buildings

MDPI

Article

Exploring the Impact of Different Clustering Algorithms on the Performance of Ensemble Learning-Based Mass Appraisal Models

Suleyman Sisman [1], Abdullah Kara [1,2,*] and Arif Cagdas Aydinoglu [1]

[1] Department of Geomatics Engineering, Gebze Technical University, 41400 Kocaeli, Türkiye
[2] Faculty of Architecture and the Built Environment, Delft University of Technology, 2600 AA Delft, The Netherlands
* Correspondence: a.kara@tudelft.nl

Abstract

Mass appraisal models are gaining use for improving valuation accuracy, yet their performance remains highly sensitive to how spatial and non-spatial data are structured before training. Clustering algorithms can be used to segment heterogeneous property groups into more homogeneous ones, potentially improving predictive performance. This study investigates the impact of different clustering algorithms, (i.e., K-Means, K-Medians and the Spatially Constrained Multivariate Clustering Algorithm (SCMCA)), on the performance of prominent ensemble learning-based mass appraisal models (i.e., Random Forest (RF), the Gradient Boosting Machine (GBM), Extreme Gradient Boosting (XGBoost) and the Light Gradient Boosting Machine (LightGBM)). Using a comprehensive real estate dataset, clustering quality is evaluated using Silhouette, Calinski–Harabasz, and Davies–Bouldin indices, and the performance of cluster-based ensemble mass appraisal models is then compared. The findings indicate that the best performance is achieved with the SCMCA–LightGBM model combination, which reached RMSE = 0.061 and R^2 = 0.722. Furthermore, it is determined that clustering-based models provide improvements of up to 7.26% in MAE, 10.61% in MAPE, and 8.40% in RMSE, depending on the combination. The results show that clustering is an effective preprocessing step that can substantially enhance the predictive performance and overall quality of mass appraisal models.

Keywords: cluster analysis; mass appraisal; machine learning; GIS; ensemble learning; artificial intelligence

Academic Editor: Pierfrancesco De Paola

Received: 3 December 2025
Revised: 20 January 2026
Accepted: 29 January 2026
Published: 2 February 2026

1. Introduction

Real estate appraisal is described as objectively determining the potential value of real estate by taking into account all rights, responsibilities and restrictions, as well as its spatial, socio-economic, environmental, planning and physical characteristics [1–4]. Mass appraisal is increasingly used in parallel with the changing real estate management needs of today's cities, which enables the efficient, consistent and cost-effective valuation of large numbers of properties [5]. It is described by International Association of Assessing Officers (IAAOs) as "the process of valuing a group of properties as of a given date using common data, standardized methods, and statistical testing" [6]. The rapid advancement of Artificial Intelligence (AI) and spatial analysis techniques has significantly contributed to the development of effective mass appraisal models, as has been the case in many other fields [7,8]. Machine Learning (ML) algorithms, evaluated as a sub-branch of AI, can provide the ability to process complex and large datasets used in mass appraisal more

quickly and accurately. The features that affect value can be measured quantitatively with the analyses provided by Geographic Information Systems (GISs). In the context of increasing data density from diverse sources, GISs play a key role in managing spatial data for mass valuation while enriching the datasets used in the process [9–11].

1.1. Literature Review

Various studies in the literature use ML algorithms to develop mass appraisal models. Iban (2022), for example, developed mass appraisal models with 1002 market samples and 43 features using tree-based ML algorithms such as Random Forest (RF), Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Gradient Boosting Machine (GBM), together with Explainable AI (XAI) methods [12]. Carranza et al. (2022) developed mass appraisal models for mapping urban land values in the city of Fortaleza, Brazil [13]. Their dataset comprises 1.320 market samples with 11 features, and their models were developed using RF, Quantile Random Forest (QRF) and GBM algorithms [13]. Deppner et al. (2023) examined the accuracy and bias of market appraisals for the U.S. commercial real estate sector [14]. They used the XGBoost algorithm and developed a model with 12,956 market samples with 50 features. They emphasized that 50 covariates can increase the appraisal accuracy and eliminate structural bias [14]. Unel and Yalpir (2023) specified various features that affect real estate value under the categories of legal, physical, locational, neighborhood and economic features [15]. They prepared a modeling dataset by analyzing the data representing the specified features using GIS and developed mass appraisal models by using the Multiple Linear Regression (MLR) algorithm for a sustainable tax system [15]. Baur et al. (2023) used MLR, Elastic Net Regression, Support Vector Regression (SVR), RF, and LightGBM algorithms to automate mass appraisal [16]. They integrated textual descriptions of real estate into models to improve the performance of the models. They conducted case studies in Berlin (30,218 market samples with 13 features) and Los Angeles (33,610 market samples with 9 features) [16]. All the aforementioned studies have focused on obtaining highly accurate value predictions using various ML algorithms.

In addition to the significant improvements achieved by ML-based mass appraisal models, most previous studies have increasingly incorporated GIS-based analyses to improve modeling performance. The following studies utilize the GIS combined with ML-based models in real estate valuation for different purposes. Cellmer et al. (2020) examined housing prices across Poland within the context of spatial heterogeneity and comparatively analyzed the performance of GWR and Multiscale Geographically Weighted Regression (MGWR) models against the Ordinary Least Squares (OLSs) model [17]. In the study, spatial autocorrelation was assessed using both global and local Moran's I statistics, confirming that housing prices were not randomly distributed but exhibited significant clustering patterns. The spatial variation in model performance was visualized through spatial coefficient distribution maps and local R^2 maps produced in the GIS. The findings revealed that while the explanatory power of the OLS model was 62.7%, this value increased substantially to 82.5% in the MGWR model. Accordingly, it was concluded that models accounting for spatial heterogeneity provide significantly more successful results in explaining urban real estate markets compared to traditional models [17]. Hosseini et al. (2024) examined spatial inequality in housing prices in Tehran and their thirty-year evolution between 1991 and 2021 using GIS-based methods [18]. Based on neighborhood-level data, methods such as Moran's I, Hot/Cold Spot Analysis (Getis-Ord Gi*), and Kriging spatial interpolation analysis were performed, revealing strong spatial clustering patterns in housing prices. The analyses demonstrated a structural spatial stratification, with high-price clusters (hot spots) concentrated in the northern parts of the city and low-price areas (cold spots) pre-

dominantly located in the south. Time-series analyses further confirmed that this spatial segregation has evolved into a persistent inequality over the three-decade period. Moreover, Kriging surface maps were used to visualize the spatial boundaries of urban rent distribution [18]. Mete (2024) designed a conceptual valuation data model by integrating ML- and GIS-based approaches [19]. In the study, a valuation dataset incorporating spatial variables was constructed using various GIS analyses, including proximity measures, topographic factors (slope and aspect), and viewshed analyses. During the modeling process, in which the RF algorithm was implemented, the integration of spatial variables resulted in a remarkably high predictive performance (R^2 = 0.86). The critical role of GIS-based analytics in improving predictive accuracy in machine learning-based assessment models was clearly highlighted [19]. Genc et al. (2025) analyzed the performance of several ML algorithms, namely MLR, RF, SVR, ANN, Gaussian Process Regression (GPR), Regression Trees (RTs), and LSBoost, in mass appraisal processes using a GIS-based approach [20]. In the study, spatial factors were analyzed through Euclidean distance analysis, while topographic variables were derived from a Digital Elevation Model (DEM) within the GIS. Model performance was evaluated by mapping the predicted values using the IDW spatial interpolation method and spatially comparing them with the distribution of actual values. The results indicated that the RF model achieved the highest performance (R^2 = 0.72); however, it was emphasized that algorithm performances varied at the neighborhood scale [20].

Improving the accuracy and performance of value predictions is possible by applying clustering techniques to group properties or geographic areas with similar characteristics in the pre-modeling phase. By identifying spatial, socio-economic and market-based homogeneity within these clusters, models can better account for localized trends and reduce prediction errors. In recent years, numerous studies have explored integrating clustering methods into mass appraisal workflows, demonstrating their potential to enhance model performance and interpretability. Aydinoglu and Sisman (2024) pointed out that clustering algorithms, another sub-branch of AI, can be used to enhance the performance of mass appraisal models [21]. They used the Spatially Constrained Multivariate Clustering Algorithm (SCMCA) and created geographical zones for mass appraisal. They developed RF-based models by using approximately 200,000 market samples with 121 features and compared the modeling performance [21]. Soltani et al. (2021) used MLR, Geographically Weighted Regression (GWR) and Geographically and Temporally Weighted Regression (GTWR) algorithms for examining spatio-temporal housing price variations in the context of mass appraisal [22]. They defined housing value clusters using SCMCA and pointed out that the features affecting the real estate value in each cluster are different, concluding that clustering algorithms are useful for policymakers in the real estate sector [22]. Wu et al. (2020) pointed out that defining housing sub-markets is critical for analyzing the urban real estate market [23]. Therefore, they used K-Means, Hierarchical and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms for determining housing sub-markets. They defined 43 sub-markets and noted that the overall forecast prediction of the models they developed in these sub-markets enhanced [23]. On the other hand, although there are some studies on clustering in mass appraisal, those that have directly integrated clustering algorithms into ML-based mass appraisal models remains limited. From a methodological perspective, creating clusters with the same characteristics provides an opportunity to assess the impact of localized model training on predictive accuracy, performance, and robustness. Additionally, creating geo-enriched datasets representing the features to be used in mass appraisal and incorporating them into the models is essential for achieving realistic results.

1.2. Aim and Methodology

This study investigates the impact of different clustering algorithms, including K-Means, K-Medians and SCMCA, on the performance of ensemble learning-based mass appraisal models such as RF, GBM, XGBoost and LightGBM. In other words, this study attempts to evaluate the effect of using the cluster analysis method and ensemble learning technique together on the performance of the valuation model, using geo-enriched datasets. This study is the first to systematically compare spatially constrained and unconstrained clustering algorithms within an ensemble learning-based mass appraisal framework and to quantify their differential impacts on valuation accuracy. It also provides empirical evidence that clustering can serve as an effective pre-processing step that significantly enhances model performance, demonstrating a methodological pathway for improving mass appraisal systems. The ultimate aim is not mass appraisal of properties but exploring the impact of different clustering algorithms on the mass appraisal models. In this study, not only the quality of clustering analysis results using established indices (i.e., Silhouette, Calinski–Harabasz, Davies–Bouldin) is evaluated, but also the performance of cluster-based ensemble learning-based mass appraisal models is compared. The methodological process followed in this context is presented in Figure 1.

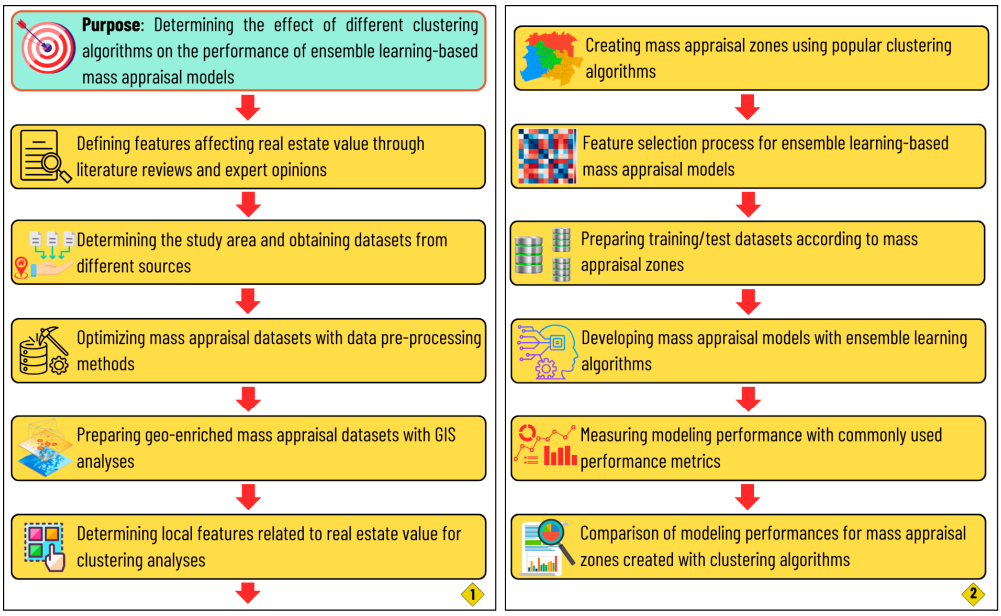

Figure 1. The methodological framework followed in the study.

The remainder of the paper is organized as follows: Section 2 describes the study area, the data-obtaining and preparing process, and the clustering and ensemble learning algorithms used. In Section 3, the conducted case study and its results are presented. The discussion and conclusions are given in the final section.

2. Materials and Methods

2.1. Study Area

The study area is a transition region that is located in the east of Istanbul and west of Kocaeli. It covers a total of five districts from two metropolitan cities of Türkiye and is surrounded by the Marmara Sea and Black Sea. Pendik and Tuzla districts are located on the Anatolian side of Istanbul, while Gebze, Çayırova and Darıca districts are located in Kocaeli (Figure 2). The surface area of the study area is approximately 792 km^2 and the total number of neighborhoods is 116 [24]. According to the official results of Address Based Population Registration System, the total district populations are 743.774 (Pendik),

293.604 (Tuzla), 407.436 (Gebze), 153.301 (Çayırova) and 227.892 (Darıca) in 2023 [25]. While Pendik is the third most populous district of Istanbul, Gebze is the most populous district of Kocaeli. Most of the study area's population and urban development are concentrated in the south, with high-density housing clusters around the airport, major motorways, metro and railway stations, and the coastline. This southern zone also hosts key urban functions such as education, health, transport, culture and industry. By contrast, the northern part of the area consists mainly of rural settlements, forests and a dam; rural activities such as agriculture and livestock farming are predominant, settlement patterns are dispersed, and population density is relatively low. According to the Socio-Economic Development Index (SEDI) Report published by the General Directorate of Development Agencies, in which all districts in the country are evaluated in six development level categories, Pendik, Tuzla and Gebze districts have the first level of development, while Çayırova and Darıca have the second level of development [26]. On the other hand, Gebze, Çayırova and Darıca districts have a cosmopolitan population structure and are important industrial areas. Tuzla district, which is located on the Istanbul–Kocaeli city boundary, has the largest shipyard areas in the country. The area between the Pendik and Gebze districts was selected as the implementation area for analyzing real estate value zones using different clustering algorithms. This was due to the fact that it encompasses diverse income groups, varying urban development patterns, and multiple real estate types, making it representative of the wider region.

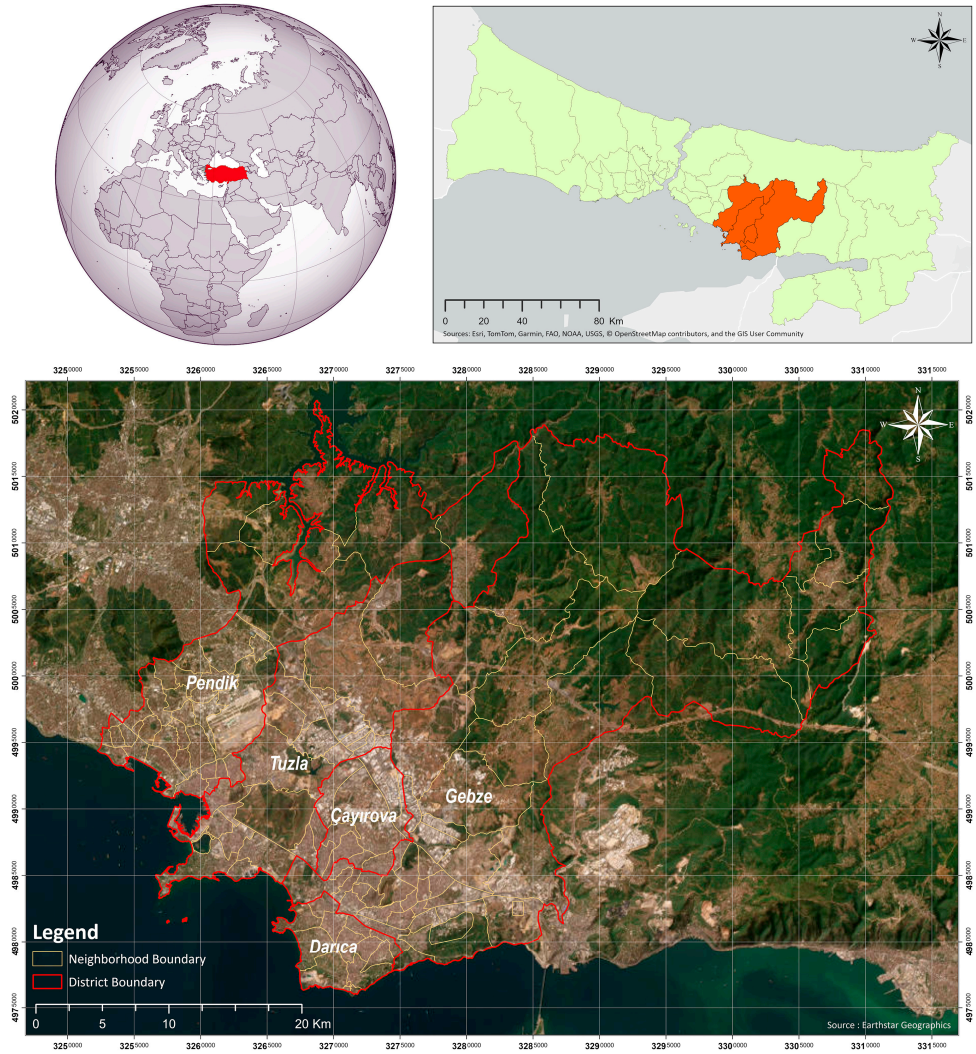

Figure 2. The study area.

2.2. Datasets

Relevant datasets obtained from different sources were used for the development of ensemble learning-based mass appraisal models and clustering analyses in the study area. The price dataset was obtained from the verified sales statements of a national real estate company [27], as, declared prices, which may not reflect market prices, are only recorded in the official register in Türkiye [28]. Hence, the dataset derived from verified market transactions is used in this study, as they are considered to substantially reflect real market prices and thus provide a more reliable representation of the market dynamics in the study area. It includes 282.953 market samples recorded between 1 January 2020 and 31 December 2023. This dataset contains a large amount of features such as verified sale price, geographical location (latitude and longitude), date of update, area size, number of rooms, number of living rooms, number of bathrooms, number of storeys in the building, storey on which the property is located, building age, type of heating system, direction of property facade, landscape, as well as the existence of elevator, car park, gym, security, generator and pool. This dataset was complete, with no missing values; therefore, no imputation was required. On the other hand, updating of the price feature is very important for reliable mass appraisal models. Therefore, an update was performed to the real estate price feature in the dataset as the price feature covering the period 2020–2023 contains temporal differences due to inflation. The Central Bank of the Republic of Türkiye (CBRT) publishes monthly Residential Property Price Index (RPPI) statistics for cities in order to monitor price changes in the Turkish housing market. This index, which is created using the hedonic regression method, can be used effectively for temporal updates to the data [29]. In this study, these index data [30] were obtained for Istanbul and Kocaeli cities (Figure 3) and RPPI update rates were calculated with the equation $RPPI_{December\ 2023}/RPPI_{Transfer\ Date}$. These rates are multiplied by the price feature and temporal differences were eliminated. In other words, all data were updated to December 2023, representing the most recent available period. It is imperative to acknowledge that, in accordance with this update process, the ramifications of the pandemic period, high inflation, and broader economic fluctuations are considered to be incorporated into the price variable through the official index.

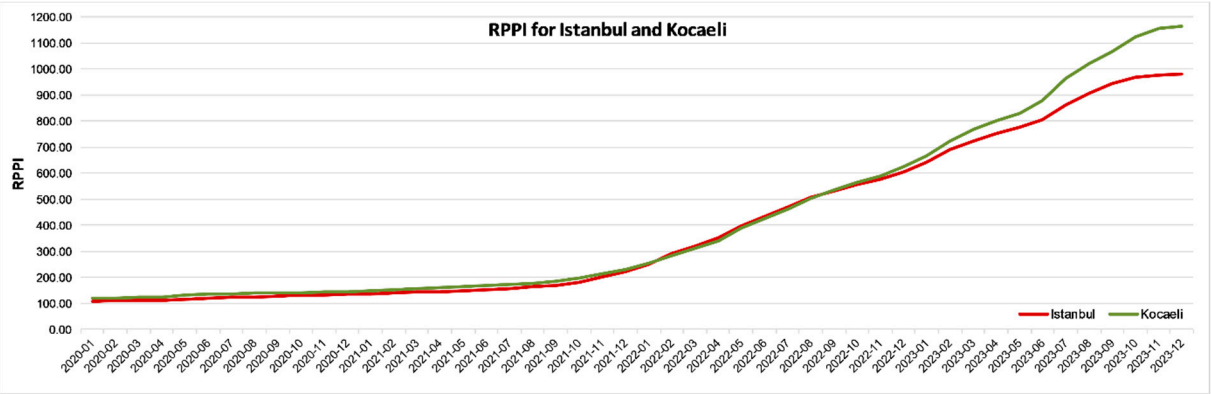

Figure 3. CBRT-RPPI for Istanbul and Kocaeli between 1 January 2020 and 31 December 2023.

The real estate market activity dataset was obtained through the Land Query Application of the General Directorate of Land Registry and Cadastre (GDLRC). The features in this dataset represent the buying and selling activity in the housing market in the study area. It includes features such as trading density, number of sales and number of mortgaged sales for condominium units [31].

The local features dataset was obtained from the database of the TUBITAK Project No. 122R021, funded by the Scientific and Technological Research Council of Türkiye

(TUBITAK). This dataset provides a local assessment of the neighborhoods in which the real estate is located, representing the data that is used to evaluate the impact of local socio-economic development characteristics on the real estate value. It includes features such as income level distribution, expenditure distribution, education level and marital status at the neighborhood and district level. Population data according to the neighborhood populations and age groups that can be considered within the scope of this dataset were obtained through the Turkish Statistical Institute (TurkStat) platform, while the Socio-Economic Development Index data of the districts were obtained from the General Directorate of Development Agencies reports [25,26].

Urban functions datasets were obtained from the relevant departments of Istanbul and Kocaeli Metropolitan Municipalities. The features in these datasets represent the data that will be used to evaluate the proximity of real estate to various urban facilities. It includes location information of urban Points of Interest (POI) such as educational facilities, health facilities, transportation facilities, shopping and trade facilities, cultural facilities, public service facilities, green areas, entertainment and sports facilities, accommodation facilities, industrial facilities, and religious facilities.

Two distinct sources were used to obtain environmental datasets containing air quality and meteorological data. The National Air Quality Monitoring Network open data portal was utilized as the source of air quality data between January 2020 and January 2024 [32]. It includes the calculated Air Quality Index (AQI) feature based on stations. Meteorological data were obtained from the Meteorological Data-Information Presentation and Sales System platform of the Turkish State Meteorological Service between January 2020 and December 2023 [33]. It includes features such as temperature, rainfall, and humidity based on stations.

The energy dataset was obtained from the Directorate General of Vocational Services for building energy statistics [34]. It includes features such as number of energy identity certificates, primary energy consumption (kWh/year), acclimatized area (m^2) and renewable energy contribution (kWh/year) according to districts. Detailed information on the datasets used in this study, including their variables and data levels, is presented in Appendix A.

All of these datasets were organized in a GIS environment and exported to a geo-database. This made all the datasets available for geographical analysis and modeling processing. It is evident that the real estate value varies depending on many spatial, environmental and local features. In order to develop reliable mass appraisal models, all these features affecting real estate value must be evaluated with geo-analytical approaches. Datasets representing features can be produced using various analysis techniques in GIS [35,36]. In this study, pixel-based geographical analyses were performed using the geodatabase. In this context, analyses such as proximity, interpolation and density were carried out in accordance with the structure of the features. A number of the analysis results are presented in Figure 4. The analysis results of all spatial and local features were geographically joined to each market sample to construct a geo-enriched dataset. Then, since the features in the dataset were in different units, the dataset was standardized using the maximum normalization method. In this way, a geo-enriched dataset suitable for developing ensemble learning-based mass appraisal models were prepared.

2.3. Clustering Algorithms

Clustering analyses are defined as grouping observations or features in a dataset by considering their similarities among themselves. Most of the clustering analysis methods use the distances between data. Although many clustering methods are used for different applications, these methods are mainly used to subdivide datasets into subsets by utilizing similarities or differences between features. Clustering analyses, which are among the

multivariate statistical methods, used to group data in datasets containing a large number of features and to compare the groups created, are commonly utilized due to their easy applicability and easy understanding of the results [37–41]. In this study, popular clustering methods were used to determine urban areas with similar characteristics in order to enhance the performance of mass appraisal models. The methods used are explained in the following subsections.

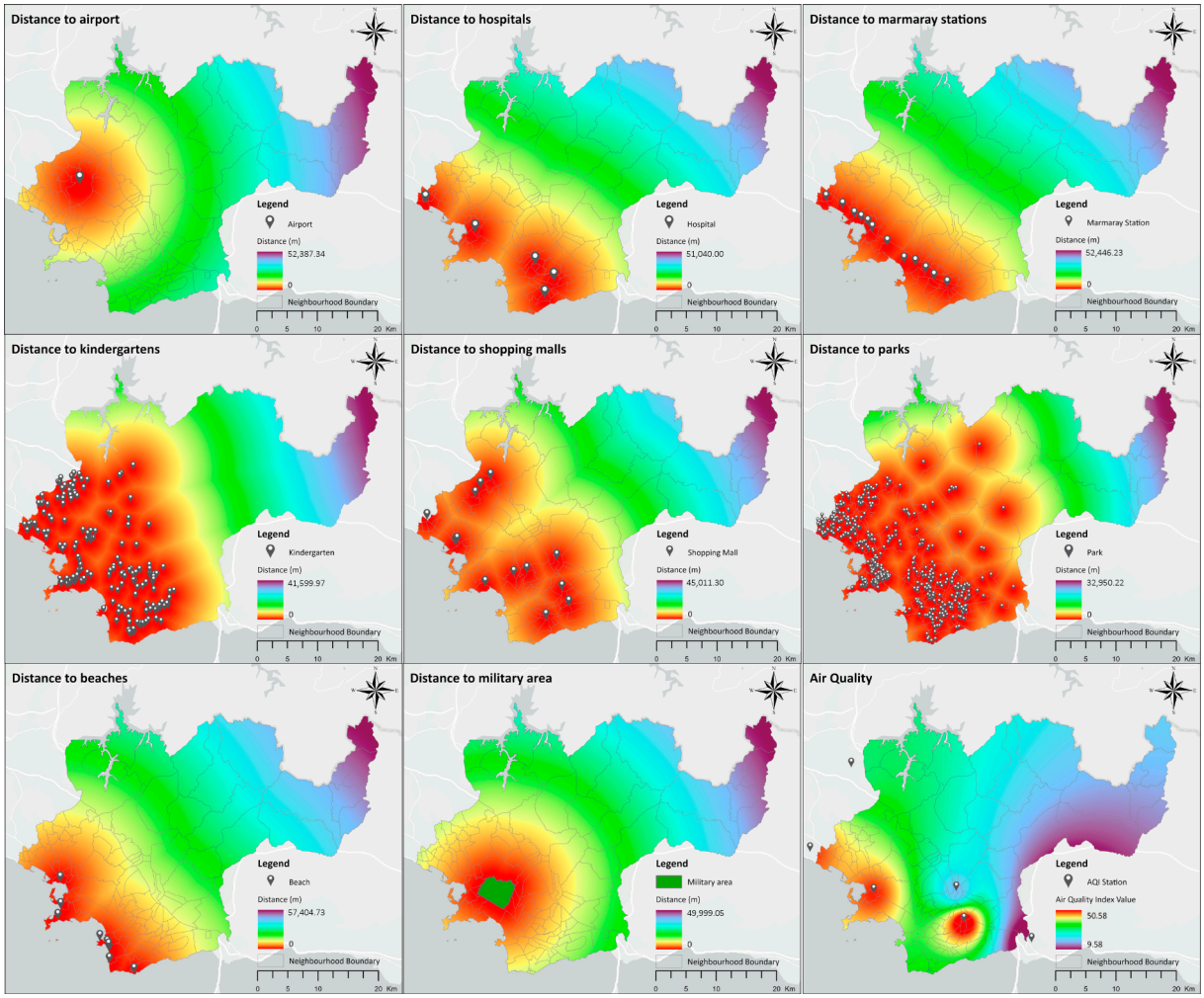

Figure 4. Examples of geographical analysis for some of the spatial features.

2.3.1. K-Means Algorithm

K-Means, developed by James MacQueen, is the most widely used unsupervised ML algorithm for identifying clustering patterns in datasets [42]. It is commonly utilized in many fields due to its ability to process big datasets quickly and effectively. The aim is to ensure that the clusters produced at the end of the clustering process have maximum similarity within clusters and minimum similarity between clusters. The clustering process develops around the mean vectors of the features, which are used as cluster centroid. It aims to divide the data points in the dataset into a specified number of clusters in a way that minimizes the sum of squares within the cluster [43,44]. The algorithm starts by associating each data point with a cluster and then tries to optimize these clusters. Through repeated iterations, it minimizes the distances of the data points to the clusters at each step.

Mathematically, it is formulated as follows: A dataset consists of n number of data points, each of which is d-dimensional, where $X = \{x_i | \, i = 1, 2, 3, \ldots, n\}$. Considering that this dataset will be divided into k number of clusters, these clusters are denoted as

$C = (C_1, C_2, C_3, \ldots, C_k)$ each cluster being C_j. Due to the aim of this algorithm, which is to minimize the total errors between the clusters, the objective function $J(C)$ is given with Equation (1):

$$J(C) = \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (1)$$

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (2)$$

where x_i indicates data points, C_j indicates the set of data points assigned to the *j-th* cluster, μ_k indicates the centroid of the *j-th* cluster and $\|x_i - \mu_j\|^2$ indicates the squared Euclidean distance between x_i and cluster centroid μ_j.

The basis of this algorithm, which is categorized under partitioning methods, is the division of a dataset containing n observations (data points) into K clusters. (1) K is the number of clusters to be created and is determined by the users. (2) Then, temporary cluster centroids (centers) are randomly selected. (3) The distance of each data point to a particular centroid is usually calculated using Euclidean distance. Each data point is assigned to the cluster represented by the nearest centroid. (4) New cluster centroids are determined using the data points included in the clusters. New centroids are calculated by averaging the data points in the created cluster with Equation (2). Steps 2 and 3 are repeated until the variation of data points between clusters decreases or until a certain stopping threshold is reached. (5) The algorithm stops working when the membership of the data points in their respective clusters stabilizes. Thus, each data point is grouped with the nearest centroid, and the final clusters are determined so that the centroids do not change any more. This process is summarized in Figure 5.

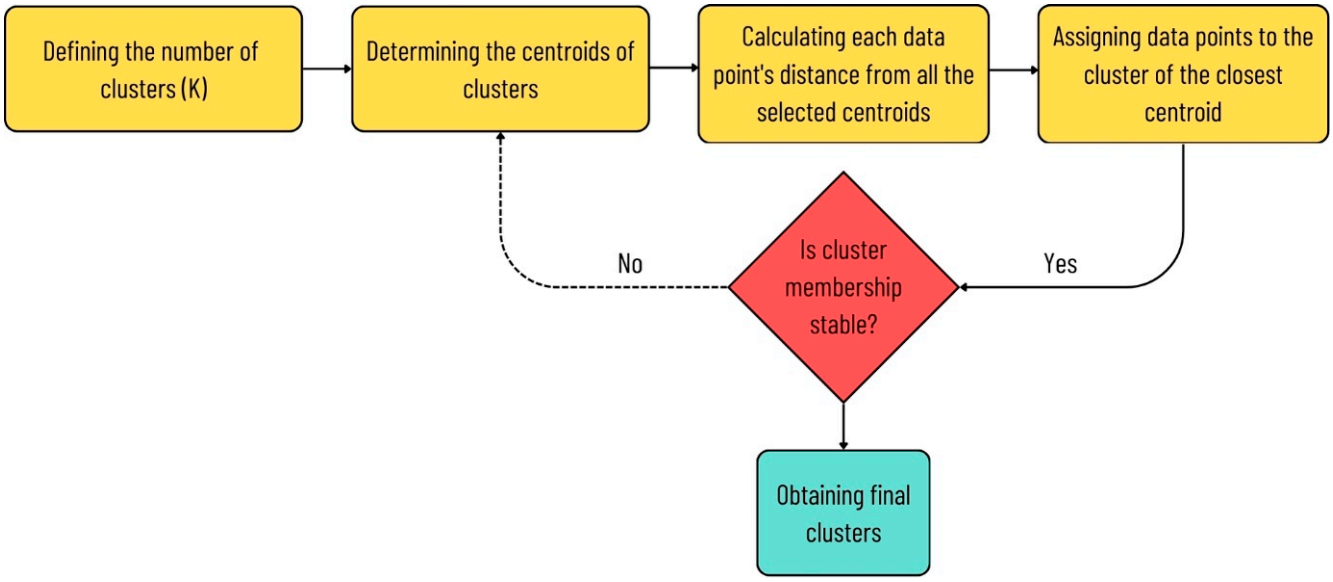

Figure 5. The general implementation process of clustering with the K-Means algorithm.

2.3.2. K-Medians Algorithm

Today, big datasets collected for many different applications may include outliers. This can make algorithms such as K-Means sensitive to outliers, which can affect the results. K-Medians is a clustering algorithm introduced by [42] and further developed by [45] to deal with this problem [45]. K-Medians is known as a special variant of the K-Means algorithm and was proposed to improve some limitations of the K-Means algorithm. It is used effectively when the features in the dataset are asymmetric. This algorithm, which has a

similar implementation principle to the K-Means algorithm, aims to divide the dataset into k clusters. However, the median values of the data within the cluster are used to determine the cluster centroids. As in the K-Means, each data point is assigned to the cluster centroid closest to the calculated median value. Here, the distance is usually calculated by using the Manhattan distance (L1 norm), which is different from K-Means [46]. Almost all other implementation steps are the same as K-Means. In this context, the objective function $J(M)$ for the K-Medians algorithm is given by Equation (3):

$$J(M) = \frac{1}{n}\sum_{i=1}^{n} \min_{j=1,2,\ldots,k} \left\| x_i - c_j \right\| \quad (3)$$

where n indicates a total number of data points, k indicates a total number of clusters, x_i indicates data points, c_j indicates cluster centroid of the *j-th* cluster and $\left\| x_i - c_j \right\|$ indicates Manhattan distance between x_i and cluster centroid c_j.

2.3.3. Spatially Constrained Multivariate Clustering Algorithm

In many clustering algorithms that have been widely used, clusters are determined by the feature values in the datasets and various distance metrics. Specifically, location information regarding the data is ignored and only the similarities of the data included in the analysis are determined based on their feature values. To overcome this locational limitation of clustering algorithms, AssunÇão et al. (2006) proposed a network-based Spatially Constrained Multivariate Clustering Algorithm (SCMCA) known as Spatial C(K)luster Analysis by Tree Edge Removal (SKATER) [47]. It is an effective algorithm that aims to group objects in a dataset by taking into account both spatial constraints and multivariate features. It is based on the Minimum Spanning Tree (MST) algorithm derived from the spatial weight matrices and focuses on pruning the created tree structure [48]. The weights in these matrices are equal to the pairwise dissimilarity between data points, which turn into the edges in the graphical representation of the weights (i.e., data points are the nodes). Considering that a network structure consists of nodes and edges, and nodes are connected to each other by edges, MST focuses on connecting nodes with the lowest cost. The edge cost calculation is performed by evaluating the differences between geographical neighboring regions for the relevant feature concerning location [49]. Although various algorithms such as Prim, Kruskal and Sollin are used in creating network structures with MST and reaching a solution, the Prim algorithm is used in SCMCA. In this context, the implementation process of SCMCA based on MST is considered. (1) Firstly, the geographical centers of the data are determined. (2) A graph explaining the neighborhood relationships between the identified centers is created. Using this graph, a spanning tree structure is created that explains both the spatial relationships and the similarity for the features included in the analysis. (3) Data with similar characteristics on the created structure, where the difference is minimized, are included in the same cluster, while data with characteristics where the difference is high are assigned to another cluster. (4) Finally, spatially constrained clusters are determined by pruning the tree structures where they begin to differ.

2.4. Clustering Validity Indices

Measuring the quality of clustering results is essential to provide that the identified clusters are meaningful, well-separated, and internally cohesive. Among the many validity indices proposed in the literature, the most commonly used measures, which are summarized in Table 1, are the Silhouette Index, Calinski–Harabasz Index, and Davies–Bouldin Index. Each of these indices focuses on different perspectives of the cluster characteristics, such as cohesion, separation, and overall compactness, providing complementary insights for evaluating and comparing clustering results.

Table 1. Commonly used validity indices for clustering analysis.

Validity Index	Description	Formula
Silhouette Index	It is a validity measure used to evaluate the cohesion and separation of clusters by analyzing how similar each data point is to its own cluster compared with other clusters. It has an index value in the range of [−1,1], and higher index values mean that data points are well matched to their assigned clusters while being distinctly separated from neighboring clusters [39,50]	$SI = \frac{1}{n}\sum_{i=1}^{n} s(x_i)$ where $s(x_i) = \frac{b(x_i)-a(x_i)}{max\{a(x_i),b(x_i)\}}$ $a(x_i)$: the average distance between x_i and all other data points in its own cluster (within-cluster dissimilarity) $b(x_i)$: the minimum average distance between x_i and all data points in any other cluster (nearest-cluster dissimilarity)
Calinski–Harabasz Index	It is a statistical measure used to evaluate the validity and internal consistency of clustering results. The index simultaneously assesses how distinctly the clusters are separated and how tightly the data points within each cluster are grouped. A high index value indicates strong within-cluster cohesion and clear separation between clusters [51,52]	$CH = \frac{\sum_{i=1}^{k} n_i \lVert m_i - m \rVert^2}{\sum_{i=1}^{k} \sum_{x \in C_i} \lVert x - m_i \rVert^2} \times \frac{N-k}{k-1}$ n_i: the number of data points in cluster C_i m_i: the centroid of cluster C_i m: the overall mean of all data points x: a data point belonging to cluster C_i k: the number of clusters N: the total number of data points
Davies–Bouldin Index	It is a clustering validity measure that evaluates the quality of a clustering solution by combining two components: the first term penalizes high intra-cluster variance, while the second term rewards large inter-cluster separation. It quantifies how compact clusters are and how distinct they are from each other, with lower index values indicating better clustering performance [53,54]	$DB_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \max_{i \neq j}\left\{ \frac{S_i + S_j}{d(\overline{x}_i, \overline{x}_j)} \right\}$ where $S_i = \frac{1}{n_i}\sum_{j=1}^{n} d\left(x_j, \overline{x}_i\right)$ k: the number of clusters S_i: the average distance between the cluster's center and all of its elements (within-cluster scatter) n_i: number of data points in cluster i $d\left(\overline{x}_i, \overline{x}_j\right)$: the distance between the centroids of clusters i and j (between-cluster separation)

2.5. Ensemble Learning Algorithms

Although there are numerous algorithms available for developing regression models, ensemble learning algorithms, which have recently gained considerable popularity, are widely used due to their high modeling performance and strong generalization capabilities [55–58]. Ensemble learning is defined as a ML approach that combines two or more learners (e.g., decision trees, artificial neural networks, etc.) during the modeling process to produce more accurate predictions. In other words, an ensemble learning model combines multiple individual models to achieve more accurate predictions than a single model could produce. Ensemble learning is particularly effective in overcoming the weaknesses of individual models, especially in complex and highly multivariate datasets. Ensemble learning approaches are generally categorized into three main types: Bootstrap Aggregating (Bagging), Boosting, and Stacking [55,59]. While there are various algorithms within these three main categories, the most popular ones are the RF, GBM, LightGBM, and XGBoost algorithms, which are utilized in this study.

2.5.1. Random Forest (RF)

Random Forest (RF) is one of the popular ensemble learning algorithms based on decision trees, developed by Breiman and widely used for various classification and regression applications. In the modeling process of the RF algorithm, a decision forest of decision trees is created by combining a large number of uncorrelated trees generated independently of one another [21,60,61]. By means of the generated decision forest, the

predictions made by each individual tree are combined to produce a more accurate final prediction. When the fundamentals of the RF algorithm are examined, it can be seen that it combines the Bagging approach, in which decision trees are developed independently based on the training datasets, with the Random Subspace technique, in which a smaller subset of randomly selected variables from the entire dataset is used and the variables that achieve the best split at each tree node are chosen [62]. The RF algorithm stands out due to its advantages, including the ability to assess variable importance and model interactions, the independent development of decision trees on their respective data subsets to minimize the risk of overfitting, and strong generalization capability. Models can be developed that produce effective predictions, particularly when datasets contain variables with high explanatory power [63].

2.5.2. Gradient Boosting Machine (GBM)

Gradient Boosting Machine (GBM) is an ensemble learning algorithm proposed by Friedman (2001) that demonstrates high modeling performance for regression and classification problems [64]. It can also be considered a flexible adaptation of the AdaBoost algorithm for both regression and classification tasks. The GBM algorithm generally relies on combining weak learners, such as decision trees, to construct a stronger predictive model [65]. In contrast to the RF algorithm, the decision trees in GBM are dependent on one another. In the GBM algorithm, a sequence of weak learners, each forming a single predictive model, is trained iteratively. For each model in the sequence, prediction errors are calculated, and the subsequent model in the sequence is trained to optimize the errors of the preceding model. This modeling process is performed iteratively, with each iteration aiming to improve the prediction error of the model. Prediction errors are determined using differentiable loss functions, taking into account the structure of the regression or classification problem. The GBM algorithm employs the gradient descent method to minimize the loss function and thereby improve predictions. This method, commonly used in optimization problems, enables the calculation of the minimum point of the loss function and determines how to improve modeling performance by iteratively updating the model parameters using the gradient information of the function [66,67].

2.5.3. Extreme Gradient Boosting (XGBoost)

The Extreme Gradient Boosting (XGBoost) algorithm is described as a scalable and platform-integrable version of the GBM algorithm, optimized by [68] to improve computational speed and model prediction performance. In the XGBoost algorithm, algorithmic enhancements (such as regularization, cross-validation, sparsity awareness), system-level optimizations (software, hardware, etc.), and parallel and distributed computing capabilities collectively shorten the model training process and accelerate the time required for models to reach a solution. The parallel computing capability of the XGBoost algorithm enables efficient use of memory resources and processing time, and it has been reported to operate up to ten times faster compared to other ML algorithms. Furthermore, the XGBoost algorithm, which is resistant to overfitting, demonstrates high performance, particularly in modeling studies conducted with big datasets [68–70].

2.5.4. Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LightGBM) is an ensemble learning algorithm developed by Microsoft in 2017, which is among the novel decision tree-based ensemble boosting methods. Similar to other decision tree-based methods, it can be utilized effectively for classification and regression. It was designed to address the inefficiency of gradient-boosting decision tree models when dealing with high-dimensional features and large datasets. Due to its characteristics such as parallel computation and low memory

consumption, it stands out compared to other gradient boosting models and provides high modeling performance and low training time in big datasets. It converts continuous features in the dataset into discrete features to reduce the computational cost. Also, it has advantages such as lower memory usage, higher modeling performance, ability to process big data effectively, support for parallel learning and GPU learning, and parameter optimization against overfitting. It is stated that it works 20 times faster compared to other ML algorithms [71–73]. While most decision tree algorithms grow decision trees based on a level-wise (depth) strategy, LightGBM uses a leaf-wise (best-first) growth strategy. In the level-wise growth strategy, the tree grows balanced at each level. It aims to achieve a more organized structure by maintaining the balance of the tree and it allows for the expansion of all nodes at every level. In the leaf-wise growth strategy, the decision trees grow by focusing on the leaf nodes. This strategy ignores the stability of the tree and splits leaf nodes at each step, increasing the learning rate of the model and focusing on reducing the prediction error. At each splitting step, it chooses the division that will reduce the prediction error the most, which allows the model to learn more effectively. In some cases, with small datasets the leaf-wise growth strategy can cause overfitting; therefore, LightGBM has the max_depth parameter to limit the tree depth. This parameter can be specified by users. Also, LightGBM utilizes two novel approaches, Gradient-based One Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), to address the constraints associated with the conventional histogram-based method employed in gradient boosting decision tree algorithms. To overcome these constraints and enhance the algorithm's overall performance, GOSS and EFB are utilized [74,75]. Considering all these reasons, the LightGBM algorithm is widely used in practice due to its performance against traditional statistical approaches for mass appraisal models [76,77].

2.6. Performance Metrics

In regression models developed using ensemble algorithms, there are various performance measurement metrics to measure the prediction accuracy. Among these metrics, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), Root Mean Squared Error (RMSE) and R-Squared (R^2) are widely used [12,78–81]. These metrics are utilized in mass appraisal to evaluate the differences between the known market value and the value predicted from the developed models. They are formulated with the equations given below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

$$MAPE = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \tag{5}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{6}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2} \tag{7}$$

In the equations, y_i denotes the market values, $\hat{y}_i$ denotes the predicted values, $\overline{y_i}$ denotes the average market values, $i = 1, 2, 3, \ldots, n$ and n denotes the number of samples used in the model.

2.7. Overview of the Datasets and Applied Methodological Steps

To improve the clarity and transparency of the methodological framework, the datasets, data pre-processing steps, clustering algorithms, and ensemble learning-based modeling

algorithms employed in this study are summarized in an integrated manner. Rather than aiming to provide an exhaustive state-of-the-art review, this section is intended to clearly present the analytical workflow and the methodological components adopted in the proposed framework. Accordingly, Table 2 summarizes the datasets used in the study together with the corresponding data pre-processing procedures, feature selection and standardization methods, clustering algorithms applied to define valuation zones, and ensemble learning algorithms used for mass appraisal modeling. This structured overview enhances the methodological transparency of the study and provides a concise and reproducible reference framework for future research and comparative analyses.

Table 2. Summary of datasets and methods used in the methodological framework.

Dataset	Data Pre-Processing	Clustering	Modeling
The Price Dataset, The Local Features Dataset, Urban Functions Datasets, Real Estate Market Activity Dataset, Air Quality and Meteorological Datasets, The Building Energy Statistics Dataset	**GIS-Based Dataset Preparing Methods** (XY Table To Point, Analysis, Euclidean Distance Analysis, Spatial Interpolation Analysis, Extract Multi Values to Points Analysis)	**Clustering Algorithms** (K-Means, K-Medians, SCMCA)	**Ensemble Learning Algorithms** (RF, GBM, XGBoost, LightGBM)
	Outlier Detection Methods (Boxplot, COA)	**Clustering Validity Indices** (Silhouette Index, Calinski–Harabasz Index, Davies–Bouldin Index)	**Hyperparameter Tuning** (Grid Search)
	Feature Selection Method (Pearson Correlation Analysis)		**Performance Metrics** (MAE, MAPE, RMSE, R^2)
	Feature Standardization Method (Max Normalization)		

3. Results

3.1. Data Pre-Processing

Within the scope of this study, there is insufficient market data on housing types in industrial, protected, military and rural residential areas in Pendik, Tuzla, Gebze and Çayırova districts; therefore, the neighborhoods in these areas were excluded from the study. It is important to note that the majority of the northern part of the study area is excluded in this step (see Figure 2). The study area for analysis consisted of a total of 90 neighborhoods: 31 from Pendik district, 13 from Tuzla district, 23 from Gebze district, 9 from Çayırova district and 14 from Darıca district.

Samples that differ significantly from the majority of samples in the dataset are referred to as outlier data. Outlier data are defined as sample points that deviate from the general distribution, where the values of any feature in the dataset are in harmony with a large part of the dataset. Detecting and removing outliers from the dataset, which is an important data pre-processing step, enables the development of more reliable ML models [76,82,83]. Although there are various methods for detecting spatial and non-spatial outliers, the graph-based Boxplot and GIS-based Clustering and Outlier Analysis (COA) are utilized in this study.

The dataset was analyzed separately according to room type categories using the Boxplot method. Analyses and visualization were carried out with the gglot2 library in the open-source R software version 4.5.2. Analyses were performed for price (₺), area (m^2) and price per m^2 (₺/m^2) features. While there were no outliers below the lower limit in

room types for the price (₺) feature, outliers were detected in all room type categories on the upper limits. A total of 12,067 outliers were detected. When the area (m^2) feature was examined, outliers were detected below the lower limit and on the upper limit in all room types except the 2 + 2 room type (In Türkiye, apartment layouts are described using the 'x + y' format, where 'x' represents the number of bedrooms and 'y' represents the number of living rooms. For instance, a '2 + 1' apartment has two bedrooms and one living room). A total of 12.012 outliers were detected. While there were no outliers below the outlier lower limit for room types for the price per m^2 (₺/m^2) feature, outliers were detected in all room type categories on the outlier data upper limits. A total of 4.462 outliers were detected (Figure 6).

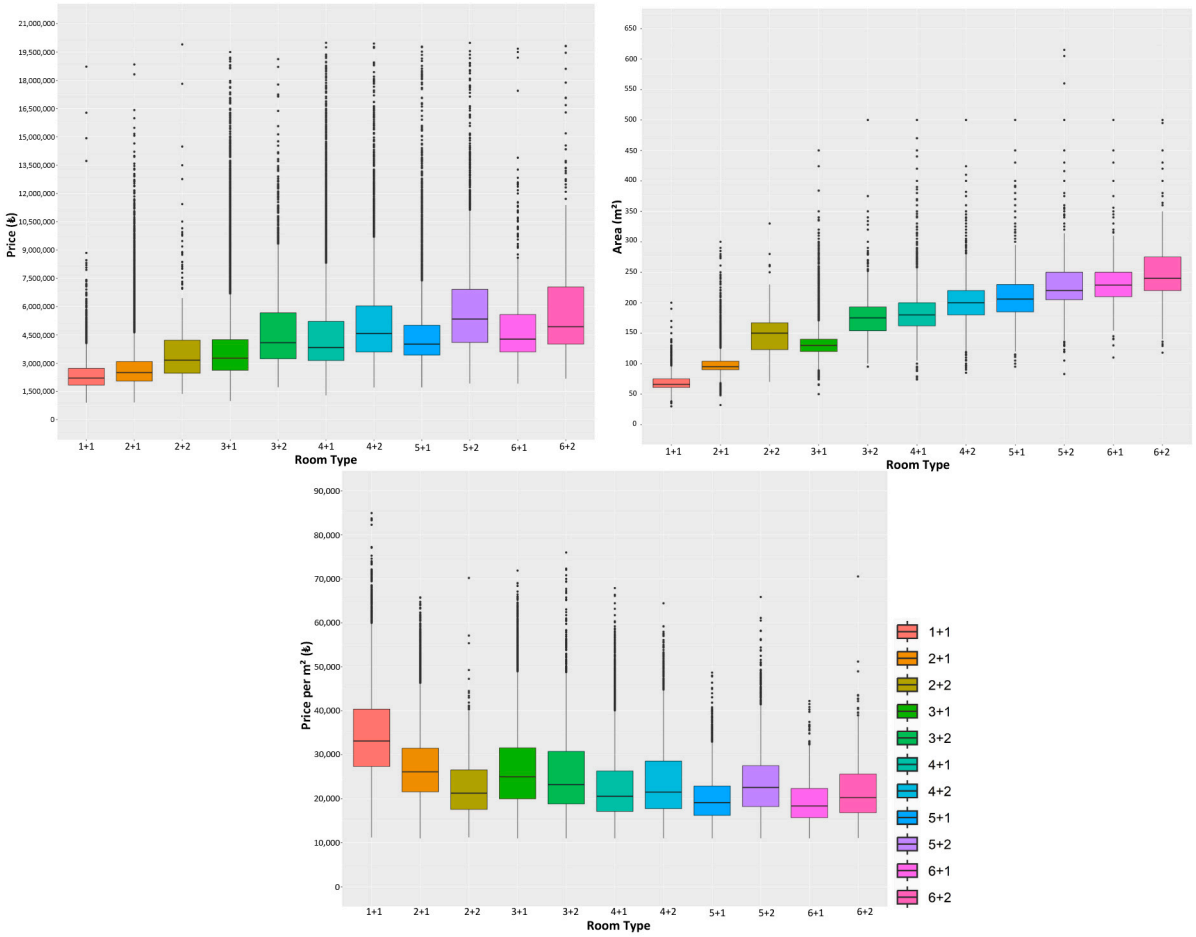

Figure 6. Detecting non-spatial outliers with Boxplot graphics for all room types.

Similar to non-spatial outliers, spatial outliers were detected with the COA method based on Anselin Local Moran's I statistic for each room type category. Analyses were performed in ArcGIS Pro 3.2 software for the unit price feature [84]. The COA method produces results expressed in four different categories. While the High-High (HH) category shows clusters with high values surrounded by samples with high values for the feature analyzed in the dataset, the Low-Low (LL) category shows clusters with low values surrounded by samples with low values for the feature analyzed. While the Low-High (LH) category shows samples with low values surrounded by samples with high values for the analyzed feature, High-Low (LH) shows samples with high values surrounded by samples with low values for the analyzed feature. Therefore, samples in these categories are considered spatial outliers [85]. In this study, outliers were detected in both categories for all room types. When the results were examined, spatial outliers were detected in the

following room types: 2.123 for 1 + 1 (i.e., one bedroom and one living room, respectively), 20.429 for 2 + 1, 46 for 2 + 2, 13.233 for 3 + 1, 168 for 3 + 2, 2.141 for 4 + 1, 638 for 4 + 2, 530 for 5 + 1, 196 for 5 + 2, 56 for 6 + 1, and 10 for 6 + 2. Thus, a total of 39.570 spatial outliers were detected. The analyses results for several different room types are given in Figure 7. Consequently, the detected all outliers were removed from the dataset and the mass appraisal dataset was optimized.

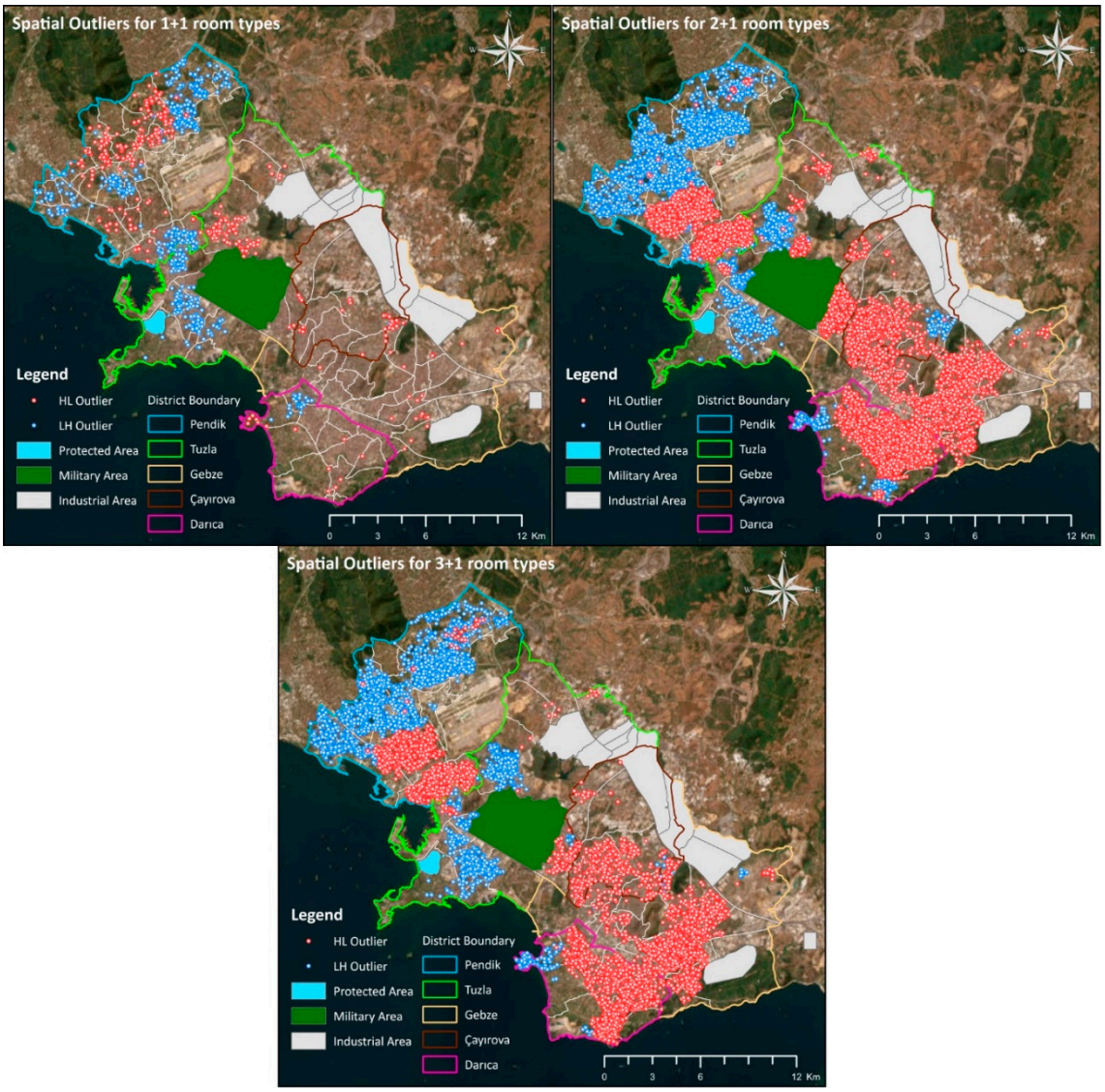

Figure 7. Detecting spatial outliers with COA for the room types.

3.2. Determining the Features Related to Real Estate Value for Clustering Analysis

To examine regional differences in mass appraisal, different spatial and non-spatial clustering algorithms are utilized to determine urban areas with similar characteristics. Determining the input features for clustering analyses is of great importance [86]. In this study, a sub-dataset containing neighborhood-level features in nine different categories such as population, marital status, income level, education level, household expenditure behavior, socio-economic development, real estate sales statistics, local inventory and land value was created to mass appraisal zones with clustering algorithms. Then, average housing unit prices were calculated for the neighborhoods in GIS. Pearson Correlation Analysis was performed to determine the input features correlated with the neighborhood housing unit prices for clustering analyses. According to the results given in Table 3,

11 features with correlation coefficients greater than 0.5 and statistically significant and average housing unit price were determined as input features for clustering analyses.

Table 3. Pearson Correlation Analysis results for clustering analysis input features.

Category	Feature	Correlation Coefficient	Sign (Two-Tailed)
Population	Total population	0.079	0.458
	Women residing in the neighborhood (%)	0.311 **	0.003
	Man residing in the neighborhood (%)	−0.311 **	0.003
	Population density (person/km^2)	−0.097	0.362
	Children population in the neighborhood (%)	−0.468 **	0.000
	Young population in the neighborhood (%)	−0.231 *	0.028
	Adult population in the neighborhood (%)	0.432 **	0.000
	Old population in the neighborhood (%)	0.406 **	0.000
Marital Status	Single person residing in the neighborhood (%)	0.023	0.830
	Married couple residing in the neighborhood (%)	−0.365	0.000
	Divorced couple residing in the neighborhood (%)	0.284 **	0.007
	Widows residing in the neighborhood (%)	0.617 **	0.000
Income Level	Population with A+ income level (%)	0.300 **	0.004
	Population with A income level (%)	0.298 **	0.004
	Population with B income level (%)	−0.026	0.809
	Population with C income level (%)	−0.277 **	0.008
	Population with D income level (%)	−0.298 **	0.004
Education Level	Population with unknown educational status (%)	0.344 **	0.001
	Illiterate population (%)	−0.362 **	0.000
	Literate population without a diploma (%)	−0.514 **	0.000
	Primary school graduate population (%)	−0.555 **	0.000
	Primary education graduate population (%)	−0.653 **	0.000
	Secondary school graduate population (%)	−0.731 **	0.000
	High school graduate population (%)	0.199	0.060
	Population with undergraduate degree (%)	0.740 **	0.000
	Population with MSc degree (%)	0.703 **	0.000
	Population with PhD degree (%)	0.666 **	0.000
Household Expenditure Behavior	Individual household income per capita (₺)	0.326 **	0.002
	Savings per capita (₺)	0.336 **	0.001
	Total household consumption expenditure per capita (₺)	0.322 **	0.002
	Food and non-alcoholic beverages expenditures per capita (₺)	0.263 *	0.012
	Alcoholic beverages, cigarette and tobacco expenditures per capita (₺)	0.279 **	0.008
	Clothing and footwear expenditures per capita (₺)	0.331 **	0.001
	Housing, water, electricity, gas and other fuel expenditures per capita (₺)	0.271 **	0.010
	Furniture, houses appliances and home care services expenditures per capita (₺)	0.331 **	0.001
	Health expenditures per capita (₺)	0.329 **	0.002
	Transportation expenditures per capita (₺)	0.335 **	0.001
	Communication expenditures per capita (₺)	0.319 **	0.002
	Entertainment and culture expenditures per capita (₺)	0.336 **	0.001
	Educational services expenditures per capita (₺)	0.330 **	0.002
	Restaurant, food services and hotel expenditures per capita (₺)	0.330 **	0.001
	Various good and services expenditures per capita (₺)	0.336 **	0.001

Table 3. *Cont.*

Category	Feature	Correlation Coefficient	Sign (Two-Tailed)
Socio-Economic Development	Socio-economic development score of the district	0.473 **	0.000
	General development ranking of the district by country	−0.370 **	0.000
	Development ranking of the district within the city	0.642 **	0.000
	District development level	−0.359 **	0.001
Sales Statistics	Real estate trading density	0.047	0.657
	Number of main real estate mortgage sales	−0.090	0.401
	Number of main real estate sales	−0.241 *	0.022
	Number of condominium mortgage sales	0.119	0.263
	Number of condominium unit sales	0.130	0.223
Local Inventory	Number of education facilities	0.313 **	0.003
	Number of religious facilities	0.053	0.622
	Number of cultural facilities	0.306 **	0.003
	Number of healthcare facilities	0.378 **	0.000
	Number of industrial facilities	0.046	0.668
	Number of private workplace facilities	0.231 *	0.029
	Number of shopping facilities	0.189	0.074
	Number of accommodation facilities	0.483 **	0.000
	Number of sport facilities	0.422 **	0.000
	Number of transportation facilities	0.543 **	0.000
	Number of entertainment facilities	0.025	0.813
	Number of public service facilities	0.296 **	0.005
	Number of personal care facilities	0.382 **	0.000
	Number of households	0.151	0.156
	Number of buildings	0.175	0.098
Land Value	Average land market value (₺)	0.666 **	0.666 **

Note: ** Correlation is significant at the 0.01 level (two-tailed); * Correlation is significant at the 0.05 level (two-tailed).

3.3. Creating Appraisal Zones with the Clustering Algorithms and Evaluating Clustering Results

K-Means, K-Medians and SCMCA clustering analyses were performed with the variables determined statistically significant for creating appraisal zones in the study area. In order to understand how the mass appraisal of the neighborhoods changes according to the district boundaries, five clusters were formed in all cluster analyses. Five appraisal zones with different characteristics were then identified for the study area using these algorithms, and the results were visualized using ArcGIS Pro 3.2, as can be seen in Figure 8.

Considering the results of K-Means, there are 12 neighborhoods in Cluster-1, 21 neighborhoods in Cluster-2, 13 neighborhoods in Cluster-3, 3 neighborhoods in Cluster-4 and a total of 41 neighborhoods in the cluster determined as Cluster-5. Considering the results of K-Medians, there are 15 neighborhoods in Cluster-1, 17 neighborhoods in Cluster-2, 16 neighborhoods in Cluster-3, 14 neighborhoods in Cluster-4 and a total of 28 neighborhoods in the cluster determined as Cluster-5. As can be seen in Figure 8, the K-Means and K-Medians algorithms generally exhibited a similar spatial pattern. However, changes were detected in some neighborhoods according to the district boundaries, and both algorithms exhibited a heterogeneous spatial pattern. For example, the İçmeler and Aydınlı neighborhoods in the Tuzla district and the Bayramoğlu neighborhood in the Darıca district were included in Cluster-1, which was created in the Pendik district. Considering the results of SCMCA, there are 22 neighborhoods in Cluster-1, 10 neighborhoods in Cluster-2, 10 neighborhoods in Cluster-3, 16 neighborhoods in Cluster-4 and a total of 32 neighborhoods in the cluster determined as Cluster-5. Compared to other algorithms,

it exhibited a more homogeneous spatial pattern that matched the district boundaries (Figure 8).

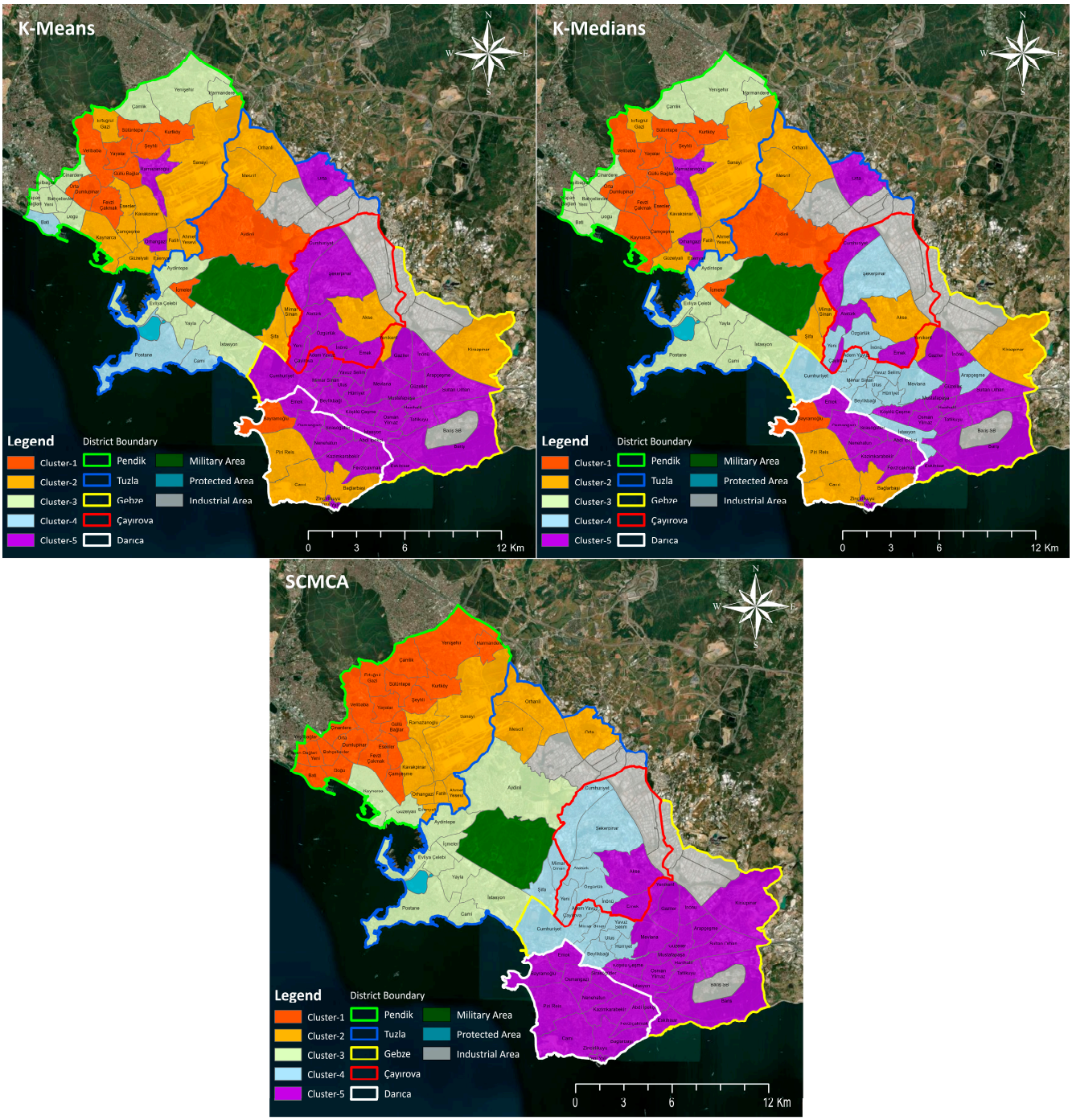

Figure 8. Clustering analysis results for the study area.

The results of the clustering analyses were evaluated separately using the indices expressed in Table 1, and the results are given in Table 4. When the results are examined, the K-Means algorithm demonstrates superior performance compared to the other algorithms. In terms of the Silhouette Index, K-Means achieved the highest value (0.549), indicating that data points are well assigned to their respective clusters and that the separation between clusters is clear. For the Calinski–Harabasz Index, K-Means also obtained the highest value (432.320), reflecting clusters that are both compact and well-separated. In the Davies–Bouldin Index evaluation, lower values indicate better performance, with K-

Means showing the best result (0.574). On the other hand, the SCMCA algorithm exhibited the weakest performance across all indices, while K-Medians demonstrated moderate performance. Overall, based on these three indices, the K-Means algorithm stands out as the method that best ensures both clear inter-cluster separation and strong intra-cluster cohesion for the given dataset.

Table 4. Comparison of clustering algorithm performance using common validity indices.

Clustering Algorithm	Silhouette Index	Calinski–Harabasz Index	Davies–Bouldin Index
K-Means	0.549	432.320	0.574
K-Medians	0.464	262.481	0.648
SCMCA	0.032	46.641	2.702

3.4. Developing Mass Appraisal Models with the Ensemble Learning Algorithms and Evaluating Results

In developing mass appraisal models, firstly, Pearson Correlation Analysis, was performed to avoid the multicollinearity problem in ML-based mass appraisal models and to select the features to be included in the models. Features were evaluated as pairwise. As a result of the evaluation, one of the features with a correlation coefficient greater than 0.5 was removed from the dataset. In this approach, which is widely used in modeling studies, the mass appraisal studies in the literature were considered in determining the threshold value as 0.5 [21,29]. Thus, a total of 120 features were made ready for developing mass appraisal models. Then, the optimized dataset was split into 70% training and 30% test ratios for the study area and the appraisal zones created by the clustering algorithms. The numbers of training and test samples corresponding to each cluster are summarized in Table 5. For the entire study area, a total of 128.797 samples were used for training, and 55.199 samples were used for testing.

Table 5. Number of training and test samples for clusters.

Clustering Algorithm	Training/Test Data Ratio	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Cluster-5
K-Means	Training (70%)	31.913	25.857	23.361	1.971	45.698
	Test (30%)	13.678	11.082	10.013	846	19.585
K-Medians	Training (70%)	36.743	20.193	25.332	13.736	32.797
	Test (30%)	15.748	8.654	10.857	5.888	14.056
SCMCA	Training (70%)	41.477	9.485	20.767	15.171	41.902
	Test (30%)	17.776	4.065	8.900	6.503	17.958

On the other hand, when developing mass appraisal models with ensemble learning algorithms, there are some hyperparameters for each algorithm. Determining optimal hyperparameter values is crucial for achieving high-performance models. In this study, the optimal values of the hyperparameters for the ensemble learning algorithms were determined with the Grid-Search approach with 10-fold cross-validation. These operations were performed on the dataset representing the entire study area using the "caret" package of the open-source R software version 4.5.2. The tuning ranges and optimal values applied for the hyperparameters are given in Table 6. The tuning ranges and optimal values applied for the hyperparameters are summarized in Table 6. Then, all mass appraisal models were trained using these optimal hyperparameter values for the clusters and the study area. While training the models, the real estate price per m^2 (₺/m^2) was defined as the dependent feature, and the features influencing the value were evaluated as independent features. The

models were trained using the relevant libraries of the R software, and a total of 64 mass appraisal models were developed.

Table 6. Hyperparameter tuning ranges and optimal parameter values for models.

Model	Parameter	Range	Optimal Value
RF	mtry	30, 40, 50	40
	ntree	50, 100, 150	150
GBM	n.trees	50, 100, 150	150
	interaction.depth	4, 5, 6	6
	shrinkage	0.01, 0.05, 0.10	0.05
	bag.fraction	0.6, 0.8, 1.0	0.8
	n.minobsinnode	10, 15, 20	15
XGBoost	eta	0.01, 0.05, 0.10	0.10
	max_depth	4, 5, 6	6
	min_child_weight	10, 15, 20	15
	Subsample	0.7, 0.8, 0.9	0.8
	colsample_bytree	0.7, 0.8, 0.9	0.9
	gamma	0, 1, 5	1
	alpha (L1)	0, 1, 5	0
	lambda (L2)	1, 5, 10	1
	Nrounds	50, 100, 150	150
LightGBM	learning_rate	0.01, 0.05, 0.10	0.05
	num_leaves	40, 50, 60	60
	max_depth	4, 5, 6	6
	min_child_samples	10, 15, 20	20
	subsample	0.7, 0.8, 0.9	0.8
	colsample_bytree	0.7, 0.8, 0.9	0.8
	reg_alpha (L1)	0, 0.1, 0.5	0.1
	reg_lambda (L2)	0, 0.1, 0.5	0.1
	Nrounds	50, 100, 150	150

The performance of all models was measured separately with the performance metrics described in Section 2.6. The goodness of fit of the ensemble learning-based mass appraisal models is evaluated holistically by jointly considering error-based performance metrics (MAE, MAPE, and RMSE) and explanatory power (R^2) on the test data, rather than relying on a single metric. The performance analysis results of models developed using different ensemble learning algorithms for the entire study area are summarized in Table 7. The results presented for all study area models were used as reference model performances in subsequent comparisons. The modeling results indicate that the LightGBM and XGBoost algorithms generally achieved superior performance. Notably, the LightGBM algorithm outperformed the other algorithms in terms of consistency and robustness, achieving lower error rates (MAE, MAPE, RMSE) and higher explanatory power (R^2). In this context, the ensemble learning algorithms can be ranked in terms of overall performance for mass appraisal as follows: LightGBM > XGBoost > GBM > RF. Furthermore, the observed performance ranking of the ensemble learning algorithms reflects their developmental

progression. Thus, the performance differences among the models can be explained not only by the characteristics of the datasets employed but also by the structural and historical evolution of the algorithms.

Table 7. Performance analysis results of the ensemble learning algorithm for the entire study area.

Model	MAE	MAPE	RMSE	R^2
RF	0.063	15.679	0.083	0.654
GBM	0.055	13.270	0.072	0.737
XGBoost	0.046	11.071	0.061	0.814
LightGBM	0.045	10.770	0.059	0.823

The results of the performance analysis of the models in the clusters, created using different clustering algorithms according to ensemble learning algorithms, are given in Tables 8–11.

Table 8. Performance analysis results for the RF algorithm.

Clustering Algorithm	Cluster	MAE	MAPE	RMSE	R^2
K-Means	Cluster-1	0.066	13.519	0.086	0.404
	Cluster-2	0.059	16.160	0.078	0.565
	Cluster-3	0.072	13.287	0.090	0.531
	Cluster-4	0.069	11.830	0.086	0.402
	Cluster-5	0.045	13.910	0.060	0.404
K-Medians	Cluster-1	0.064	14.303	0.083	0.504
	Cluster-2	0.057	15.434	0.076	0.515
	Cluster-3	0.072	13.124	0.090	0.518
	Cluster-4	0.054	13.110	0.069	0.389
	Cluster-5	0.047	13.889	0.063	0.428
SCMCA	Cluster-1	0.069	14.212	0.088	0.539
	Cluster-2	0.060	13.942	0.080	0.633
	Cluster-3	0.068	14.190	0.086	0.638
	Cluster-4	0.054	13.183	0.069	0.480
	Cluster-5	0.053	14.688	0.070	0.472

Table 9. Performance analysis results for the GBM algorithm.

Clustering Algorithm	Cluster	MAE	MAPE	RMSE	R^2
K-Means	Cluster-1	0.059	11.933	0.078	0.515
	Cluster-2	0.051	13.805	0.067	0.673
	Cluster-3	0.064	11.672	0.081	0.620
	Cluster-4	0.064	10.858	0.080	0.479
	Cluster-5	0.040	12.047	0.053	0.545
K-Medians	Cluster-1	0.057	12.475	0.075	0.603
	Cluster-2	0.051	13.509	0.067	0.614
	Cluster-3	0.065	11.622	0.081	0.605
	Cluster-4	0.049	11.654	0.063	0.495
	Cluster-5	0.042	12.120	0.056	0.555
SCMCA	Cluster-1	0.061	12.369	0.078	0.638
	Cluster-2	0.054	12.451	0.072	0.696
	Cluster-3	0.060	12.284	0.077	0.715
	Cluster-4	0.047	11.533	0.062	0.590
	Cluster-5	0.046	12.593	0.061	0.597

Table 10. Performance analysis results for the XGBoost algorithm.

Clustering Algorithm	Cluster	MAE	MAPE	RMSE	R^2
K-Means	Cluster-1	0.050	9.778	0.067	0.634
	Cluster-2	0.044	12.020	0.060	0.744
	Cluster-3	0.056	10.016	0.072	0.698
	Cluster-4	0.063	10.668	0.079	0.491
	Cluster-5	0.037	11.119	0.048	0.620
K-Medians	Cluster-1	0.049	10.474	0.065	0.697
	Cluster-2	0.045	11.807	0.060	0.694
	Cluster-3	0.056	10.019	0.073	0.683
	Cluster-4	0.047	11.153	0.061	0.530
	Cluster-5	0.038	11.116	0.051	0.631
SCMCA	Cluster-1	0.050	9.902	0.066	0.744
	Cluster-2	0.049	11.157	0.066	0.744
	Cluster-3	0.054	10.969	0.070	0.762
	Cluster-4	0.045	10.887	0.059	0.630
	Cluster-5	0.042	11.487	0.056	0.667

Table 11. Performance analysis results for the LightGBM algorithm.

Clustering Algorithm	Cluster	MAE	MAPE	RMSE	R^2
K-Means	Cluster-1	0.048	9.461	0.066	0.652
	Cluster-2	0.043	11.563	0.058	0.756
	Cluster-3	0.054	9.670	0.070	0.711
	Cluster-4	0.062	10.602	0.079	0.495
	Cluster-5	0.036	10.869	0.047	0.636
K-Medians	Cluster-1	0.046	9.950	0.063	0.717
	Cluster-2	0.043	11.364	0.058	0.712
	Cluster-3	0.055	9.745	0.072	0.695
	Cluster-4	0.046	10.970	0.060	0.541
	Cluster-5	0.037	10.820	0.050	0.646
SCMCA	Cluster-1	0.048	9.433	0.064	0.760
	Cluster-2	0.048	10.937	0.065	0.754
	Cluster-3	0.053	10.615	0.069	0.771
	Cluster-4	0.044	10.735	0.058	0.637
	Cluster-5	0.041	11.163	0.055	0.682

In the clusters defined with clustering algorithms, model performance metrics were calculated separately for each ensemble learning algorithm. However, to directly compare the impact of different clustering algorithms on modeling performance and to generalize the results across the entire study area, cluster-based weighted average values of each

performance metric were calculated with Equation (8), taking into account the test sample sizes of the respective clusters.

$$PM_{Weighted} = \frac{\sum_{i=1}^{k}(n_{test,i} \times PM_i)}{\sum_{i=1}^{k} n_{test,i}} \quad (8)$$

In the equation, $PM_{Weighted}$ represents the cluster-based weighted average value of the performance metric; k indicates the total number of clusters generated by the clustering algorithms; $n_{test,i}$ indicates the number of test samples used for the performance evaluation of the i^{th} cluster; and PM_i indicates the calculated value of the corresponding performance metric for the i^{th} cluster.

According to the analysis results presented in Table 12, the overall performance ranking of the ensemble learning algorithms used in the development of cluster-based models, similar to the models developed for the entire study area, can be expressed as LightGBM > XGBoost > GBM > RF. On the other hand, it is clearly observed that the models developed for the clusters generated by different clustering algorithms generally achieve higher performance levels compared to the reference models at the scale of the entire study area.

Table 12. Cluster-based weighted average values of performance analysis results according to clustering and ensemble learning algorithms.

Clustering Algorithm	Model	MAE	MAPE	RMSE	R^2
K-Means	RF	0.058	14.120	0.076	0.459
	GBM	0.052	12.285	0.067	0.576
	XGBoost	0.045	10.761	0.060	0.661
	LightGBM	0.044	10.438	0.059	0.675
K-Medians	RF	0.059	14.016	0.077	0.477
	GBM	0.053	12.291	0.069	0.581
	XGBoost	0.047	10.830	0.062	0.659
	LightGBM	0.045	10.461	0.060	0.675
SCMCA	RF	0.061	14.222	0.079	0.533
	GBM	0.054	12.336	0.070	0.636
	XGBoost	0.047	10.798	0.062	0.709
	LightGBM	0.046	10.450	0.061	0.722

While the best performance in terms of error metrics (MAE, MAPE and RMSE) was observed in the K-Means algorithms in the models developed with the RF algorithm, the explanatory power of the model remained relatively limited ($R^2 \approx 0.46$). In contrast, SCMCA significantly increased the explanatory power of the models ($R^2 = 0.533$) despite the limited increase in error metrics.

In the models developed using the GBM algorithm, overall higher performance was observed compared to the RF models, and the best performance in terms of error metrics was again obtained with the K-Means algorithm. While the K-Means algorithm highlighted model performance with its low error values, its explanatory power remained at a moderate level ($R^2 \approx 0.58$). As observed in the RF models, SCMCA also enhanced the spatial consistency of the models by increasing their explanatory power ($R^2 = 0.636$).

In the models developed using the XGBoost algorithm, overall high performance was observed, and the best performance in terms of error metrics was obtained with the K-Means clustering algorithm. While this algorithm highlighted regional model performance through its low error values, the explanatory power of the models remained at a high level ($R^2 \approx 0.66$). Similar to the other ensemble learning models, SCMCA was observed to enhance the spatial consistency of the models by increasing their explanatory power ($R^2 = 0.709$).

In the models developed using the LightGBM algorithm, the highest overall performance was observed compared to the other ensemble learning algorithms, and the best performance in terms of error metrics was achieved with the K-Means algorithm. While this algorithm highlighted regional model performance through its low error values, the explanatory power of the models was also at a high level ($R^2 = 0.675$). Similar to the other models, SCMCA was observed to strengthen spatial consistency by increasing the explanatory power of the models ($R^2 = 0.722$).

When the performance analysis results are evaluated holistically across the clustering algorithms, it is observed that the classical K-Means and K-Medians clustering algorithms exhibit high performance in minimizing error metrics, whereas the SCMCA algorithm enhances spatial explanatory power in the models where it is implemented. In particular, the results of the SCMCA algorithm indicate that jointly evaluating spatially proximate and similar areas leads to a more spatially consistent distribution of predicted values, especially in regions with high spatial heterogeneity.

To enable comparison of the cluster-based weighted average performance metrics ($PM_{Weighted}$) obtained from the clustering algorithms with the results of the reference models metrics ($PM_{Reference}$) generated for the entire study area (as presented in Table 7), the percentage changes (%ΔPM) in each performance metric were calculated using Equation (9), and the results are given in Table 13.

$$\%\Delta\text{PM} = \frac{PM_{Weighted} - PM_{Reference}}{PM_{Reference}} \quad (9)$$

Table 13. Percentage changes in cluster-based weighted average performance metrics compared to the models for the entire study area.

Clustering Algorithm	Model	MAE (%ΔPM)	MAPE (%ΔPM)	RMSE (%ΔPM)	R^2 (%ΔPM)
K-Means	RF	−7.26	−9.95	−8.40	−29.81
	GBM	−5.48	−7.42	−6.90	−21.83
	XGBoost	−0.84	−2.80	−1.15	−18.90
	LightGBM	−1.07	−3.09	−1.14	−17.90
K-Medians	RF	−5.73	−10.61	−7.26	−27.13
	GBM	−3.30	−7.38	−4.83	−21.11
	XGBoost	2.22	−2.18	1.74	−19.06
	LightGBM	1.57	−2.87	1.55	−17.96
SCMCA	RF	−3.05	−9.29	−4.16	−18.55
	GBM	−1.69	−7.04	−2.88	−13.71
	XGBoost	3.20	−2.46	2.95	−12.99
	LightGBM	2.74	−2.97	2.84	−12.27

When the results presented in Table 13 are examined, it is observed that the general trend for ensemble learning algorithms is an improvement in error metrics; however, decreases at varying levels are identified in the R^2 values that represent the explanatory power of the models. RF models achieved the highest improvement in error metrics but exhibited the greatest loss in explanatory power, with an approximately 25% decrease in the average R^2 values. Similarly, although GBM models provided significant improvements in error metrics, they experienced the second-highest decline in explanatory power, with an approximately 19% reduction in the average R^2 values. This indicates that, despite their strength in minimizing error metrics, RF and GBM models demonstrate limited performance in preserving spatial variation following regional modeling.

In the high-performing LightGBM and XGBoost models, improvements in error metrics remained relatively modest; however, the comparatively lower decreases in R^2 values ($R^2 \approx 16$–17) suggest that these models are able to enhance predictive accuracy after the clustering analyses while also maintaining spatial consistency in a more balanced manner.

In the cluster-based models developed using the RF algorithm, notable improvements were achieved in the error metrics, with the K-Means algorithm providing the highest performance gain. However, the approximately 30% decrease in the R^2 metric compared to the reference model indicates a limited reduction in explanatory power despite the improvements in error metrics. Nevertheless, SCMCA preserved spatial explanatory power more effectively by limiting the decrease in R^2 values to around 19%.

In the models developed using the GBM algorithm, the improvements in error metrics were more limited compared to the RF models, yet the overall predictive accuracy was maintained. The best results were again obtained with the K-Means algorithm. The decrease in R^2 values was approximately 22%, which is lower than that observed in the RF models. SCMCA also strengthened spatial consistency by limiting the reduction in R^2 to around 14%. Furthermore, the results indicate that GBM models exhibit a more balanced performance than RF models in reducing error metrics while preserving explanatory power.

In the models developed using the XGBoost algorithm, changes in error metrics were limited, and predictive accuracy was largely preserved. The highest performance improvement was observed with the K-Means algorithm. The decrease in the R^2 metric was approximately 19%, while in SCMCA, this reduction was limited to around 13%. Furthermore, the results clearly indicate that when XGBoost models are used in combination with SCMCA, they tend to preserve spatial explanatory power in a manner similar to other ensemble learning models.

In the models developed using the LightGBM algorithm, overall decreases or limited increases in error metrics were observed, while predictive accuracy was largely preserved. The K-Means algorithm again produced the best results in these models. Changes in R^2 values were approximately 18% compared to the reference models, whereas in SCMCA, this reduction was limited to around 12%. Therefore, when LightGBM models are used in combination with the SCMCA algorithm, they exhibit the most balanced and consistent performance in terms of preserving spatial explanatory power and balancing error metrics.

Although RF and GBM models demonstrated significant improvements, particularly in error metrics, when overall performance and ranking results are considered, the use of LightGBM and XGBoost algorithms in combination with both classical and spatially constrained clustering algorithms provides a more balanced and reliable modeling performance in mass appraisal, in terms of both predictive accuracy and spatial consistency.

4. Discussion and Conclusions

The rapid evolution of AI technologies has recently brought notable improvements to mass appraisal models, reflecting the transformative impact that AI is having across

many sectors. ML techniques within AI have provided the ability to process big and complex datasets used in mass appraisal more quickly and accurately. Mass appraisal models are developed by using ML algorithms and geo-datasets enriched with GIS analyses, enabling effective price predictions. Furthermore, clustering algorithms can be used to determine similar urban residential areas in terms of mass appraisal using various socio-economic development parameters such as income level, expenditure distribution, and education level. In this context, it is important to use different clustering algorithms to identify areas with similar characteristics in terms of mass appraisal and evaluate their effect on model performance.

This study investigates the impact of popular clustering algorithms (i.e., K-Means, K-Medians and SCMCA) on the performance of ensemble learning-based mass appraisal models. For this purpose, a case study was conducted in the study area, which consists of neighborhoods from five districts with different socio-economic characteristics, located within the provincial boundaries of Istanbul and Kocaeli, Türkiye. Firstly, variables reflecting the neighborhood-level socio-economic development characteristics, which are correlated with property values in the study area, were determined, and clustering analyses were conducted using these variables. The quality of the cluster analysis results was evaluated using established indices (i.e., Silhouette, Calinski–Harabasz, Davies–Bouldin). The results show that, based on the indices, the K-Means algorithm outperformed the other clustering methods by providing the clearest separation between clusters and the strongest cohesion within clusters, while SCMCA showed the lowest performance and K-Medians achieved moderate results. Subsequently, mass appraisal models were trained using ensemble learning algorithms (i.e., RF, GBM, XGBoost and LightGBM) for the five clusters determined by each clustering algorithm, as well as for the entire study area. The performance of the models was comparatively evaluated on the test dataset using commonly used performance metrics (MAE, MAPE, RMSE and R^2). In terms of developing mass appraisal models using ensemble learning algorithms, the results indicate that LightGBM and XGBoost generally performed best, with LightGBM showing the highest consistency and robustness, followed by XGBoost, GBM, and RF in overall performance. The performance ranking of cluster-based models is similar to the ranking of models developed for the entire study area (LightGBM > XGBoost > GBM > RF); however, models developed on clusters formed with different clustering algorithms generally outperform the entire study area models. In terms of clustering algorithms, it was observed that K-Means and K-Medians performed best in minimizing errors, while SCMCA improved spatial consistency by effectively grouping spatially proximate and similar areas, particularly in heterogeneous regions.

This study offers significant implications for the United Nations Sustainable Development Goals (SDGs), particularly SDG 11 (Sustainable Cities and Communities). The proposed ensemble learning-based mass appraisal framework, supported by different clustering algorithms, provides local authorities and urban planners with a data-driven decision-support tool that enables more accurate, transparent, and consistent real estate valuation processes. Such evidence-based valuation systems strengthen institutional capacity for effective urban planning and land management through supporting sustainable urban development strategies.

The effectiveness of clustering algorithms is closely related to the spatial scale of the study area, the degree of heterogeneity in the dataset, the scope of the explanatory variables and the specific modeling objectives. While this study demonstrated the significant benefits of integrating clustering algorithms into ensemble learning-based mass appraisal models, there are several limitations and areas for future research to enhance the generalizability, validation, algorithmic expansion and robustness of the proposed framework.

The methodology used in this study should be tested in diverse geographical regions with varying levels of property heterogeneity and market dynamics to validate the generalizability of our findings. The geographical scope of the study is confined to five districts within the Istanbul–Kocaeli region. This limitation is a crucial factor that may restrict the direct generalizability of the results to other cities or countries with different market conditions, socio-economic structures, and real estate price dynamics, although the selected study area is considered as an area representative of the wider region with the same characteristics such as income levels, mixed urban forms, varied, property types and heterogeneous real estate market. Therefore, it is determined that further investigation is necessary in order to ascertain the generalizability of the proposed model in other areas in the future.

A range of dynamic factors, including macroeconomic, infrastructural and policy dynamics, such as interest rates, national housing programs, zoning regulations and large-scale infrastructure investments, exhibit temporal variation and heterogeneous regional effects. The analysis of these factors typically necessitates the use of temporal or panel data-based modeling approaches. In future research, the integration of such variables into the proposed clustering-based modeling framework would facilitate its extension to different time periods and spatial contexts.

The spatial analysis (e.g., distance) is conducted solely within the designated study area. In more detail, the distances are calculated exclusively for the POI within the study area, as no POI information exists outside of this area in the dataset. A future study, in the case that a more comprehensive dataset is obtained, may extend the analysis to encompass the eastern and western regions adjacent to the study area, thus facilitating a more realistic evaluation of the results. Furthermore, calculating accessibility or transportation network-based measures instead of Euclidean distance analysis may provide a more realistic reflection of the effects of physical barriers and transportation infrastructure. Therefore, the integration of network-based accessibility measures and movement potential-based variables into the proposed modeling framework is planned as a future work. Apart from this, house price time-adjustments rely on city-level official indices, which may be seen as an issue given that the study covers five districts across two cities. Calculating and using finer-grained residential price indices could improve adjustment precision and represents a potential direction for future work.

The role of institutional ownership structures and developer-level characteristics in revealing additional heterogeneity in price formation, particularly in metropolitan markets with large-scale or branded residential projects, was not examined in this study. This can be seen as a limitation of the current work and a relevant direction for future research.

In this study, the Silhouette, Calinski–Harabasz and Davies–Bouldin indices are used to comparatively evaluate the outcomes of different clustering algorithms, with the analytical focus placed on relative performance rather than absolute statistical inference. Assessing the stability and uncertainty of clustering results through resampling-based procedures (e.g., bootstrap) or repeated runs with different initialization settings (e.g., varying random seeds) is methodologically important and this can be considered as a limitation of the current study and a relevant direction for future work. Moreover, different indices, such as the Dunn Index, Ball–Hall Index, C Index, Gamma Index, Ratkowsky–Lance Index, Wemmert–Gancarski Index, Kendall's Tau Index, and spatial autocorrelation-based measures (e.g., Moran's I), can be also employed in subsequent studies in order to evaluate the validity of clusters.

Another future work is determined as integrating the proposed clustering-based modeling framework with explainable artificial intelligence (XAI) techniques (e.g., SHAP),

which would enhance the applicability of the approach, particularly in decision-support contexts such as tax policy and spatial planning.

Last but not least, the scope of the investigation could be broadened by incorporating other spatially constrained clustering methods (e.g., Regionalization with Dynamically Constrained Agglomerative Clustering and Partitioning—REDCAP and Spatially Constrained Hierarchical Clustering—SCHC) alongside alternative clustering techniques (e.g., spectral clustering or self-organizing maps) and a broader set of ensemble learning algorithms (e.g., CatBoost or Stacking Ensembles) to identify the most effective combination. This may enable the development of a more robust, adaptive and transferable methodology for implementing cluster-based mass appraisal systems.

Author Contributions: Conceptualization, A.C.A.; Methodology, S.S. and A.K.; Software, S.S.; Investigation, S.S. and A.K.; Writing—original draft, S.S., A.K. and A.C.A.; Writing—review and editing, S.S., A.K. and A.C.A.; Visualization, S.S.; Supervision, A.C.A.; Project administration, A.C.A.; Funding acquisition, A.C.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Scientific and Technological Research Council of Turkey, grant number 122R021.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. The detailed content of the datasets used in this study.

Dataset	Variables	Data Level
Price Dataset	Sales Price (₺), Area (m^2), Number of Rooms, Number of Living Rooms, Total Number of Rooms, Number of Bathrooms, Total Number of Floors, Floor Level, Building Age, Direction (North, South, East, West), Room Type, Terrace Area (m^2), Inside a Residential Complex, Sports Facility/Gym, Children's Playground, Elevator, Generator, Security Service, Open Parking Lot, Indoor Parking Garage, Outdoor Swimming Pool, Indoor Swimming Pool, Thermal Insulation, Air Conditioning, Fireplace, Heating System, Landscape (City, Nature, Sea, etc.), Presence of Terrace, Being Duplex, Eligibility for Bank Loan, Furnished Status, etc.	Location (Latitude and Longitude)
Local Features Dataset	Population (Neighborhood Population, Neighborhood Population Density (People/km^2), Percentage of Children (0–14) (%), Percentage of Youth (15–24) (%), Percentage of Adults (25–65) (%), Percentage of Elderly (65+) (%), Percentage of Females (%), Percentage of Males (%)), Marital Status (Percentage of Married Couples (%), Percentage of Single Individuals (%), Percentage of Divorced Couples (%), Percentage of Widowed Individuals (%)), Income Level (Percentage of A+ Group Individuals (%), Percentage of A Group Individuals (%), Percentage of B Group Individuals (%), Percentage of C Group Individuals (%), Percentage of D Group Individuals (%), Household Income Per Capita, Savings Per Capita), Education Level (Percentage of Individuals with Unknown Education Level (%), Percentage of Illiterate Individuals (%), Percentage of Literate but Uneducated Individuals (%), Percentage of Primary School Graduates (%), Percentage of Secondary School Graduates (%), Percentage of High School Graduates (%), Percentage of Undergraduate/Bachelor's Graduates (%), Percentage of Master's Graduates (%), Percentage of PhD Graduates (%)), Household Expenditure Behaviors (Food Expenditures, Healthcare Expenditures, Transportation Expenditures, Education Expenditures, Housing Expenditures, Clothing Expenditures, Restaurant Expenditures, Entertainment Expenditures, Alcohol Expenditures, Furniture Expenditures, Communication Expenditures, Total Expenditure), Socio-Economic Development (Provincial Socio-Economic Development Level, District Socio-Economic Development Level, Socio-Economic Development Ranking of the District within the Province)	Neighborhood/District/ Province

Table A1. *Cont.*

Dataset	Variables	Data Level
Urban Functions Datasets	**Educational Facilities** (Kindergartens, Primary Schools, High Schools, Universities, Public Education Centers), **Health Facilities** (Local Health Units, Hospitals, Pharmacies, Emergency Health Stations, Clinics, etc.), **Shopping and Commercial Facilities** (City Centers/Bazaars, Shopping Malls, Markets, Restaurants, Post Offices), **Cultural Facilities** (Exhibition Centers, Cinemas and Theaters, Museums, Convention and Cultural Centers, Libraries, Historical Buildings), **Public Service Facilities** (Administrative Facilities, Courthouses, Banks, ATMs, Fire Stations, Security Units), Green Spaces Parks, Forests), **Entertainment and Sports Facilities** (Beaches, Amusement Parks, Sports Facilities), **Accommodation Facilities** (Guesthouses, Hotels), **Industrial Facilities** (Fuel Stations, Industrial Facilities, Water Treatment Plants), **Religious Facilities** (Places of Worship, Cemeteries), **Transportation Facilities** (Rail System Stations, Airports, Bus Stops, Taxi Stands, Sea Piers, EV Charging Stations, Bike Stations, Bus Terminals, etc.), **Public Law Restriction Areas** (Military Area, Natural Protected Area, Specially Protected Environment Areas, Water Basin Areas, etc.)	Location (Latitude and Longitude)
Real Estate Market Activity Dataset	Trading Density, Number of Sales, Number of Mortgages, Sales for Condominium Units	Neighborhood
Air Quality and Meteorological Dataset	Air Quality Index, Temperature (°C), Wind Speed (m/sn), Pressure (hPa), Humidity (%)	Location (Latitude and Longitude)
The Building Energy Statistics Dataset	Number of Building Energy Certificates, Primary Energy Consumption (kWh/year), Air-Conditioned area (m^2), Renewable Energy Contribution (kWh/year)	District

References

1. Bovkir, R.; Aydinoglu, A.C. Providing Land Value Information from Geographic Data Infrastructure by Using Fuzzy Logic Analysis Approach. *Land Use Policy* **2018**, *78*, 46–60. [CrossRef]
2. IVSC. *International Valuation Standards*; IVSC: London, UK, 2022; ISBN 9780993151347.
3. Kang, Y.; Zhang, F.; Peng, W.; Gao, S.; Rao, J.; Duarte, F.; Ratti, C. Understanding House Price Appreciation Using Multi-Source Big Geo-Data and Machine Learning. *Land Use Policy* **2021**, *111*, 104919. [CrossRef]
4. Robin, E. Performing Real Estate Value(s): Real Estate Developers, Systems of Expertise and the Production of Space. *Geoforum* **2022**, *134*, 205–215. [CrossRef]
5. Dimopoulos, T.; Moulas, A. A Proposal of a Mass Appraisal System in Greece with CAMA System: Evaluating GWR and MRA Techniques in Thessaloniki Municipality. *Open Geosci.* **2016**, *8*, 675–693. [CrossRef]
6. IAAO. *Standard on Mass Appraisal of Real Property—A Criterion for Measuring Fairness, Quality, Equity and Accuracy*; IAAO: Kansas City, MO, USA, 2017.
7. Wang, D.; Li, V.J. Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review. *Sustainability* **2019**, *11*, 7006. [CrossRef]
8. Tepe, E. A Random Forests-Based Hedonic Price Model Accounting for Spatial Autocorrelation. *J. Geogr. Syst.* **2024**, *26*, 511–540. [CrossRef]
9. Aydinoglu, A.C.; Bovkir, R.; Colkesen, I. Implementing a Mass Valuation Application on Interoperable Land Valuation Data Model Designed as an Extension of the National GDI. *Surv. Rev.* **2021**, *53*, 349–365. [CrossRef]
10. Jafary, P.; Shojaei, D.; Rajabifard, A.; Ngo, T. Automating Property Valuation at the Macro Scale of Suburban Level: A Multi-Step Method Based on Spatial Imputation Techniques, Machine Learning and Deep Learning. *Habitat Int.* **2024**, *148*, 103075. [CrossRef]
11. Jafary, P.; Shojaei, D.; Rajabifard, A.; Ngo, T. Automated Land Valuation Models: A Comparative Study of Four Machine Learning and Deep Learning Methods Based on a Comprehensive Range of Influential Factors. *Cities* **2024**, *151*, 105115. [CrossRef]
12. Iban, M.C. An Explainable Model for the Mass Appraisal of Residences: The Application of Tree-Based Machine Learning Algorithms and Interpretation of Value Determinants. *Habitat Int.* **2022**, *128*, 102660. [CrossRef]
13. Carranza, J.P.; Piumetto, M.A.; Lucca, C.M.; Da Silva, E. Mass Appraisal as Affordable Public Policy: Open Data and Machine Learning for Mapping Urban Land Values. *Land Use Policy* **2022**, *119*, 106211. [CrossRef]
14. Deppner, J.; von Ahlefeldt-Dehn, B.; Beracha, E.; Schaefers, W. Boosting the Accuracy of Commercial Real Estate Appraisals: An Interpretable Machine Learning Approach. *J. Real Estate Financ. Econ.* **2025**, *71*, 314–351. [CrossRef] [PubMed]
15. Unel, F.B.; Yalpir, S. Sustainable Tax System Design for Use of Mass Real Estate Appraisal in Land Management. *Land Use Policy* **2023**, *131*, 106734. [CrossRef]
16. Baur, K.; Rosenfelder, M.; Lutz, B. Automated Real Estate Valuation with Machine Learning Models Using Property Descriptions. *Expert Syst. Appl.* **2023**, *213*, 119147. [CrossRef]

17. Cellmer, R.; Cichulska, A.; Bełej, M. Spatial Analysis of Housing Prices and Market Activity with the Geographically Weighted Regression. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 380. [CrossRef]
18. Hosseini, S.M.; Bahadori, B.; Charkhan, S. Spatial Analysis of Housing Prices in Tehran City. *Int. J. Hous. Mark. Anal.* **2024**, *17*, 475–497. [CrossRef]
19. Mete, M.O. Developing GeoAI Integrated Mass Valuation Model Based on LADM Valuation Information Great Britain Country Profile. *Trans. GIS* **2025**, *29*, e13273. [CrossRef]
20. Genc, N.; Colak, H.E.; Ozbilgin, F. Spatial Performance Approach to Machine Learning Algorithms: A GIS-Based Comparison Analysis for Real Estate Valuation. *Trans. GIS* **2025**, *29*, e13303. [CrossRef]
21. Aydinoglu, A.C.; Sisman, S. Comparing Modelling Performance and Evaluating Differences of Feature Importance on Defined Geographical Appraisal Zones for Mass Real Estate Appraisal. *Spat. Econ. Anal.* **2024**, *19*, 225–249. [CrossRef]
22. Soltani, A.; Pettit, C.J.; Heydari, M.; Aghaei, F. Housing Price Variations Using Spatio-Temporal Data Mining Techniques. *J. Hous. Built Environ.* **2021**, *36*, 1199–1227. [CrossRef]
23. Wu, Y.; Wei, Y.D.; Li, H. Analyzing Spatial Heterogeneity of Housing Prices Using Large Datasets. *Appl. Spat. Anal. Policy* **2020**, *13*, 223–256. [CrossRef]
24. DGM. District Areas, Directorate General for Mapping (DGM)—National Mapping Agency. Available online: https://www.harita.gov.tr/il-ve-ilce-yuzolcumleri (accessed on 17 November 2025).
25. TurkStat. Population and Demography. Available online: https://data.tuik.gov.tr/Kategori/GetKategori?p=nufus-ve-demografi-109&dil=1 (accessed on 17 November 2025).
26. GDDA. *General Directorate of Development Agencies (GDDA), Socio-Economic Development Index of Districts (SEDI)-2022*; GDDA: Ankara, Turkey, 2023.
27. Endeksa. Value, Sell, Invest in Property—Endeksa. Available online: https://www.endeksa.com/en/ (accessed on 17 November 2025).
28. Çağdaş, V. An Application Domain Extension to CityGML for Immovable Property Taxation: A Turkish Case Study. *Int. J. Appl. Earth Obs. Geoinf.* **2013**, *21*, 545–555. [CrossRef]
29. Yilmazer, S.; Kocaman, S. A Mass Appraisal Assessment Study Using Machine Learning Based on Multiple Regression and Random Forest. *Land Use Policy* **2020**, *99*, 104889. [CrossRef]
30. EVDS. Residential Property Price Index (RPPI) Statistics-Electronic Data Delivery System (EVDS). Available online: https://www.tcmb.gov.tr/ (accessed on 17 November 2025).
31. GDLRC. Land Query Application, General Directorate of Land Registry and Cadastre (GDLRC). Available online: https://parselsorgu.tkgm.gov.tr/ (accessed on 17 November 2025).
32. NAQIMN. National Air Quality Monitoring Network (NAQIMN), Ministry of Environment, Urbanisation and Climate Change. Available online: https://www.turkiye.gov.tr/cevre-ve-sehircilik-ulusal-hava-kalite-izleme-agi (accessed on 17 November 2025).
33. TSMS. Meteorological Data-Information Presentation and Sales System, Turkish State Meteorological Service (TSMS). Available online: https://www.mgm.gov.tr/eng/forecast-cities.aspx (accessed on 17 November 2025).
34. GVS. Directorate General of Vocational Services (GVS). Available online: https://meslekihizmetler.csb.gov.tr/en (accessed on 17 November 2025).
35. Velumani, P.; Priyadharshini, B.; Mukilan, K.; Shanmugapriya. A Mass Appraisal Assessment Study of Land Values Using Spatial Analysis and Multiple Regression Analysis Model (MRA). *Mater. Today Proc.* **2022**, *66*, 2614–2625. [CrossRef]
36. Zhang, R.; Du, Q.; Geng, J.; Liu, B.; Huang, Y. An Improved Spatial Error Model for the Mass Appraisal of Commercial Real Estate Based on Spatial Analysis: Shenzhen as a Case Study. *Habitat Int.* **2015**, *46*, 196–205. [CrossRef]
37. Ezugwu, A.E.; Ikotun, A.M.; Oyelade, O.O.; Abualigah, L.; Agushaka, J.O.; Eke, C.I.; Akinyelu, A.A. A Comprehensive Survey of Clustering Algorithms: State-of-the-Art Machine Learning Applications, Taxonomy, Challenges, and Future Research Prospects. *Eng. Appl. Artif. Intell.* **2022**, *110*, 104743. [CrossRef]
38. Montero, G.; Caruso, G.; Hilal, M.; Thomas, I. A Partition-Free Spatial Clustering That Preserves Topology: Application to Built-up Density. *J. Geogr. Syst.* **2023**, *25*, 5–35. [CrossRef]
39. Shutaywi, M.; Kachouie, N.N. Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy* **2021**, *23*, 759. [CrossRef]
40. Demirhan, H.; Baser, F. Hierarchical Fuzzy Regression Functions for Mixed Predictors and an Application to Real Estate Price Prediction. *Neural Comput. Appl.* **2024**, *36*, 11545–11561. [CrossRef]
41. Chen, M.; Arribas-Bel, D.; Singleton, A. Understanding the Dynamics of Urban Areas of Interest through Volunteered Geographic Information. *J. Geogr. Syst.* **2019**, *21*, 89–109. [CrossRef]
42. MacQueen, J. Multivariate Observations. In Proceedings of the 5th Berkeley Symposium on Mathematical Statisticsand Probability, Berkley, CA, USA, 18–21 June 1965 and 27 December 1965–7 January 1966; pp. 281–297.
43. Ikotun, A.M.; Ezugwu, A.E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-Means Clustering Algorithms: A Comprehensive Review, Variants Analysis, and Advances in the Era of Big Data. *Inf. Sci.* **2023**, *622*, 178–210. [CrossRef]

44. Ye, S.; Song, C.; Shen, S.; Gao, P.; Cheng, C.; Cheng, F.; Wan, C.; Zhu, D. Spatial Pattern of Arable Land-Use Intensity in China. *Land Use Policy* **2020**, *99*, 104845. [CrossRef]
45. Kaufman, L.; Rousseeuw, P. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2009.
46. Godichon-Baggioni, A.; Surendran, S. A Penalized Criterion for Selecting the Number of Clusters for K-Medians. *J. Comput. Graph. Stat.* **2024**, *33*, 1298–1309. [CrossRef]
47. Assunção, R.M.; Neves, M.C.; Câmara, G.; Da Costa Freitas, C. Efficient Regionalization Techniques for Socio-Economic Geographical Units Using Minimum Spanning Trees. *Int. J. Geogr. Inf. Sci.* **2006**, *20*, 797–811. [CrossRef]
48. Anselin, L.; Amaral, P. Endogenous Spatial Regimes. *J. Geogr. Syst.* **2024**, *26*, 209–234. [CrossRef]
49. Ma, Y.; Lin, H.; Wang, Y.; Huang, H.; He, X. A Multi-Stage Hierarchical Clustering Algorithm Based on Centroid of Tree and Cut Edge Constraint. *Inf. Sci.* **2021**, *557*, 194–219. [CrossRef]
50. Rousseeuw, P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
51. Ertunç, E.; Karkınlı, A.E.; Bozdağ, A. A Clustering-Based Approach to Land Valuation in Land Consolidation Projects. *Land Use Policy* **2021**, *111*, 105739. [CrossRef]
52. Caliński, T.; Harabasz, J. A Dendrite Method Foe Cluster Analysis. *Commun. Stat.* **1974**, *3*, 1–27. [CrossRef]
53. Ros, F.; Riad, R.; Guillaume, S. PDBI: A Partitioning Davies-Bouldin Index for Clustering Evaluation. *Neurocomputing* **2023**, *528*, 178–199. [CrossRef]
54. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [CrossRef]
55. Deng, L.; Zhang, X. Boosting the Accuracy of Property Valuation with Ensemble Learning and Explainable Artificial Intelligence: The Case of Hong Kong. *Ann. Reg. Sci.* **2025**, *74*, 32. [CrossRef]
56. Mora-Garcia, R.T.; Cespedes-Lopez, M.F.; Perez-Sanchez, V.R. Housing Price Prediction Using Machine Learning Algorithms in COVID-19 Times. *Land* **2022**, *11*, 2100. [CrossRef]
57. Sagi, O.; Rokach, L. Ensemble Learning: A Survey. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2018**, *8*, e1249. [CrossRef]
58. Soltani, A.; Heydari, M.; Aghaei, F.; Pettit, C.J. Housing Price Prediction Incorporating Spatio-Temporal Dependency into Machine Learning Algorithms. *Cities* **2022**, *131*, 103941. [CrossRef]
59. Chen, Y.; Liu, X.; Li, X.; Liu, Y.; Xu, X. Mapping the Fine-Scale Spatial Pattern of Housing Rent in the Metropolitan Area by Using Online Rental Listings and Ensemble Learning. *Appl. Geogr.* **2016**, *75*, 200–212. [CrossRef]
60. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
61. Biau, G.; Scornet, E. A Random Forest Guided Tour. *TEST* **2016**, *25*, 197–227. [CrossRef]
62. Talukdar, S.; Eibek, K.U.; Akhter, S.; Ziaul, S.; Towfiqul Islam, A.R.M.; Mallick, J. Modeling Fragmentation Probability of Land-Use and Land-Cover Using the Bagging, Random Forest and Random Subspace in the Teesta River Basin, Bangladesh. *Ecol. Indic.* **2021**, *126*, 107612. [CrossRef]
63. Čeh, M.; Kilibarda, M.; Lisec, A.; Bajat, B. Estimating the Performance of Random Forest versus Multiple Regression for Predicting Prices of the Apartments. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 168. [CrossRef]
64. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. Available online: https://www.jstor.org/stable/2699986 (accessed on 17 November 2025). [CrossRef]
65. Antolí-Martínez, J.M.; Cortijo, A.; Martines, V.; Andrés, S.M.; Sánchez-López, E.; Sirignano, F.M.; Padrón, A.L.; Kim, C.; Park, T. Predicting Determinants of Lifelong Learning Intention Using Gradient Boosting Machine (GBM) with Grid Search. *Sustainability* **2022**, *14*, 5256. [CrossRef]
66. Konstantinov, A.V.; Utkin, L.V. Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines. *Knowledge-Based Syst.* **2021**, *222*, 106993. [CrossRef]
67. Touzani, S.; Granderson, J.; Fernandes, S. Gradient Boosting Machine for Modeling the Energy Consumption of Commercial Buildings. *Energy Build.* **2018**, *158*, 1533–1543. [CrossRef]
68. 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, 13–17 August 2016; pp. 785–794. [CrossRef]
69. Abedi, R.; Costache, R.; Shafizadeh-Moghadam, H.; Pham, Q.B. Flash-Flood Susceptibility Mapping Based on XGBoost, Random Forest and Boosted Regression Trees. *Geocarto Int.* **2022**, *37*, 5479–5496. [CrossRef]
70. Dong, Y.; Qiu, L.; Lu, C.; Song, L.; Ding, Z.; Yu, Y.; Chen, G. A Data-Driven Model for Predicting Initial Productivity of Offshore Directional Well Based on the Physical Constrained EXtreme Gradient Boosting (XGBoost) Trees. *J. Pet. Sci. Eng.* **2022**, *211*, 110176. [CrossRef]
71. Kilic, B.; Bayrak, O.C.; Gülgen, F.; Gurturk, M.; Abay, P. Unveiling the Impact of Machine Learning Algorithms on the Quality of Online Geocoding Services: A Case Study Using COVID-19 Data. *J. Geogr. Syst.* **2024**, *26*, 601–622. [CrossRef]

72. Microsoft Corporation. Welcome to LightGBM's Documentation!—LightGBM 4.6.0 Documentation. Available online: https://lightgbm.readthedocs.io/en/stable/ (accessed on 27 October 2025).
73. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 3149–3157.
74. Ahmed Soomro, A.; Akmar Mokhtar, A.; B Hussin, H.; Lashari, N.; Lekan Oladosu, T.; Muslim Jameel, S.; Inayat, M. Analysis of Machine Learning Models and Data Sources to Forecast Burst Pressure of Petroleum Corroded Pipelines: A Comprehensive Review. *Eng. Fail. Anal.* **2024**, *155*, 107747. [CrossRef]
75. Tian, L.; Feng, L.; Yang, L.; Guo, Y. Stock Price Prediction Based on LSTM and LightGBM Hybrid Model. *J. Supercomput.* **2022**, *78*, 11768–11793. [CrossRef]
76. Zaki, J.; Nayyar, A.; Dalal, S.; Ali, Z.H. House Price Prediction Using Hedonic Pricing Model and Machine Learning Techniques. *Concurr. Comput. Pract. Exp.* **2022**, *34*, e7342. [CrossRef]
77. Sibindi, R.; Mwangi, R.W.; Waititu, A.G. A Boosting Ensemble Learning Based Hybrid Light Gradient Boosting Machine and Extreme Gradient Boosting Model for Predicting House Prices. *Eng. Rep.* **2023**, *5*, e12599. [CrossRef]
78. Ho, W.K.O.; Tang, B.S.; Wong, S.W. Predicting Property Prices with Machine Learning Algorithms. *J. Prop. Res.* **2021**, *38*, 48–70. [CrossRef]
79. Hong, J.; Choi, H.; Kim, W.S. A House Price Valuation Based on the Random Forest Approach: The Mass Appraisal of Residential Property in South Korea. *Int. J. Strateg. Prop. Manag.* **2020**, *24*, 140–152. [CrossRef]
80. Yalpır, Ş. Enhancement of Parcel Valuation with Adaptive Artificial Neural Network Modeling. *Artif. Intell. Rev.* **2018**, *49*, 393–405. [CrossRef]
81. Wan, X.; Yang, W. A Divide-and-Conquer Method for Predicting the Fine-Grained Spatial Distribution of Population in Urban and Rural Areas. *J. Geogr. Syst.* **2025**, *27*, 283–299. [CrossRef]
82. Alghushairy, O.; Alsini, R.; Soule, T.; Ma, X. A Review of Local Outlier Factor Algorithms for Outlier Detection in Big Data Streams. *Big Data Cogn. Comput.* **2021**, *5*, 1. [CrossRef]
83. Ur Rehman, A.; Belhaouari, S.B. Unsupervised Outlier Detection in Multidimensional Data. *J. Big Data* **2021**, *8*, 80. [CrossRef]
84. ESRI. Cluster and Outlier Analysis (Anselin Local Moran's I) (Spatial Statistics)—ArcGIS Pro | Documentation. Available online: https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-statistics/cluster-and-outlier-analysis-anselin-local-moran-s.htm (accessed on 17 November 2025).
85. Anselin, L.; Sridharan, S.; Gholston, S. Using Exploratory Spatial Data Analysis to Leverage Social Indicator Databases: The Discovery of Interesting Patterns. *Soc. Indic. Res.* **2007**, *82*, 287–309. [CrossRef]
86. Sisman, S.; Aydinoglu, A.C. Improving Performance of Mass Real Estate Valuation through Application of the Dataset Optimization and Spatially Constrained Multivariate Clustering Analysis. *Land Use Policy* **2022**, *119*, 106167. [CrossRef]