geographical analysis

Geographical Analysis (2022) **0**, n/a

A Hybrid Approach for Mass Valuation of Residential Properties through Geographic Information Systems and Machine Learning Integration

Muhammed Oguzhan Mete and Tahsin Yomralioglu

Geomatics Engineering Department, Istanbul Technical University, Istanbul, Turkey

Geographic Information Systems (GIS) and Machine Learning methods are now widely used in mass property valuation using the physical attributes of properties. However, locational criteria, such as as proximity to important places, sea or forest views, flat topography are just some of the spatial factors that affect property values and, to date, these have been insufficiently used as part of the valuation process. In this study, a hybrid approach is developed by integrating GIS and Machine Learning for mass valuation of residential properties. GIS-based Nominal Valuation Method was applied to carry out proximity, terrain, and visibility analyses using Ordnance Survey and OpenStreetMap data, than land value map of Great Britain was produced. Spatial criteria scores obtained from the GIS analyses were included in the price prediction process in which global and spatially clustered local regression models are built for England and Wales using Price Paid Data and Energy Performance Certificates data. Results showed that adding locational factors to the property price data and applying a novel nominally weighted spatial clustering algorithm for creating a local regression increased the prediction accuracy by about 45%. It also demonstrated that Random Forest was the most accurate ensemble model.

Introduction

Housing and modelling studies have great importance for both governments and citizens since property values are used in many transactions like planning, taxation, purchase, and sale (World Bank 1993; UN-GGIM 2019). House price prediction models not only provide property value estimation but also enable systematic planning and policy development. Hedonic Pricing, Nominal Valuation, Multiple Regression Analysis (MRA), Ensemble Regression Methods, and Artificial Neural Networks (ANN) are some of the most used methods for mass property valuation (Pagourtzi et al. 2003; Jahanshiri, Buyong, and Shariff 2011; Wang and Li 2019; Mohd et al. 2020).

Correspondence: Muhammed Oguzhan Mete, Geomatics Engineering Department, Istanbul Technical University, Istanbul, Turkey.
e-mail: metemu@itu.edu.tr

Submitted: January 28, 2022. Revised version accepted: September 14, 2022.

doi: 10.1111/gean.12350

It is possible to categorise mass valuation methods into two groups: Hedonic-based regression and Machine Learning regression approaches. Hedonic pricing is still one of the most used methods for housing price prediction, evaluating both internal and external characteristics of the properties to identify the price. Hedonic models provide robust prediction results with accurate data and can be adapted to any value criterion. However, it may require an expertise in statistics and model specification process.

There is also a remarkable increase in the use of Machine Learning methods in housing price prediction applications. Ensemble Machine Learning methods like Random Forest (Dimopoulos et al. 2018; Yilmazer and Kocaman 2020; Aydinoglu, Bovkir, and Colkesen 2021; Ho, Tang, and Wong 2021), XGBoost (Peng, Huang, and Han 2019; Li 2022), CatBoost (Wang, Wang, and Liu 2021; Wang and Zhao 2022), LightGBM (Truong et al. 2020; Liu et al. 2021) provide more accurate predictions by aggregating outputs from several learners. Yet, ensembles are less interpretable, can be computationally expensive, and lack considering spatial effects of determinants during the prediction process. There is a need for integrating the spatial variables into the predictive model using Geographic Information Systems (GIS).

Open governmental data is one of the most useful data sources used for the understanding of cities and there is an increasing action toward making public data openly available (Arribas-Bel 2014). Open data has many benefits like economic value, transparency, efficiency, and public service improvement (The European Data Portal 2021). Thanks to open government and municipality data portals, users can reach updated and accurate data easily. The United Kingdom has published numerous open governmental data on several topics through a data portal since 2010. HM Land Registry - Price Paid Data (PPD) is one such data set that contains sale records and basic physical attributes of property sales transactions in England and Wales (Price Paid Data 2022). Energy Performance Certificates (EPC) data is also open data which holds many informative attributes such as total floor area, energy ratings, carbon dioxide emissions, heating costs of the buildings in England and Wales (EPC 2022).

Although the PPD data set contains full address information, it does not include a unique address identifier or coordinates to match with each other and with other data sets. There are several studies that developed different address matching methods for linking the two data (Fuerst et al. 2016; Powell-Smith 2018; Chi et al. 2021a). Chi et al. (2021a) published the linked PPD-EPC data set openly for the residential property transactions between 2011 and 2019. Yet, it is difficult to carry out detailed (street or parcel level) data analysis/analytics in GIS since the precise location information is missing. Locational factors like proximity to important places, sea or forest views, flat topography have a significant effect on the property value (Wyatt 1997; Kiel and Zabel 2008; Mete and Yomralioglu 2019). So although the linked PPD-EPC data contains many attributes for properties in England and Wales, it still lacks the locational factors needed to build more accurate regression models (Clark and Lomax 2018).

Most of the house price prediction studies generate a global regression model which omits spatial variability and geographic influence of the valuation factors. Implicitly, global models are assumed that the estimation process is location-independent and that property attributes have the same characteristics for the entire study area. In contrast, local models group properties with similar features based on their location. In this paper, we propose two-step approach with nominally weighted spatial clustering and spatially enriched machine learning regression methods. This novel method uses nominal criteria weights obtained by geographical analyses and enables defining value regions (similar to Broad Rental Market Areas of Valuation Office Agency) according to spatial variates and price per meter square. Weighting the features during

cluster analysis provides more accurate house price prediction and reveals the influence of the spatially dependent factors. Local SHapley Additive exPlanations (SHAP) values increase model explainability and give more detailed information about the influencing factors in different value zones.

The main objective of this research is to create an accurate mass property valuation model that integrates GIS and Machine Learning algorithms. In this sense, the article covers the use of GIS-based Nominal Valuation Method for creating a land value map of Great Britain. It then compares the performance of Machine Learning regression models for mass valuation of residential properties in England and Wales. It also explains batch geocoding process of PPD-EPC address data for matching the Unique Property Reference Number (UPRN). Lastly, it discusses the feature enrichment of PPD-EPC data by adding proximity, terrain, and visibility analysis scores and examines the spatial dependence of the factors by clustering analysis. This study proposes a hybrid method for accurate mass property valuation that can be implemented anywhere in the world.

Real estate valuation

Real estate valuation methods can be divided into two groups: Classical and mass valuation. Classical methods consist of Sale Comparison, Income, and Cost approaches which are frequently used for single property valuation at point in a time (IVSC 2020). Mass property valuation methods like Hedonic Pricing, Nominal Valuation, MRA, Decision Trees, Ensemble Methods, and ANN are applied for multiple property valuation.

Mass valuation

Mass valuation can be defined as analyzing a set of factors to estimate the value of a large number of properties using statistical methods and standards (IAAO 2013a). There are a wide range of mass valuation methods like MRA (Zurada, Levitan, and Guan 2011; Benjamin, Guttery, and Sirmans 2020; Yilmazer and Kocaman 2020), Hedonic Pricing (Peterson and Flanagan 2009; Lisi 2019; Yamani, Ettarid, and Hajji 2019), Nominal Valuation (Yomralioglu 1993; Yomralioglu and Nisanci 2004; Mete and Yomralioglu 2019), Geographically Weighted Regression (GWR) (Huang, Wu, and Barry 2010; Dimopoulos and Moulas 2016; Wang, Li, and Yu 2020), Ensemble Methods (Alfaro-Navarro et al. 2020; Aydinoglu, Bovkir, and Colkesen 2021; Gnat 2021), and ANN (Demetriou 2017; Lee 2022; Yalpir 2018). With the improvements in computing power, GIS and Artificial Intelligence, Computer-Assisted Mass Appraisal (CAMA) applications have become widespread and Automated Valuation Models (AVMs) have been adopted in many countries (Wang and Li 2019; Renigier-Biłozor et al. 2022).

GIS-based valuation approach

As a decision support system, GIS is an essential component of the CAMA approach. GIS can be utilized for revealing the effect of locational criteria on property values. In addition to monitoring and visualization functions, GIS also provides robust spatial analysis capabilities for mass valuation studies (Longley, Higgs, and Martin 1994). GIS-based mass valuation methods like GWR, Spatial Analysis, and Nominal Valuation offer fast and accurate assessment of properties located in broad areas.

Nominal Valuation is a stochastic method based on the Weighted Linear Combination approach which is used to aggregate criteria with different importances. It can be applied to both

land and buildings by assigning scores for each criterion. The total nominal value of a property is calculated as the weighted sum of criterion scores multiplied by the parcel or pixel area (1).

$$V_i = S_i * \sum_{j=1}^{k} \left(f_{ji} * w_j\right), \tag{1}$$

where V is total nominal value, S is parcel or pixel area, f is factor score, w is factor weight, and k is total number of the factors.

Nominal Valuation has many advantages in comparison with other mass valuation methods: Since it is not directly dependent on market prices, it can be easily applied to real estate across large areas to reveal value differences based on spatial analyses. Using pixel-based valuation, Nominal Value maps can be created with high granularity, such as subparcel level, and it is also possible to calculate market prices from nominal values using reliable property prices.

Machine learning-based valuation approach

Machine learning methods have applications across finance, investment, real estate valuation (Groth and Muntermann 2009; Baldominos et al. 2018; Dixon, Halperin, and Bilokon 2020). In mass property valuation practices, regression analysis and neural networks are mostly utilized for unbiased and accurate price prediction. Regression analysis is used to estimate relationships between a dependent variable and one or more independent variables. Several regression methods such as Linear, Ridge, Lasso, Polynomial, and Bayesian Regression, as well as ensemble regression models, are used in property price prediction.

Multiple Linear Regression is a method that fits a linear model to the data for explaining the relationship between dependent and independent variables (2). The Ordinary Least Squares approach, regression minimizes the residual sum of squares between the actual and predicted values. Although linear regression is one of the most used regression models, it does not perform well on complex data which cannot be expressed in a linear form.

$$\hat{y} = b_0 + b_1X_1 + b_2X_2 + ... + b_nX_n, \tag{2}$$

where $\hat{y}$ is dependent variable, X_n are independent variables, b_0 is y-intercept, and b_n are coefficients.

Ensemble methods can be defined as the combination of multiple weak models to create a stronger predictive model. Many ensemble learning algorithms are developed by aggregation of several methods. Bagging, boosting, and stacking are three main ensemble approaches.

The Random Forest is an ensemble algorithm composed of many individual Decision Trees (Breiman 2001). It is a flexible algorithm that can accomplish both classification and regression problems with high accuracy. As an extension of the bagging approach, Random Forests generate random samples of features for each decision tree and combine them by aggregating the decision (bootstrap and aggregation). For regression tasks, the aggregate decision becomes the average of all decisions. Hence, Random Forest regression handles the overfitting problem by reducing model complexity (variance).

Bagging approach fits the model independently by combining several weak learners. In contrast, boosting uses a sequential method to fit the the model iteratively and allocates weights to each resulting model. Then it readjusts model weights utilizing the error of the prior tree to improve the accuracy. Boosting algorithms reduce bias by increasing model complexity in order to prevent underfitting. By optimizing the loss function of the boosting algorithms, Gradient

Boosting ensemble methods have been developed recently. XGBoost (Chen and Guestrin 2016), LightGBM (Ke et al. 2017), and CatBoost (Prokhorenkova et al. 2017) stand out as the most used and most successful gradient boosting algorithms.

Stacking is another ensemble approach which uses heterogeneous weak learners as opposed to bagging and boosting. It combines different learning algorithms by training a meta model for better predictive models. In order to reduce the bias of the model prediction, the output of each individual model are stacked together and used as input to a final estimator.

Material and methods

We implement a hybrid approach for mass valuation of residential properties using a combination of GIS and machine learning methods. In this study, PPD and EPC open data are used for mass valuation of residential properties in England and Wales. As a study area, where possible, we include all three countries that constitute Great Britain: England, Wales, and Scotland.

Price paid data

The United Kingdoms has made several governmental data openly available and accessible in the last decade (Open Data Barometer 2017; Open Knowledge International 2018). PPD is being published in March 2012 monthly by HM Land Registry for supporting open government and data transparency (Price Paid Data 2022). It covers residential property transactions since January 1995 in England and Wales. Sales that were not for value or sales have not been lodged with HM Land Registry are excluded from PPD data. PPD is being used extensively by proptech startups and SMEs (Small- and Medium-Sized Enterprises) to develop innovative property products and services (Hogge 2016).

PPD includes several attributes like address, date of transfer, property type, ownership type which are very useful in price prediction for mass valuation studies (Table A1). PPD is available in both Comma-separated values (CSV) and linked data formats with the four-star rating that enables running SPARQL queries for searching the data and generating reports. The data set is licensed under the Open Government Licence v3.0.

Energy performance of buildings certificates data

Since 2012, residential and commercial properties must have an Energy Performance Certificate (EPC) when constructed, sold, or let (EPC Regulations 2012). EPCs were introduced in 2007 with the aim of improving energy efficiency by reducing carbon emissions. A data set of domestic and nondomestic certificates in England and Wales is being published by United Kingdom Department for Levelling Up, Housing and Communities (formerly Ministry for Housing, Communities and Local Government) for every quarter since Q4 2008 to date. It covers many features from physical building properties to energy use and costs for different building types (Table A2). Even though EPC data has several informative features about properties, it contains remarkable errors affecting up to 80% of records. When using this data, one should take care of this limitation (Hardy and Glew 2019). The data set is available in both CSV format and Application Programming Interface (API) service for developers (EPC 2022).

Data geocoding

PPD and EPC contain numerous distinct attributes including full address information retrieved from OS AddressBase Premium data. Chi et al. (2021a) linked those two separate open data

sets for the period from 2011 to 2019 so as to benefit from rich attributes in housing studies. They carried out the data linkage process for 5,732,838 transactions in England and Wales using 251 matching rules with over 90% success and published the resulting data set on the UK Data Service ReShare repository (Chi et al. 2021b).

The linked data set was published in CSV format with multilevel spatial units like Middle Layer Super Output Area, Lower Layer Super Output Area, and Local Authority District codes. However, GIS analyses requires precise location information (geographic/projection coordinates or unique address identifier) for producing property value maps. Since Ordnance Survey AddressBase Plus/Premium products are not free and openly available, we followed an alternative method to match the text address information with a UPRN, unique identifier for every addressable location in the United Kingdom. In this sense, batch geocoding process carried out for linked PPD-EPC data in geopy.geocoders Python client using the HERE Geocoding & Search API service. 5,732,838 properties were geocoded with 99.8% success using 115 chunks in Jupyter Notebook environment.

After having the geocoded PPD-EPC data, each property is assigned to the nearest UPRN point with spatial join (Join Attributes by Nearest) operation in QGIS. In order to match the property with the correct UPRN point, maximum nearest neighbours specified as 1 and maximum distance limited to 50 m. Both the LSOA codes of the Office for National Statistics and UPRN codes of the recently published EPC data are used for validating the geocoding results.

Mass property valuation using GIS and machine learning

To make use of this enhanced data set, we created a hybrid mass valuation method which brings spatial analysis power together with the machine learning algorithms by feature enrichment approach (Fig. 1).

Valuation criteria

Criteria which affect property values are numerous. Proximity to important buildings, topographic features, view, physical and legal status are some of the internal and external determinants of the property value. Broadly, these factors can be classified as spatial, physical, and legal (Yomralioglu 1993; Wyatt 2013; Bünyan Ünel and Yalpir 2019; Mete, Guler, and Yomralioglu 2022).

Determination of valuation criteria is one of the important steps of the valuation process. In addition to PPD-EPC data attributes, several locational factors that may affect the property value are included in this study. In order to carry out proximity, terrain, and visibility analysis for spatial criteria, Ordnance Survey, OpenStreetMap, and European Environment Agency (EEA) Copernicus Land Monitoring Service open data sources are used (Table A3).

Creation of nominal land value map

Land value maps provide an understanding of price variations in a region and facilitate management, planning, and taxation activities. Using the aforementioned open data sources, a land value map of Great Britain is produced with the GIS-based Nominal Valuation method. First, a valuation database is created in PostgreSQL (v14.0) and PostGIS (v3.1.4) open-source database management system and all raster and vector data are imported into the spatial database. In order to carry out proximity, terrain, and visibility analyses, QGIS Graphical Modeler and the Python API (PyQGIS) are used (Fig. 2).

Proximity to important places is one of the most influential factors in land valuation (Wyatt 1997; Waltert and Schläpfer 2010; Tajani, Morano, and Ntalianis 2018; Bünyan Ünel and

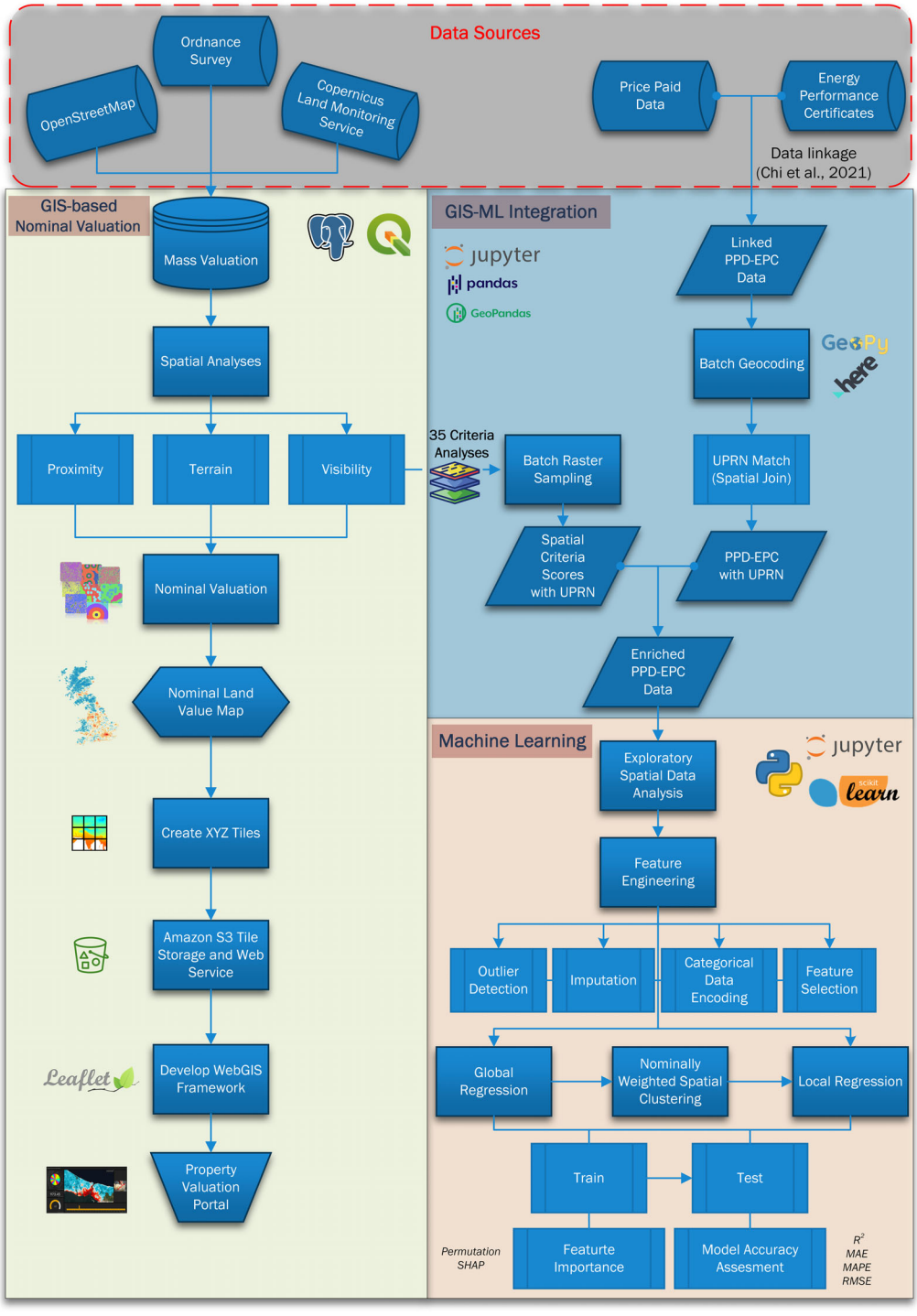

Figure 1. Workflow diagram of the proposed approach. [Colour figure can be viewed at wileyonlinelibrary.com].

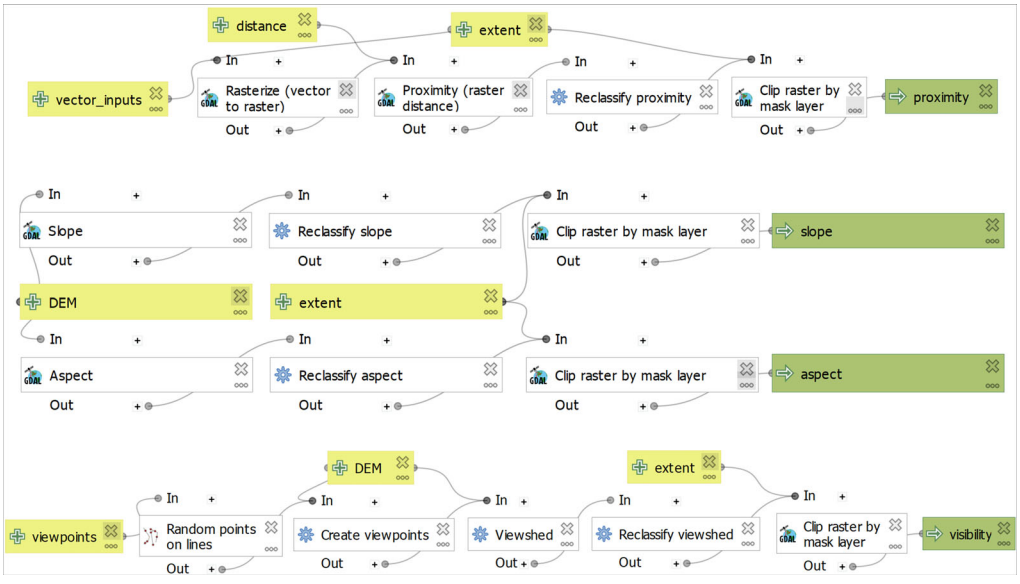

Figure 2. Spatial analyses model created in QGIS Graphical Modeler. [Colour figure can be viewed at wileyonlinelibrary.com].

Yalpir 2019; Mete and Yomralioglu 2019). Euclidean Distance, which represents the length of a line segment between the points, is used for proximity analysis. Then reclassification intervals of distances are determined separately for walking and driving options. For instance, if the walking distance to the destination (e.g. subway station) is between 0 and 400 m, proximity score of the property will be 100 (maximum score). On the other hand, a property with a driving distance up to 1 km to the destination (e.g. shopping mall) will have 100 proximity score.

Terrain features like slope and aspect also have a significant effect on land value. Having a property located on flat terrain and benefiting from long sunlight duration (south-facing slope) is preferable (Huang and Hewings 2021). Using EU-DEM open Digital Elevation Model data, percent slope and aspect analyses are carried out and terrain scores are reclassified between 0 and 100.

Pleasant views such as sea, river, or forest add great value to the property (Yu, Han, and Chai 2007; Wallner 2012). In order to determine views across the region, Visibility Analysis Plugin is used in QGIS. Firstly, random points are created on the sea and river line since the viewshed tool accepts only point vector data for viewpoints as an input. After creating viewpoints for each view type, viewshed analysis is carried out and visibility scores are reclassified between 0 and 100.

Consequently, to produce GIS-based land value map of Great Britain, Nominal Valuation method is used to calculate weighted sum of all 35 criteria (Fig. 3). The criteria weights are retrieved from feature importance results of spatial factors given in the next subsection. In order to share and visualize the land value map online, 256×256 pixels raster tiles (XYZ Tiles) are generated in QGIS and uploaded to the Amazon Web Services (AWS) S3 bucket for map service on serverless cloud architecture (Mete and Yomralioglu 2021). Then Leaflet open-source JavaScript web mapping library is used to develop a Web GIS application for consuming and visualizing the map tile service on the browser (https://web.itu.edu.tr/metemu/nominal/uk.html).

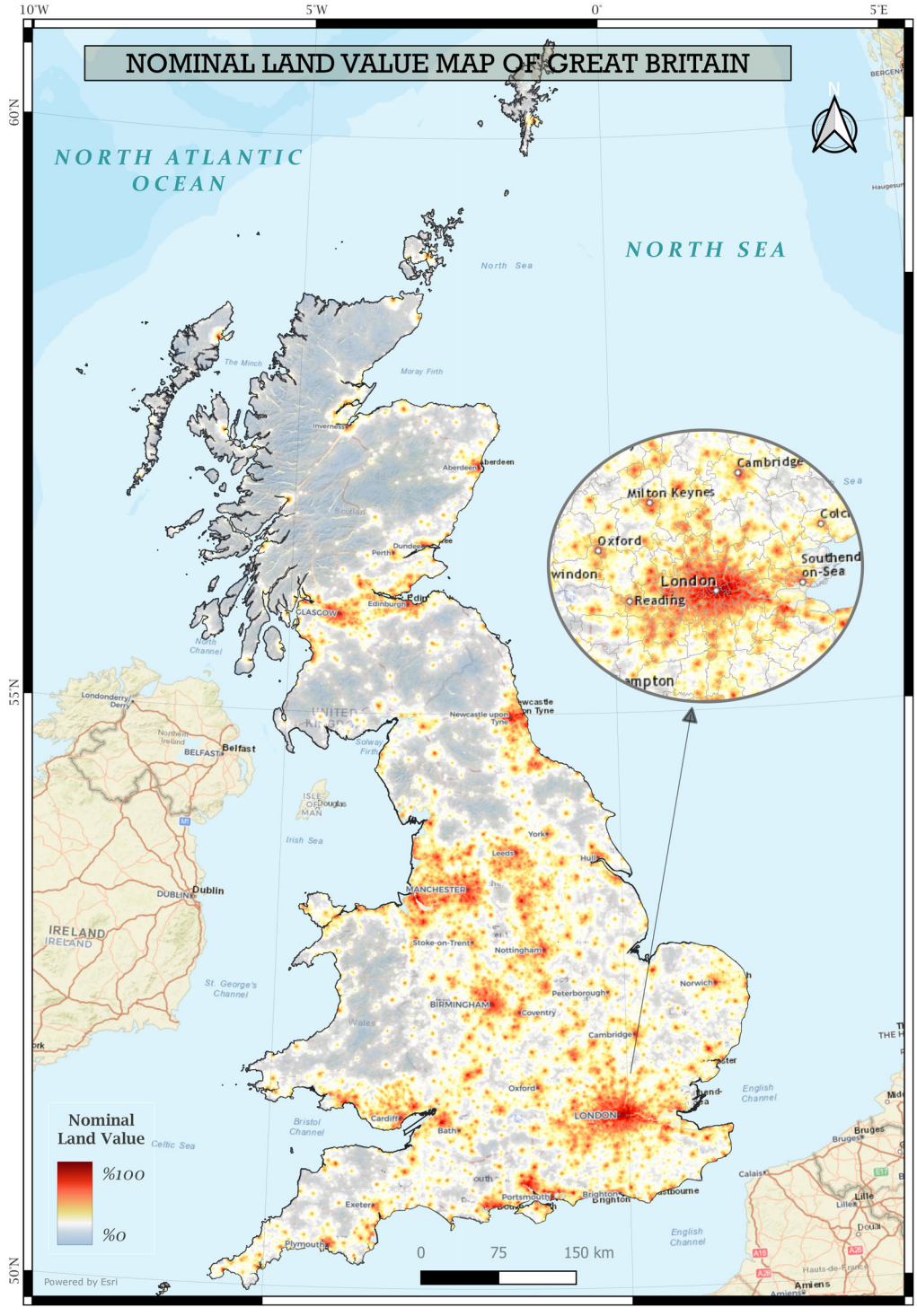

Figure 3. Nominal land value map of Great Britain, United Kingdom. [Colour figure can be viewed at wileyonlinelibrary.com].

Regression analysis

Regression analysis is one of the fundamental machine learning methods and it is widely used for property price prediction. In this paper, several machine learning regression models like multiple linear regression, XGBoost, CatBoost, LightGBM, random forest are used for residential property valuation in England and Wales.

Firstly, PPD-EPC data is inspected in the phase of Exploratory Data Analysis (EDA). EDA is essential for data science projects in terms of understanding the data well. In this step, descriptive statistics (minimum, maximum, mean, standard deviation, etc.) are calculated, null values are checked, correlation matrix, box plots, scatter plots, pair plots, histogram of the dependent variable are created. Initially, the data had 5,627,022 rows and 23 columns with different categorical and numerical data types. We used Python (Van Rossum and Drake 2009) with the Pandas (McKinney 2010) and scikit-learn (Pedregosa et al. 2011) libraries, and the code is available as a Jupyter notebook (Kluyver et al. 2016) in the Data Availability Statement section.

In the feature engineering phase, outlier detection, missing data imputation, categorical data encoding, and feature selection processes are performed. Outliers can skew the data and cause poor fitting in regression analysis. Outlier detection can be conducted by observing the boxplots, standard deviation, and interquartile range as well as with automated methods like Isolation Forest, Minimum Covariance Determinant, One-Class SVM. In this study, outliers are detected and removed by applying the 3σ distribution standard deviation. In addition, missing values are imputed with reasonable values by checking the median and mean statistics. For example, the floor-level attribute had null values and they are replaced with zero.

Almost all of the regression algorithms require input and output data to be numerical variables. Therefore, the categorical features need to be encoded. We applied One Hot Encoding to features including property type (detached, flats, semi-detached terraced), old/new, and duration (freehold, leasehold).

After data preprocessing steps, the data is split as 90% train, 5% validation, 5% test and several regression models like linear regression, XGBoost, CatBoost, LightGBM, random forest are built. Each model is trained with default parameters to see their prediction success on the PPD-EPC data and to continue optimizing the most accurate ones. In order to measure model performances, several accuracy metrics are used such as R^2, Adjusted R^2, Mean Bias Error (MBE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE).

Coefficient of determination (R^2) indicates how dependent and independent variables are correlated in the regression (3). By calculating the proportion of total variance, it demonstrates the goodness of fit of the model. On the other hand, Adjusted R-squared is a modified form of R-squared which adjusts for the number of variables in a model (4).

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y}_i)^2}, \tag{3}$$

$$\text{Adjusted } R^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}, \tag{4}$$

where y is actual value, $\hat{y}$ is predicted value, n, is sample size, and p is number of the independent variable.

Bias in the model measures the difference between average prediction and true value. High bias causes underfitting and high model error. On the other hand, variance shows how much

a random variable differs from the expected value. It indicates the amount of variation in the prediction when different set of training data is used. A model with high variance learns a lot and performs well with the training data set, but it does not generalize well with unseen data, and may show high error in the test data due to overfitting. Simple models like linear regression generally have high bias and low variance. Conversely, complex models like random forest generally have low bias and high variance. Bagging and boosting ensemble learning algorithms configure the bias variance trade-off (Bühlmann 2012).

MBE basically refers to the average bias in the model. It is calculated as average of differences of predicted and true values (5). MAE refers to the difference between the actual value and predicted value (6). It measures the average errors in the regression analysis by summation of the absolute value of the residual considering the number of observations. MAPE is the sum of prediction errors divided by the actual values (7), and shows the mean absolute deviation as a percentage.

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i). \tag{5}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|. \tag{6}$$

$$\mathrm{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{7}$$

RMSE is a metric that defines the difference between the actual value and predicted value by calculating the square root of the average of squared errors (8). Lower RMSE means a better fit for the prediction.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{8}$$

Making predictions with trained regression models, several accuracy metrics are calculated on validation data. For the ensemble methods, cross-validated randomized search is applied on specified hyper parameters like number of trees, maximum number of levels in a tree, minimum number of samples required at each leaf node. Then, GridSearchCV is used to reveal the best estimator parameters by focusing on the parameter grid based on the results of the randomized search. Thus, the best model parameters are determined to train the regression model for an accurate prediction. Results showed that random forest algorithm had better performance compared to other regression models in terms of accuracy (Table 1).

Table 1. Accuracy Metrics of the Global Regression Models

Model	R^2	Adjusted R^2	MAE	MBE	MAPE	RMSE
Linear	0.300	0.300	113,075	1,205	0.567	227,530
XGBoost	0.434	0.434	98,505	942	0.469	204,625
CatBoost	0.455	0.455	98,811	964	0.474	200,791
LightGBM	0.437	0.437	101,291	1039	0.494	204,170
Random forest	0.469	0.469	95,934	927	0.461	198,211

Feature enrichment for machine learning model

In order to build more accurate regression models, there is a need for the inclusion of locational factors, which are highly correlated with land price (Wyatt 1997; Kiel and Zabel 2008; Mete and Yomralioglu 2019). GIS provides numerous spatial analysis tools that can be utilized for revealing locational criteria effects on the property price. Thus, nominal scores of proximity, terrain, and visibility analyses are extracted for feature enrichment of regression models built for PPD-EPC data.

There are several ways of calculating pixel value information for raster data such as zonal statistics, raster sampling, raster to point conversion. In this study, each of the 35 criteria outputs raster data are sampled on UPRN point data in order to get the corresponding pixel value (nominal analysis score) by using Batch Raster Sampling processing algorithm in PyQGIS.

```
root = QgsProject.instance().layerTreeRoot()
input_layer = 'uprn.gpkg'
result_layer = input_layer
unique_field = 'fid'
for layer in root.children():
   prefix = layer.name()
   params = {'RASTERCOPY': layer.name(),
   'INPUT': input_layer,
   'COLUMN_PREFIX': prefix,
   'OUTPUT': 'memory:'
   }
   result = processing.run("native:rastersampling", params)
   rastersampling = result['OUTPUT']
   params = { 'INPUT': result_layer, 'FIELD':unique_field,
     'INPUT_2': rastersampling, 'FIELD_2': unique_field,
     'FIELDS_TO_COPY': prefix + '1',
     'OUTPUT': 'memory:'}
   result = processing.run("native:joinattributestable", params)
   result_layer = result['OUTPUT']
QgsProject.instance().addMapLayer(result_layer)
```

To enrich the valuation data with 35 new spatial features created with spatial analyses, extracted nominal scores of all criteria are added as new attribute columns in the UPRN data and joined with the PPD-EPC data set. Examining the feature correlations, it is seen that spatial factors derived from GIS are positively correlated with the price variable.

After deriving the enriched data, the same workflow is followed for building global regression models. Regression models are trained with the enriched data by using Randomized Search and GridSearchCV methods. After getting the best parameters for the models, the best estimator is used in the prediction phase of test data. Table 2 shows the training execution times and accuracy metrics of the models.

Feature importance is a part of feature selection and it is used to interpret the model for better understanding how much a feature affected the prediction process. Thus, permutation feature importance scores are calculated for random forest regression (Fig. 4).

Table 2. Accuracy Metrics of the Global Regression Models after Feature Enrichment

Model	R^2	Adjusted R^2	MAE	MBE	MAPE	RMSE
Linear	0.447	0.447	101,442	821	0.499	200,851
XGBoost	0.791	0.791	65,367	527	0.311	123,585
CatBoost	0.838	0.838	56,668	568	0.266	108,740
LightGBM	0.806	0.806	65,567	592	0.319	119,042
Random forest	0.858	0.858	44,888	502	0.194	101,847

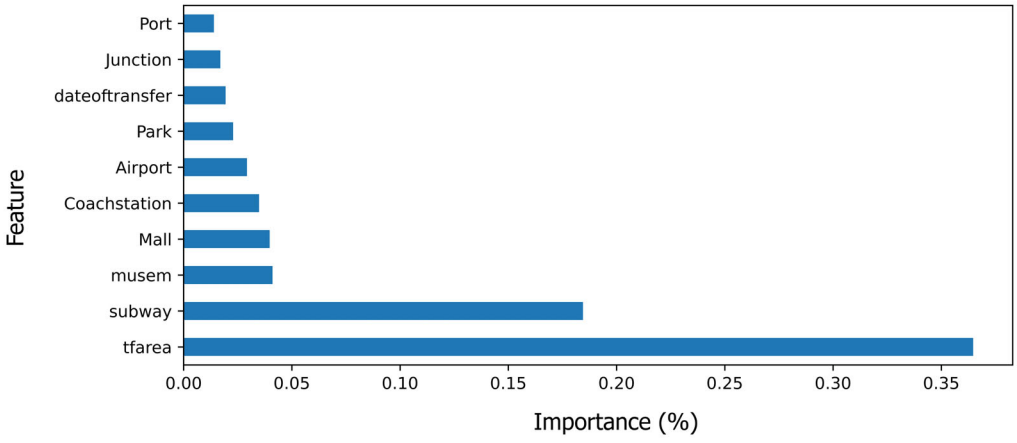

Figure 4. Feature importance scores after feature enrichment process. [Colour figure can be viewed at wileyonlinelibrary.com].

Moreover, SHAP is a state-of-the-art eXplainable Artificial Intelligence (XAI) algorithm which quantify the contribution of the features to the predictive Machine Learning model using shapley values of game theory (Lundberg and Lee 2017). Besides feature importance, SHAP values are calculated to understand and explain decision making behavior of the models (Fig. 5). Based on the SHAP summary plot, the top 20 features are shown and their impacts on the prediction are explained.

After completing the mass valuation of residential properties, ratio studies are carried out to ensure the quality of the assessment. Hence, level of appraisal, coefficient of dispersion (COD), and price-related differential (PRD) performance indicators of International Association of Assessing Officers are calculated (IAAO 2013b; IAAO 2018). Using 281,350 test samples, COD is calculated as 14.92 and PRD is calculated as 1.03.

Nominally weighted multivariate spatial clustering

In global regression models, there is a single weight for each factor that affects the value across the entire study area. Representing the criterion weight with a fixed value can cause a high bias value in mass valuation studies. For example, the importance of the criteria such as proximity to airports, proximity to prayers, or forest views may vary in different geographical regions. The assumption that these criteria affect the value with the same degree of importance negatively affects prediction model accuracy.

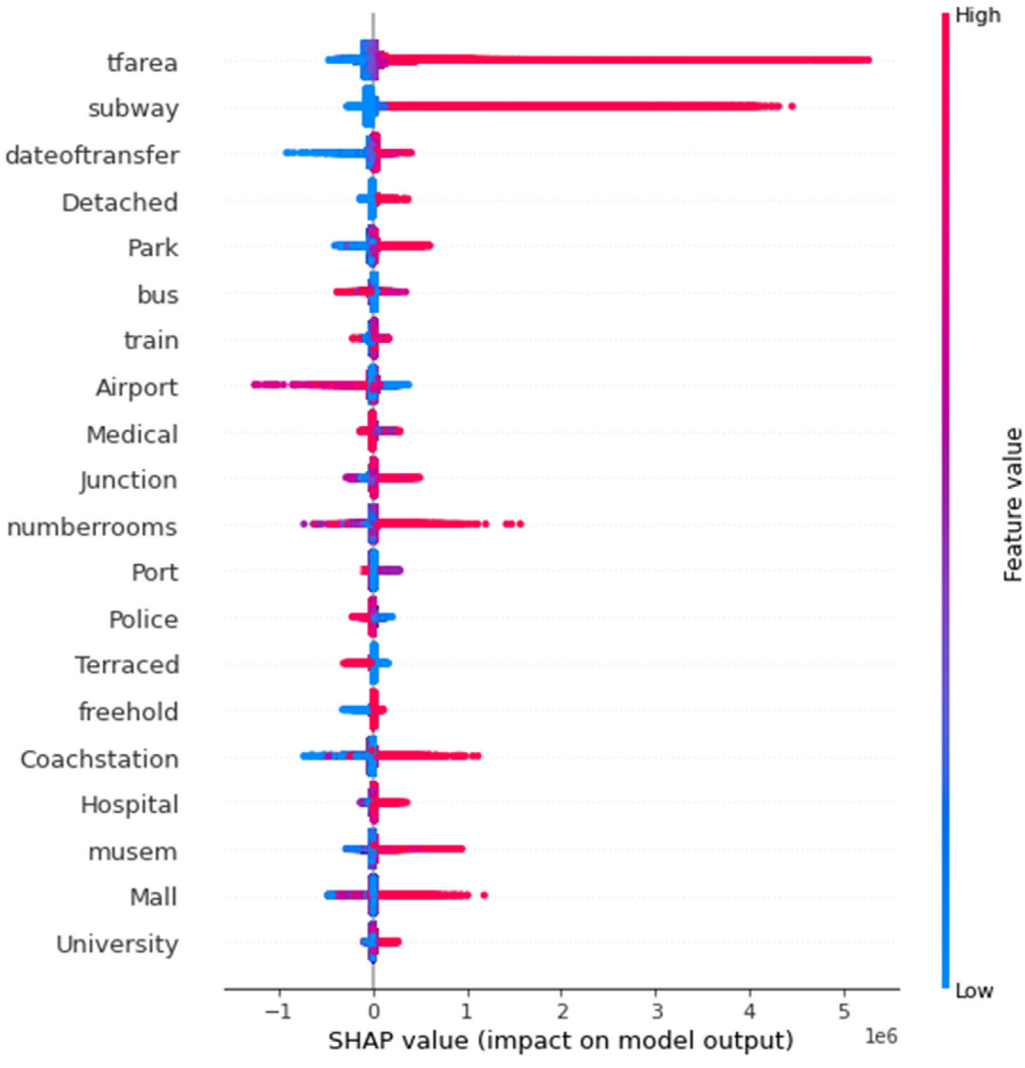

Figure 5. SHAP values after feature enrichment process. [Colour figure can be viewed at wileyonlinelibrary.com].

According to the first law of geography, near things are more related than distant ones (Tobler, 1970). In order to build consistent and accurate house price prediction models, it is necessary to create value regions by clustering properties that are close to each other and that have similar characteristics. Thus, local criterion weights can be assigned for each cluster.

Spatial autocorrelation and spatial heterogeneity of the data should be examined before spatial modeling. General distributions of variables, trends, spatial dependence, clustering status and detection of outliers should also be discussed. In this context, Global Moran's I analysis can be used (9). The global spatial autocorrelation index summarizes geographical similarities based on the neighborhood relationship of the data. The Global Moran's I index together with P-value and z-score explains whether there is clustering in the study region. Statistically significant P-value, positive z-score and positive Moran's I index value indicate clustering in the study area (Moran 1950; Li, Calder, and Cressie 2007).

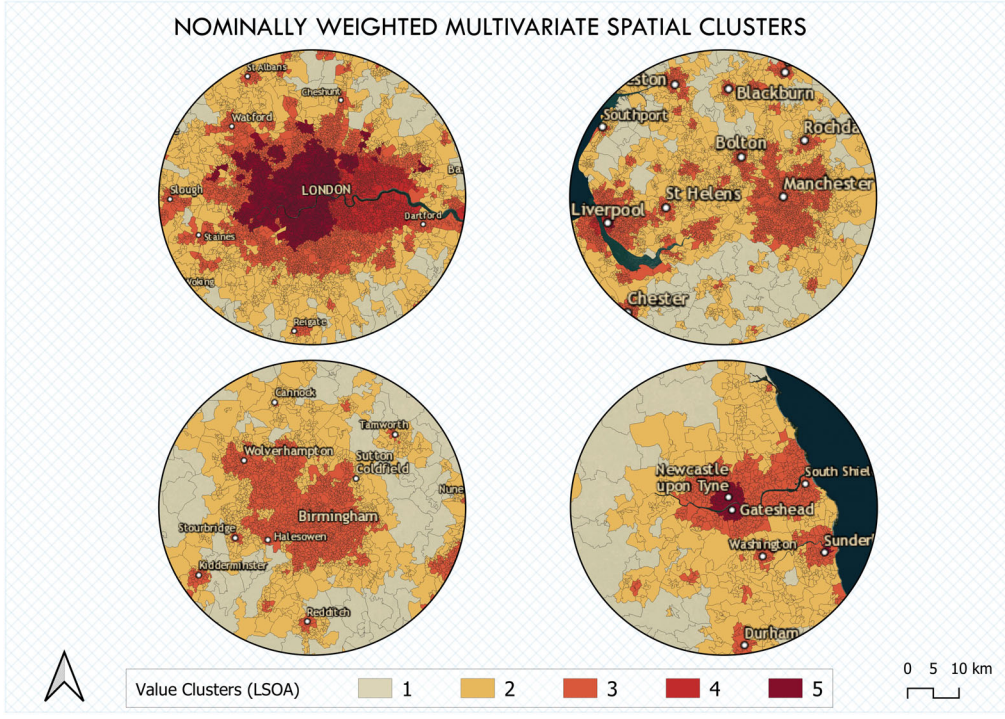

Figure 6. Value regions created with nominally weighted multivariate spatial clustering analysis. [Colour figure can be viewed at wileyonlinelibrary.com].

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}(x_i - \bar{x})(x_j - \bar{x})}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j}\right) \sum_{i=1}^{n} (x_i - \bar{x})^2}, \tag{9}$$

where I is Moran's I statistic, x_i is value of the variables in any location, and w is weight.

Spatial autocorrelation is determined as a result of the analysis carried out on PPD data in England and Wales (Moran's I Index: 0.7110, z-score: 1,832.3314, P-value: 1×10^{-6}). Due to the nonstationary nature of the house price data, regional value changes are observed. Therefore it is concluded that spatial regression analysis is needed in order to model the relations of the geographic variables. In this context, Nominally Weighted Multivariate Spatial Clustering analysis was carried out by customizing the k-means method to determine the value regions. The weights of the variables are obtained by calculating the importance scores of the global regression model that was applied in the previous step. Therefore, properties that have similar characteristics and are close to each other in terms of geographical location are included in the same cluster.

In clustering studies, the number of clusters can be determined by various methods such as Elbow Method, Average Silhouette, Gap Statistic, or Dendrogram Diagram. In this study, the elbow method was used to determine the number "k" of clusters. The method has an intuitive approach: It is defined as a function of the variance value of the variables and the number of clusters. The number of clusters was determined to be five according to the elbow point where the variance values changes suddenly. The distribution of the value regions obtained as a result

Table 3. Accuracy Metrics of the Local Random Forest Regression Models

Cluster	Sample	R^2	Adjusted R^2	MAE	MBE	MAPE	RMSE
1	639,866	0.877	0.877	68,491	433	0.168	71,468
2	2,457,593	0.872	0.872	33,160	464	0.178	61,798
3	1,963,335	0.863	0.862	31,199	254	0.140	65,611
4	286,208	0.826	0.826	38,587	557	0.190	98,211
5	279,927	0.871	0.871	65,367	426	0.161	70,178

of the clustering analysis is shown in Fig. 6. When the Greater London region is examined, it is seen that the City of London, City of Westminster, Kensington and Chelsea, Hammersmith and Fulham stay in the highest value region which is located around the Thames River, while other value regions are spatially clustered toward the outer peripheries.

Local regression models are trained and tested for each value region and various performance metrics. Looking at the results of the clustered regression, it is observed that the local random forest algorithm achieved the highest prediction accuracy in all value regions (Table 3). On the other hand, local regression models performed better prediction performance in comparison with global regression.

SHAP values for features explaining the influence of the spatial criteria are given in the Data Availability Statement section for each cluster.

Results and discussion

There is a well-known argument that 80% of the world's data have some kind of location aspect (Williams 1987; Franklin and Hane 1992). Hence, spatial data should be used and managed effectively in order to deal with most real-life problems. Criteria that affect property prices mostly contain geographic factors as well. That is why valuation data should be capable of expressing spatial effects on the price (Clark and Lomax 2018).

Although the PPD-EPC data has numerous attributes about physical factors, it does not contain locational or environmental features of the property. In order to enrich the data and increase prediction accuracy, nominal scores of proximity, terrain, and visibility criteria are included by using GIS analyses. After the data enrichment process, R^2 increased by 39% (0.47–0.86), RMSE decreased by 49% (from 198,211 to 101,847) in global random forest model. Creating local regression models resulted in a significant further reduction in prediction error.

Ensemble machine learning methods like XGBoost, LightGBM, random forest are widely used for house price prediction. In this study, several regression methods are used on PPD-EPC data for mass property valuation purpose. Comparing regression model metrics, random forest algorithm had higher accuracy than other approaches. Multiple linear regression had a moderate performance which was not enough for explaining the complexity of the variables. Although the XGBoost, CatBoost, and LightGBM regression models were more accurate than the linear regression, they could not compete with the random forest algorithm. However, comparing the training durations, random forest model was the slowest and LightGBM was the fastest model among other regression algorithms. XGBoost, Catboost, and LightGBM are therefore reasonable alternatives in regression analysis since they perform high accuracy in a short training time.

Local regression analyzes are performed on five different clusters instead of a single global model created for the entire study area. The GWR method is frequently used as spatial regression

in housing valuation studies. However, since this method is a linear regression type, it may be insufficient in modeling nonlinear variables. For this reason, an alternative method is developed and ensemble learning regressions like random forest, XGBoost, CatBoost, and LightGBM are created in the local clusters. Regression models are trained and tested for each value region and various performance metrics. Looking at the results of the clustered regression analysis, it is observed that the local random forest algorithm achieved the highest prediction accuracy in all clusters.

The explainability of a machine learning model is crucial for understanding how and why decisions have been made. It provides trust and transparency by identifying model accuracy and feature importance in an interpretable way (Hagras 2018).

Feature importance and SHAP values indicated that spatial factors contribute greatly to the model during the property price prediction. After creating local regression models, SHAP summary plots are created for five clusters to understand the effect of the features in different value zones. Cluster 1 has the lowest mean value and 11.16% of properties fall within this value zone which is mostly located in rural areas. Therefore, physical features like total floor area and number of rooms are the most important factors of this cluster. Clusters 2 and 3 contain 77.07% of all properties. In terms of feature importance, they have similar SHAP value results. After total floor area and the number of rooms features, proximity to car parks, train stations, metro stations, and universities are criteria that highly affect the property value positively in those regions. Cluster 4 covers a relatively small group of properties from prominent English cities such as London, Liverpool, Manchester, and Newcastle. The most important features of this value region are proximity to metro stations, coach stations, shopping malls, car parks, and ferry docks. Unlike other value regions, proximity to the museums has great importance in cluster 5. There are more than 150 museums in the Greater London region and they are mostly located in notable neighborhoods. That is why proximity to the museums has a high SHAP value in the summary plot.

It is inferred that proximity to public transportation stations and sociocultural centers positively affects property prices across England and Wales. On the other hand, it is surprising to find that proximity to airports criterion generally has a negative effect on house prices.

Hyperparameter optimization is an essential part of the learning process. By tuning the model with the optimum parameters, model accuracy can be increased remarkably. Using RandomSearch and GridSearchCV selection models together, exhaustive parameter search is completed with fewer model fits in a shorter time compared to using GridSearchCV solely. During the Randomized Search and Grid Search operations, we limited the total number of fits, since large data size and scale caused out of memory error. To overcome this issue encountered in hyperparameter optimization, Cloud Computing or parallel processing can be utilized.

In order to measure the performance of the mass valuation study, level of appraisal, COD, and PRD indicators are calculated according to the IAAO standards by using both predicted and actual values. Level of appraisal examines the closeness of the predicted value to the actual value. It consists of mean ratio, weighted mean ratio, and median ratio. Those ratios are desired to be 1.00 theoretically, yet appraisal level value between 0.90 and 1.10 is acceptable. COD is another performance indicator of mass valuation study which can be expressed as the average percentage difference between the ratios and the median ratio. COD measures appraisal uniformity, and it should be between 5 to 10 for single-family residential properties, “newer or more homogeneous areas”; 5 to 15 for “older or more heterogeneous areas”; and 5 to 20 for “other residential properties.” PRD is used to measure the vertical equity by the mean ratio and weighted mean

ratio. PRD ratio between 0.98 and 1.03 is statistically meaningful in terms of vertical equity (IAAO 2013b). According to the ratio study results, it was seen that mass valuation study of residential properties in England and Wales was carried out in accordance with the IAAO ratio standards.

Several studies link the PPD and EPC data to add more useful attributes to the property sales data (Fuerst et al. 2016; Powell-Smith 2018; Chi et al. 2021a). Deriving spatial criteria scores from GIS, we increased the prediction accuracy of the regression models dramatically. Moreover, property value regions are created with a novel weighted spatial clustering approach and local regression models are generated to deal with spatially dependent, nonstationary data. This paper aims to fill a key gap in taking care of neglected spatial variables in housing and modeling studies with integrated GIS and machine learning approaches.

Conclusion

Adoption of mass valuation techniques in value-based activities (taxation, zoning applications, planning, etc.) has been increasing in recent years. Hedonic Pricing, Nominal Valuation, Spatial Analysis, and Regression Analysis are some of the prominent methods of mass valuation practices. Although there are studies using Machine Learning algorithms for mass property valuation, the spatial factors that affect property value are neglected. In this study, GIS-based Nominal Valuation method and Ensemble Machine Learning regression algorithms were used for mass valuation of residential properties in England and Wales.

Benefiting from the robust spatial analysis capability of GIS, the effects of the locational factors on the property value can be unveiled. In order to enrich the PPD-EPC data with significant attributes, proximity, terrain, and visibility analyses were carried out. By applying the nominal valuation method, a property value map of Great Britain was created in QGIS and published online. For the data enrichment process, spatial analysis results were assigned to the UPRN data with the spatial join operation so that precise location information could be added to the PPD-EPC.

By applying the same procedures to the enriched PPD-EPC data for mass property valuation, global regression analyses were carried out with higher accuracy. Results showed that the random forest algorithm had better performance in price prediction compared to other regression models. Adding spatial features to the PPD-EPC data, model accuracy were increased substantially. Examining the feature importances and SHAP values of the regression model showed that locational factors were contributed to the prediction process greatly. To conclude, if random forest regression model is used on PPD-EPC data without proximity, terrain, and visibility spatial analyses, the model accuracy is mediocre (R^2: 0.47, MAPE: 0.46, RMSE: 198,211). If additional spatial features are included in the random forest regression analysis, R^2 increases by 38.9%, MAPE and RMSE decrease by 26.73% and 48.6%, respectively.

Global regression analysis may be insufficient to model spatially dependent variables in housing studies. After observing spatial variations in the determinants by Global Moran's I, a nominally weighted multiscale spatial clustering method was developed and value regions were created. Thus, the use of local regression analysis for each cluster was enabled so that influence of the geographically dependent factors can be unveiled for those value regions. Using local regression models for nonstationary spatial determinants resulted in better model performance and a more detailed interpretation of the criteria effects on the value.

In this study, mass valuation of residential properties in England and Wales was carried out accurately by developing a hybrid approach through GIS and machine learning integration. Countries or valuation organizations can implement this approach while conducting automated valuation of immovable assets to use as a reference in value-based applications. This hybrid approach can be adapted to any property price data and applied anywhere in the world.

As a future study, the PPD-EPC data linkage process can be rebuilt to include the period from January 2020 to date. In this manner, it will be possible to observe the COVID-19 effects on the U.K. house price market and to improve the price prediction models for recent property transactions. Time series analyses such as ANOVA, LSTM, Transformers methods can also be utilized for both price forecasting and pattern analysis (seasonality, trends). Besides, the effect of view criterion on the condominium value can be revealed for different storey levels by carrying out 3D GIS analyses with CityGML, CityJSON, or BIM data.

Acknowledgements

Contains HM Land Registry data. Crown copyright and database right 2022. PPD is licensed under the Open Government Licence v3.0.

Conflict of interest

No potential conflict of interest was reported by the authors.

Data availability statement

The data and codes that support the findings of this study are openly available in "figshare" at https://doi.org/10.6084/m9.figshare.17711363.v3 .

Appendix

Three tables give information about PPD and EPC attributes and valuation criteria weights.

Table A1. PPD Attributes (Price Paid Data, 2021)

Attribute name	Explanation
Transaction ID	A unique reference number which is generated automatically recording each published sale.
Price	Sale price stated on the transfer deed.
Date of transfer	Date when the sale was completed, as stated on the transfer deed.
Property type	D: Detached, S: Semi-Detached, T: Terraced, F: Flats/Maisonettes, O: Other
Old/new	Indicates the age of the property. Y: A newly built property, N: An established residential building
Duration	Relates to the tenure. F: Freehold, L: Leasehold
Address	Primary Addressable Object Name (PAON), Secondary Addressable Object Name (SAON), Postcode, Street, Locality, Town/City, District, County

Table A2. EPC Data Attributes (EPC, 2022)

Attribute name	Explanation
LMK Key	Individual lodgement identifier.
Building reference number	Unique identifier for the property.
Current energy efficiency	Based on cost of energy, i.e. energy required for space heating, water heating and lighting [in kWh/year] multiplied by fuel costs. ($£/m^2/year$).
Property type	Type of property such as House, Flat, Maisonette.
Built form	Detached, Semi-Detached, Terrace.
Environment impact current	The Environmental Impact Rating. A measure of the property's current impact on the environment in terms of carbon dioxide (CO_2) emissions. The higher the rating the lower the CO_2 emissions (CO_2 emissions in tonnes/year).
Energy consumption current	Current estimated total energy consumption for the property in a 12-month period (kWh/m^2).
CO_2 emissions current	CO_2 emissions per year in tonnes/year.
Lighting cost current	Current estimated annual energy costs for lighting.
Heating cost current	Current estimated annual energy costs for heating.
Hot water cost current	Current estimated annual energy costs for water.
Total floor area	The total useful floor area is the total of all enclosed spaces measured to the internal face of the external walls (m^2).
Floor level	Floor level relative to the lowest level of the property (0 for ground floor).
Extension count	The number of extensions added to the property.
Number habitable rooms	Habitable rooms include any living room, sitting room, dining room, bedroom, study and similar.
Address	Postcode, Address 1 (SAON), Address 2 (PAON), Address 3 (Street).

Table A3. Valuation Criteria that Used to Carry out Spatial Analyses

No	Criterion	Weight	Data set
1	Proximity to Airports	0.02992	OS OpenMap – Local
2	Proximity to Art Centres	0.01048	OS OpenMap – Local
3	Proximity to Bus Stations	0.00864	OS OpenMap – Local
4	Proximity to City Centres	0.00643	OSM City Centres
5	Proximity to Roads (A Road)	0.00250	OS Open Roads
6	Proximity to Roads (B Road)	0.00378	OS Open Roads
7	Proximity to Highway Junctions	0.01802	OS Open Roads
8	Proximity to Coach Stations	0.03681	OS OpenMap – Local
9	Proximity to Ferries	0.00317	OS OpenMap – Local
10	Proximity to Passenger Ferries	0.00250	OS OpenMap – Local

(Continued)

Table A3. *Continued*

No	Criterion	Weight	Data set
11	Proximity to Fire Stations	0.00499	OS OpenMap – Local
12	Proximity to Green Spaces	0.00611	EEA Tree Density Cover
13	Proximity to Hospitals	0.00781	OS OpenMap – Local
14	Proximity to Other Health Centres	0.00772	OS OpenMap – Local
15	Proximity to Libraries	0.00480	OS OpenMap – Local
16	Proximity to Museums	0.04106	OS OpenMap – Local
17	Proximity to Parks	0.02289	OSM Car Parks
18	Proximity to Police Centres	0.00822	OS OpenMap – Local
19	Proximity to Ports	0.01401	OS OpenMap – Local
20	Proximity to Post Offices	0.00630	OS OpenMap – Local
21	Proximity to Primary Education Centres	0.00436	OS OpenMap – Local
22	Proximity to Secondary Education Centres	0.00633	OS OpenMap – Local
23	Proximity to Further Education Centres	0.01125	OS OpenMap – Local
24	Proximity to Universities	0.01125	OS OpenMap – Local
25	Proximity to Malls	0.03709	OSM Shopping Centres
26	Proximity to Sport Centres	0.00767	OS OpenMap – Local
27	Proximity to Subway Stations	0.18522	OSM Railway Stations
28	Proximity to Train Stations	0.01054	OSM Railway Stations
29	Proximity to Tram Stations	0.00416	OSM Railway Stations
30	Proximity to Touristic Attractions	0.00664	OS OpenMap – Local
31	Proximity to Places of Worship	0.00376	OS OpenMap – Local
32	Aspect	0.00175	EU-DEM (v1.1)
33	Slope	0.00780	EU-DEM (v1.1)
34	River View	0.00281	OS Open Rivers
35	Sea View	0.00553	OS Watermark Boundary

References

Alfaro-Navarro, J. L., E. L. Cano, E. Alfaro-Cortés, N. García, M. Gámez, and B. Larraz. (2020). "A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems." *Complexity* 2020, 1–12 Available from: https://www.hindawi.com/journals/complexity/2020/5287263/

Arribas-Bel, D. (2014). "Accidental, Open and Everywhere: Emerging Data Sources for the Understanding of Cities." *Applied Geography* 49(2014), 45–53.

Aydinoglu, A. C., R. Bovkir, and I. Colkesen. (2021). "Implementing a Mass Valuation Application on Interoperable Land Valuation Data Model Designed as an Extension of the National GDI." *Survey Review* 53(379), 349–65. https://doi.org/10.1080/00396265.2020.1771967

Baldominos, A., I. Blanco, A. J. Moreno, R. Iturrarte, Ó. Bernárdez, and C. Afonso. (2018). "Identifying Real Estate Opportunities Using Machine Learning." *Applied Sciences* 8(11), 2321 Available from: http://www.mdpi.com/2076-3417/8/11/2321

Benjamin, J. D., R. S. Guttery, and C. F. Sirmans. (2020). "Mass Appraisal: An Introduction to Multiple Regression Analysis for Real Estate Valuation." *Journal of Real Estate Practice and Education* 7(1), 65–77.

Breiman, L. (2001). "Random Forests." *Machine Learning* 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bühlmann, P. (2012). *Bagging, Boosting and Ensemble Methods* 985–1022. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-21551-3_33

Bünyan Ünel, F., and S. Yalpir. (2019). "Reduction of Mass Appraisal Criteria with PCA and Integration to GIS." *International Journal of Engineering and Geosciences* 4(3), 94–105.

Chen, T., and C. Guestrin. (2016). "XGBoost: A Scalable Tree Boosting System." In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–94. New York, NY: ACM. https://doi.org/10.1145/2939672.2939785

Chi, B., A. Dennett, T. Oléron-Evans, and R. Morphet. (2021a). "A New Attribute-Linked Residential Property Price Dataset for England and Wales, 2011–2019." *UCL Open Environment* 2(7), 1–25. https://doi.org/10.14324/111.444/ucloe.000019

Chi, B., Dennett, A., Oléron-Evans, T., and Morphet, R. (2021b). A New Attribute-Linked Residential Property Price Dataset for England and Wales 2011-2019. Available from: https://reshare.ukdataservice.ac.uk/854942/.

Clark, S. D., and N. Lomax. (2018). "A Mass-Market Appraisal of the English Housing Rental Market Using a Diverse Range of Modelling Techniques." *Journal of Big Data* 5(1), 1–21. https://doi.org/10.1186/s40537-018-0154-3

Demetriou, D. (2017). "A Spatially Based Artificial Neural Network Mass Valuation Model for Land Consolidation." *Environment and Planning B: Urban Analytics and City Science* 44(5), 864–83. https://doi.org/10.1177/0265813516652115

Dimopoulos, T., and A. Moulas. (2016). "A Proposal of a Mass Appraisal System in Greece with CAMA System: Evaluating GWR and MRA techniques in Thessaloniki Municipality." *Open Geosciences* 8(1), 675–93.

Dimopoulos, T., H. Tyralis, N. P. Bakas, and D. Hadjimitsis. (2018). "Accuracy Measurement of Random Forests and Linear Regression for Mass Appraisal Models That Estimate the Prices of Residential Apartments in Nicosia, Cyprus." *Advances in Geosciences* 45, 377–82 Available from: https://adgeo.copernicus.org/articles/45/377/2018/

Dixon, M. F., I. Halperin, and P. Bilokon. (2020). *Machine Learning in Finance* 1406. New York: Springer.

EPC. (2022). *Energy Performance of Buildings Data*. London: Department for Levelling Up, Housing & Communities. Available from:https://epc.opendatacommunities.org/.

EPC Regulations. (2012). *The Energy Performance of Buildings (England and Wales) Regulations (No: 3118)*. London: The National Archives. Available from:https://www.legislation.gov.uk/uksi/2012/3118/contents/made.

Franklin, C., and P. Hane. (1992). "An Introduction to Geographic Information Systems: Linking Maps to Databases [and] Maps for the Rest of US: Affordable and Fun." *Database* 15(2), 12–5.

Fuerst, F., P. McAllister, A. Nanda, and P. Wyatt. (2016). "Energy Performance Ratings and House Prices in Wales: An Empirical Study." *Energy Policy* 92, 20–33.

Gnat, S. (2021). Property Mass Valuation on Small Markets. https://doi.org/10.3390/land10040388.

Groth, S. S., and J. Muntermann. (2009). "Supporting Investment Management Processes with Machine Learning Techniques." In *Wirtschaftsinformatik Proceedings 2009*. *Wirtschafts informatik at AIS Electronic Library (AISeL)*. Atlanta: AIS Electronic Library (AISeL). Available from: http://aisel.aisnet.org/wi2009/107

Hagras, H. (2018). "Toward Human-Understandable, Explainable AI." *Computer* 51(9), 28–36.

Hardy, A., and D. Glew. (2019). "An Analysis of Errors in the Energy Performance Certificate Database." *Energy Policy* 129, 1168–78.

Ho, W. K., B. S. Tang, and S. W. Wong. (2021). "Predicting Property Prices with Machine Learning Algorithms." *Journal of Property Research* 38(1), 48–70. https://doi.org/10.1080/09599916.2020.1832558

Hogge, B. (2016). GOVLAB Open Data's Impact. Available from: https://odimpact.org/case-united-kingdoms-hm-land-registry-price-paid-data.html.

Huang, B., B. Wu, and M. Barry. (2010). "Geographically and Temporally Weighted Regression for Modeling Spatio-Temporal Variation in House Prices." *International Journal of Geographical Information Science* 24(3), 383–401. https://doi.org/10.1080/13658810802672469

Huang, Y., and G. Hewings. (2021). "More Reliable Land Price Index: Is There a Slope Effect?" *Land* 10(3), 261 Available from: https://www.mdpi.com/2073-445X/10/3/261/htm

IAAO. (2013a). *Standard on Mass Appraisal of Real Property*. Kansas: International Association of Assessing Officers.

IAAO. (2013b). *Standards on Ratio Studies*. Kansas: International Association of Assessing Officers.

IAAO. (2018). *Standard on Automated Valuation Models (AVMs)*. Kansas: International Association of Assessing Officers.

IVSC. (2020). *International Valuation Standards*. London: International Valuation Standards Council.

Jahanshiri, E., T. Buyong, and A. R. M. Shariff. (2011). "A Review of Property Mass Valuation Models." *Pertanika Journal of Science & Technology*, 19(1), 23–30.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Y. Liu. (2017). "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." In *Advances in Neural Information Processing Systems (NIPS 2017)* Vol 30, 3146–54. Red Hook, NY: Curran Associates Inc.

Kiel, K. A., and J. E. Zabel. (2008). "Location, Location, Location: The 3L Approach to House Price Determination." *Journal of Housing Economics* 17(2), 175–90 Available from: https://linkinghub.elsevier.com/retrieve/pii/S105113770800003X

Kluyver, T., B. Ragan-Kelley, F. Pérez, B. E. Granger, M. Bussonnier, J. Frederic, K. Kelley, et al. (2016). "Jupyter Notebooks – A Publishing Format for Reproducible Computational Workflows." In *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, 87–90, edited by F. Loizides and B. Schmidt. Amsterdam: IOS Press.

Lee, C. (2022). "Enhancing the performance of a neural network with entity embeddings: an application to real estate valuation." *Journal of Housing and the Built Environment* 37, 1057–1072. https://doi.org/10.1007/s10901-021-09885-2

Li, H., C. A. Calder, and N. Cressie. (2007). "Beyond Moran's I: Testing for Spatial Dependence Based on the Spatial Autoregressive Model." *Geographical Analysis* 39(4), 357–75.

Li, Z. (2022). "Extracting Spatial Effects from Machine Learning Model Using Local Interpretation Method: An Example of SHAP and XGBoost." *Computers, Environment and Urban Systems* 96, 101845 https://linkinghub.elsevier.com/retrieve/pii/S0198971522000898

Lisi, G. (2019). "Property Valuation: The Hedonic Pricing Model-Location and Housing Submarkets." *Journal of Property Investment & Finance* 37(6), 589–96.

Liu, Y., Y. Wu, L. Su, W. Li, and J. Lei. (2021). "Stacking-Based Ensemble Learning Method for House Price Prediction." In *Software Engineering Application in Informatics*, Vol. 232. edited by R. Silhavy, P. Silhavy and Z. Prokopova. Cham: Springer. https://doi.org/10.1007/978-3-030-90318-3_22

Longley, P., G. Higgs, and D. Martin. (1994). "The Predictive Use of GIS to Model Property Valuations." *International Journal of Geographical Information Systems* 8(2), 217–35. https://doi.org/10.1080/02693799408901995

Lundberg, S. M., and S. I. Lee. (2017). "A Unified Approach to Interpreting Model Predictions." In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–77. Red Hook, NY: Curran Associates Inc Available from: https://arxiv.org/abs/1705.07874v2

McKinney, W. (2010). "Data Structures for Statistical Computing in Python." *9th Python in Science Conference (SCIPY 2010)*, 56-61. Austin, Texas, United States. https://doi.org/10.25080/Majora-92bf1922-00a

Mete, M. O., D. Guler, and T. Yomralioglu. (2022). "Towards a 3D Real Estate Valuation Model Using BIM and GIS". In *Innovations in Smart Cities Applications Volume 5. SCA 2021. Lecture Notes in Networks and Systems*, Vol. 393, 945–62, edited by M. Ben Ahmed, A. A. Boudhir, I. R. Karas, V. Jain and S. Mellouli, Cham: Springer. https://doi.org/10.1007/978-3-030-94191-8%_77

Mete, M. O., and T. Yomralioglu. (2019). "Creation of Nominal Asset Value-Based Maps using GIS: A Case Study of Istanbul Beyoglu and Gaziosmanpasa Districts." *GI_Forum 2019* 7(2), 98–112. https://doi.org/10.1553/giscience2019_02_s98

Mete, M. O., and T. Yomralioglu. (2021). "Implementation of Serverless Cloud GIS Platform for Land Valuation." *International Journal of Digital Earth* 14(7), 836–850. https://doi.org/10.1080/17538947.2021.1889056

Mohd, T., N. S. Jamil, N. Johari, L. Abdullah, and S. Masrom. (2020). "An Overview of Real Estate Modelling Techniques for House Price Prediction." In *Charting a Sustainable Future of ASEAN in Business and Social Sciences*, 321–38, edited by N. Kaur and M. Ahmad. Singapore: Springer. https://doi.org/10.1007/978-981-15-3859-9_28

Moran, P. A. (1950). "Notes on Continuous Stochastic Phenomena." *Biometrika* 37(1/2), 17–23.
Open Data Barometer. (2017). *Open Data Barometer*, 4th ed. Washington DC: World Wide Web Foundation. Available from: https://opendatabarometer.org/4thedition/
Open Knowledge International. (2018). *Global Open Data Index - Great Britain*. London: Open Knowledge Foundation. Available from: http://2015.index.okfn.org/place/united-kingdom/
Pagourtzi, E., V. Assimakopoulos, T. Hatzichristos, and N. French. (2003). "Real Estate Appraisal: A Review of Valuation Methods." *Journal of Property Investment & Finance* 21(4), 383–401.
Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. (2011). "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12(Oct), 2825–30.
Peng, Z., Q. Huang, and Y. Han. (2019). "Model Research on Forecast of Second-Hand House Price in Chengdu Based on XGboost Algorithm." In *2019 IEEE 11th International Conference on Advanced Infocomm Technology, ICAIT 2019*, 168–72. Jinan: IEEE.
Peterson, S., and A. Flanagan. (2009). "Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal." *Journal of Real Estate Research* 31(2), 147–64 Available from: https://www.tandfonline.com/action/journalInformation?journalCode=rjer20
Powell-Smith, A. (2018). House Prices by Square Metre in England & Wales. Available from: https://houseprices.anna.ps.
Price Paid Data. (2022). *Price Paid Data*. London: HM Land Registry. Available from: https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads
Prokhorenkova L, Gusev G, Vorobev A, Veronika Dorogush A, and Gulin A. (2017). CatBoost: Unbiased Boosting with Categorical Features. *arXiv preprint*, 6638–48. Available from: https://arxiv.org/abs/1706.09516v5
Renigier-Biłozor, M., S. Źróbek, M. Walacik, R. Borst, R. Grover, and M. d'Amato. (2022). "International Acceptance of Automated Modern Tools Use Must-Have for Sustainable Real Estate Market Development." *Land Use Policy* 113, 105876 Available from: https://linkinghub.elsevier.com/retrieve/pii/S0264837721005998
Tajani, F., P. Morano, and K. Ntalianis. (2018). "Automated Valuation Models for Real Estate Portfolios: A Method for the Value Updates of the Property Assets." *Journal of Property Investment and Finance* 36(4), 324–47.
The European Data Portal. (2021). *The Open Data Maturity (ODM) Report 2021*. Luxembourg: European Union Available from: https://data.europa.eu/sites/default/files/landscaping_insight_report_n7_2021.pdf
Truong, Q., M. Nguyen, H. Dang, and B. Mei. (2020). "Housing Price Prediction via Improved Machine Learning Techniques." *Procedia Computer Science* 174, 433–42.
UN-GGIM. (2019). *Framework for Effective Land Administration (FELA)*. New York, NY: United Nations Global Geospatial Information Management (UN-GGIM).
Van Rossum, G., and F. L. Drake. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.
Wallner, R. (2012). "GIS Measures of Residential Property Views." *Journal of Real Estate Literature* 20(2), 224–5 Available from: https://www.tandfonline.com/doi/full/10.1080/10835547.2014.12090338
Waltert, F., and F. Schläpfer. (2010). "Landscape Amenities and Local Development: A Review of Migration, Regional Economic and Hedonic Pricing Studies." *Ecological Economics* 70(2), 141–52.
Wang, D., and V. J. Li. (2019). "Mass Appraisal Models of Real Estate in the 21st Century: A Systematic Literature Review." *Sustainability* 11(7006), 1–14.
Wang, D., V. J. Li, and H. Yu. (2020). "Mass Appraisal Modeling of Real Estate in Urban Centers by Geographically and Temporally Weighted Regression: A Case Study of Beijing's Core Area." *Land* 9(143), 1–18.
Wang, T., Y. Wang, and M. Liu. (2021). "A Price Prediction Method Based on CatBoost." In *Proceedings - 2021 International Conference on Culture-Oriented Science and Technology, ICCST 2021*, 403–8. Beijing: IEEE.
Wang, Y., and Q. Zhao. (2022). "House Price Prediction Based on Machine Learning: A Case of King County." In *Proceedings of the 7th International Conference on Financial Innovation and Economic Development*, 1547–55. Dordrecht: Atlantis Press. Available from: https://www.atlantis-press.com/proceedings/icfied-22/125971834

Williams, R. E. (1987). "Selling a Geographical Information System to Government Policy Makers." *URISA* 3, 150–6.

World Bank. (1993). *Housing: Enabling Markets to Work*. World Bank, Policy Paper.

Wyatt, P. (2013). *Property Valuation*, 2nd ed. Hoboken, NJ: Wiley-Blackwell. Available from: https://www.wiley.com/en-gb/Property+Valuation%2C+2nd+Edition-p-9781118624685

Wyatt, P. J. (1997). "The Development of a GIS-Based Property Information System for Real Estate Valuation." *International Journal of Geographical Information Science* 11(5), 435–50 Available from: https://www.tandfonline.com/action/journalInformation?journalCode=tgis20

Yalpir, S. (2018). "Enhancement of Parcel Valuation with Adaptive Artificial Neural Network Modeling." *Artificial Intelligence Review* 49(3), 393–405. https://doi.org/10.1007/s10462-016-9531-5

Yamani, S. E., M. Ettarid, and R. Hajji. (2019). "Building Information Modeling Potential for an Enhanced Real Estate Valuation Approach Based on the Hedonic Method." In *Building Information Modelling (BIM) in Design, Construction and Operations III* Vol 1, 305–16. Southampton, U.K.: WIT Press.

Yilmazer, S., and S. Kocaman. (2020). "A Mass Appraisal Assessment Study Using Machine Learning Based on Multiple Regression and Random Forest." *Land Use Policy* 99(104889), 1–11. https://doi.org/10.1016/j.landusepol.2020.104889

Yomralioglu, T. (1993). *A Nominal Asset Value-Based Approach For Land Readjustment and Its Implementation Using Geographical Information Systems*. Ph.D Thesis. University of Newcastle upon Tyne.

Yomralioglu, T., and Nisanci, R., (2004). *Nominal Asset Land Valuation Technique by GIS*. *FIG Working Week 2004*, Athens, Greece. FIG. Available from: https://www.fig.net/resources/proceedings/fig_proceedings/athens/papers/ts27/TS27_4_Yomralioglu_Nisanci.pdf

Yu, S. M., S. S. Han, and C. H. Chai. (2007). "Modeling the Value of View in High-Rise Apartments: A 3D GIS Approach." *Environment and Planning B: Planning and Design* 34(1), 139–53. https://doi.org/10.1068/b32116

Zurada, J., A. Levitan, and J. Guan. (2011). "A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context." *Journal of Real Estate Research* 33(3), 349–88. https://doi.org/10.1080/10835547.2011.12091311