Liu, Y.; Zuo, R.; Dong, Y. Analysis of Temporal and Spatial Characteristics of Urban Expansion in Xiaonan District from 1990 to 2020 Using Time Series Landsat Imagery. *Remote Sens.* **2021**, *13*, 4299. https://doi.org/10.3390/rs13214299.
22. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13 August 2016; ACM: New York, NY, USA, 2016; pp. 785–794.
23. Ayman, A.; Yasser, M.; Hazem, E. Applying Machine Learning Algorithms to Architectural Parameters for Form Generation. *Autom. Constr.* **2024**, *166*, 105624. https://doi.org/10.1016/j.autcon.2024.105624.
24. Ramdani, F.; Furqon, M.T. The Simplicity of XGBoost Algorithm versus the Complexity of Random Forest, Support Vector Machine, and Neural Networks Algorithms in Urban Forest Classification. *F1000 Res.* **2022**, *11*, 1069. https://doi.org/10.12688/f1000research.124604.1.
25. Singh, R.; Biswas, M.; Pal, M. Cloud Detection Using Sentinel 2 Imageries: A Comparison of XGBoost, RF, SVM, and CNN Algorithms. *Geocarto Int.* **2023**, *38*, 1–32. https://doi.org/10.1080/10106049.2022.2146211.
26. Li, X.; Deng, Y.; Liu, B.; Yang, J.; Li, M.; Jing, W.; Chen, Z. GDP Spatial Differentiation in the Perspective of Urban Functional Zones. *Cities* **2024**, *151*, 105126. https://doi.org/10.1016/j.cities.2024.105126.
27. Huang, J.; Zhang, X.; Xin, Q.; Sun, Y.; Zhang, P. Automatic Building Extraction from High-Resolution Aerial Images and LiDAR Data Using Gated Residual Refinement Network. *ISPRS J. Photogramm. Remote Sens.* **2019**, *151*, 91–105. https://doi.org/10.1016/j.isprsjprs.2019.02.019.
28. Sathyanarayana, P.; Mohanaiah, P.; GuruKumar, L. Image Texture Feature Extraction Using GLCM Approach. International Journal of Scientific and Research Publications. *Int. J. Sci. Res. Publ.* **2013**, 3, 1–5. https://doi.org/10.5772/58692.
29. Wurm, M.; Schmitt, A.; Taubenbock, H. Building Types' Classification Using Shape-Based Features and Linear Discriminant Functions. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 1901–1912. https://doi.org/10.1109/JSTARS.2015.2465131.
30. Kuang, H.; Hu, D.; Guo, B. Mapping Regional High-Resolution Anthropogenic Heat Flux with Downscaled Nighttime Light Data. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–16. https://doi.org/10.1109/TGRS.2022.3227725.
31. Sagi, O.; Rokach, L. Approximating XGBoost with an Interpretable Decision Tree. *Inf. Sci.* **2021**, *572*, 522–542. https://doi.org/10.1016/j.ins.2021.05.055.
32. Che, Y.; Li, X.; Liu, X.; Wang, Y.; Liao, W.; Zheng, X.; Zhang, X.; Xu, X.; Shi, Q.; Zhu, J.; et al. 3D-GloBFP: The First Global Three-Dimensional Building Footprint Dataset. *Earth Syst. Sci. Data Discuss.* **2024**, *16*, 5357–5374.
33. GB50096-2011. *Design Code for Residential Buildings*; Ministry of Housing and Urban-Rural Development: Beijing, China, 2011.