GIS & Machine Learning based Mass Appraisal of Residential Properties in England & Wales

Mete MO[*1], Yomralioglu T[†1]

[1]Department of Geomatics Engineering, Istanbul Technical University, Istanbul, Turkey

January 10, 2022

Summary

Property value is needed in many transactions such as purchase and sale, taxation, expropriation, capital market activities. Geographic Information Systems (GIS) and Machine Learning methods are widely used in automated mass appraisal practices. Analysing the Machine Learning based mass appraisal applications, it is seen that locational criteria are insufficiently used during the price prediction process. Whereas, locational criteria like proximity to important places, sea or green areas views, smooth topography are some of the spatial factors that extremely affect the property value. In this study, a hybrid approach is developed by integrating GIS and Machine Learning for mass appraisal of residential properties in England and Wales.

KEYWORDS: real estate valuation; property valuation; mass appraisal; GIS, Machine Learning

1. Introduction

Open data phenomenon has increased substantially with the determinant political will of countries recently (Open Data Barometer, 2017). 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. United Kingdom has made numerous data openly accessible on several topics through a data portal since 2010. HM Land Registry - Price Paid Data (PPD) is an open governmental data that contains sale records and several physical attributes of the residential properties in England and Wales. On the other hand, Energy Performance Certificates (EPC) data is also an 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.

Although PPD dataset contains full address information, it does not include a unique address identifier or coordinates to match with each other and with other datasets. There are several studies which developed different address matching methods for linking the two data (Fuerst *et al.*, 2016; Powell-Smith, 2018; Chi *et al.*, 2021). (Chi *et al.*, 2021) published the linked PPD-EPC dataset 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.

The objective of this research is to create a mass property appraisal model with GIS and Machine Learning hybrid approach. In this sense, the study covers the use of GIS-based Nominal Valuation Method for creating a land value map based on scientific and objective evaluation. Besides, it compares performances of various Machine Learning regression models for mass appraisal 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 contribution of the spatial features on the prediction accuracy.

[*] metemu@itu.edu.tr
[†] tahsin@itu.edu.tr

2. Material and Methods

In this study, we used linked PPD-EPC dataset to carry out mass property appraisal in England and Wales through GIS and Machine Learning integration (**Figure 1**).

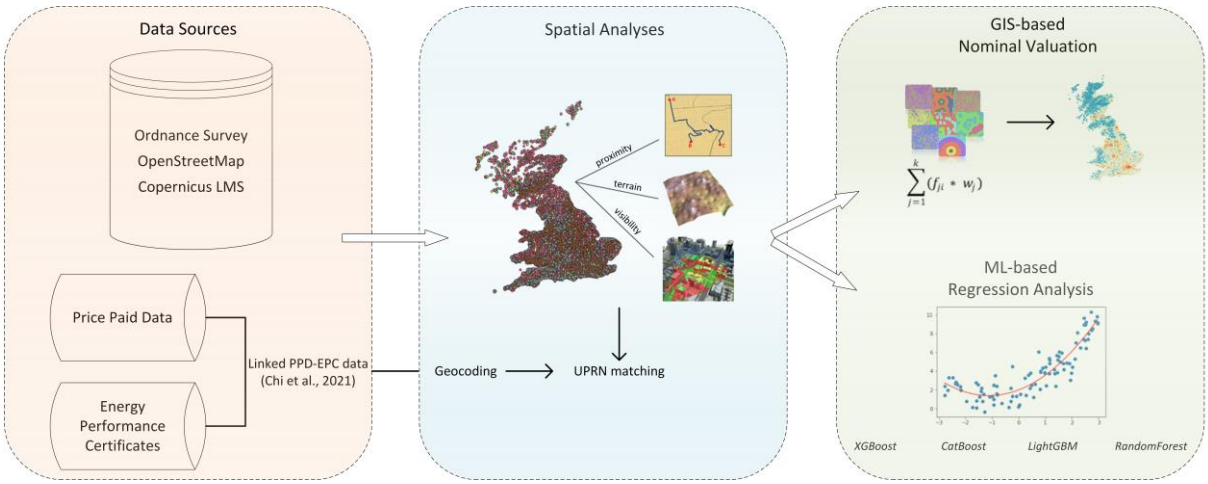

Figure 1 Workflow diagram of the study.

Using a property price dataset in GIS and Machine Learning analysis for valuation purpose requires to have precise location information such as coordinates or unique address identifier. PPD and EPC data provide informative attributes about residential properties located in England and Wales. Although linked PPD-EPC dataset contains address and LSOA information, it does not include coordinate information or unique address identifier. In order to match PPD-EPC data with the UPRN, property addresses are geocoded using HERE Maps API with 99,8% success.

After having the geocoded PPD-EPC data, each property is assigned to the nearest UPRN point by using spatial join (Join Attributes by Nearest) operation in QGIS. In order to match the property with correct UPRN point, maximum nearest neighbours specified as 1 and maximum distance limited to 50 metres. As a result, PPD-EPC data with UPRN attribute created in GeoPackage (GPKG) open and platform independent GIS data format.

Land value maps provide an understanding of price variations in a region and facilitate management, planning, and taxation activities. Using aforementioned open data sources, land value map of Great Britain is produced with GIS-based Nominal Valuation method (Yomralioglu, 1993). Firstly, a valuation database is created on Amazon Aurora PostgreSQL Serverless (Mete and Yomralioglu, 2021) 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 is used. Hence, GIS-based land value map of Great Britain is produced by using Nominal Valuation method (**Figure 2**).

Regression analysis is one of the fundamental algorithms of the Machine Learning and it is widely used for property price prediction. In this study, several ensemble regression methods like XGBoost, CatBoost, LightGBM, Random Forest are built for price prediction on PPD-EPC dataset. In order to measure the model performances, several accuracy metrics are used such as R^2, Adjusted R^2, Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Error (RMSE) (**Table 1**).

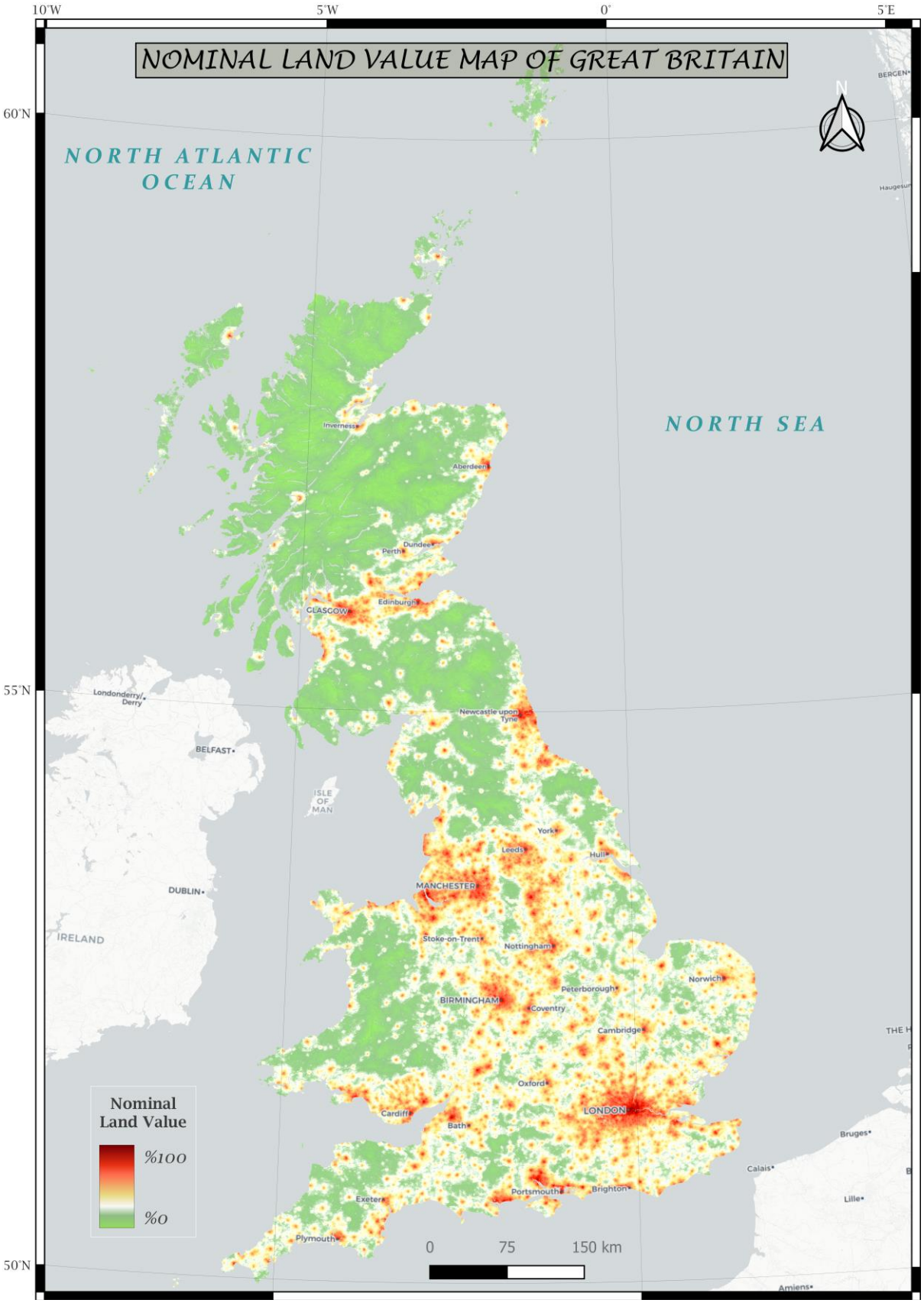

Figure 2 Nominal land value map of Great Britain, UK.

Table 1 Accuracy metrics of the regression models.

Model	Wall Time (minute)	R^2	Adjusted R^2	MAE	MAPE	RMSE
XGBoost	13.78	0.4341	0.4340	98505.9861	0.4689	204625.5031
CatBoost	4.70	0.4551	0.4550	98811.1141	0.4737	200791.2485
LightGBM	0.75	0.4366	0.4365	101291.7355	0.4937	204170.2015
RandomForest	73.90	0.4690	0.4689	95934.5759	0.4614	198211.9713

PPD-EPC data contains many physical attributes about properties which are beneficial for price prediction in Machine Learning based appraisal process. However, there is a need for inclusion of locational factors, which are highly correlated with land price, in order to build more accurate regression models (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. After having the enriched data, the same workflow is followed for building the Machine Learning regression models. **Table 2** shows training execution times and accuracy metrics of the models.

Table 2 Accuracy metrics of the regression models after feature enrichment.

Model	Wall Time (minute)	R^2	Adjusted R^2	MAE	MAPE	RMSE
XGBoost	10.52	0.7907	0.7906	65367.1736	0.3107	123585.9848
CatBoost	13.50	0.8380	0.8379	56668.9762	0.2660	108740.2277
LightGBM	1.47	0.8058	0.8058	65567.8885	0.3188	119042.0076
RandomForest	203.78	0.8579	0.8578	44888.7285	0.1941	101847.7448

Feature importance is a part of feature selection and it is used to interpret the model for better understanding how much a feature effected the prediction process. Thus, permutation importance's are calculated for Random Forest model (**Figure 3**).

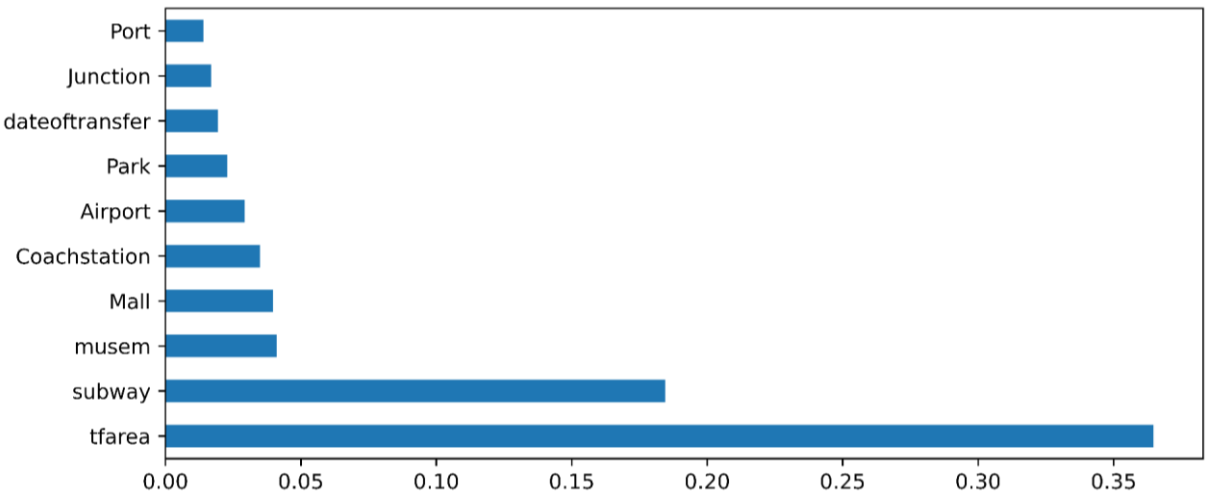

Figure 3 Feature importance scores after feature enrichment.

3. Conclusion

Adoption of mass appraisal techniques in value-based activities has been increasing in recent years. Hedonic Pricing, Nominal Valuation, Spatial Analysis, and Regression Analysis are some of the prominent methods of the mass appraisal practices. Although there are numerous studies using Machine Learning algorithms for mass property appraisal, it has been observed that the spatial factors that extremely affect the land value are neglected. In this study, GIS-based Nominal Valuation method and Ensemble Machine Learning regression algorithms were used for mass appraisal of residential properties in England and Wales. According to the prediction results, Random Forest algorithm was the most successful regression model. 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 for anywhere in the world.

4. Acknowledgements

This work was supported by Scientific Research Projects Coordination Unit of Istanbul Technical University [Grant No. MDK-2021-43080].

References

Chi, B. *et al.* (2021) 'A new attribute-linked residential property price dataset for England and Wales, 2011–2019', *UCL Open Environment*, 2(07). doi: 10.14324/111.444/UCLOE.000019.

Fuerst, F. *et al.* (2016) 'Energy performance ratings and house prices in Wales: An empirical study', *Energy Policy*, 92, pp. 20–33. doi: 10.1016/J.ENPOL.2016.01.024.

Kiel, K. A. and Zabel, J. E. (2008) 'Location, location, location: The 3L Approach to house price determination', *Journal of Housing Economics*, 17(2), pp. 175–190. doi: 10.1016/j.jhe.2007.12.002.

Mete, M. O. and Yomralioglu, T. (2019) 'Creation of Nominal Asset Value-Based Maps using GIS: A Case Study of Istanbul Beyoglu and Gaziosmanpasa Districts', *GI_Forum 2019*, 7(2), pp. 98–112. doi: 10.1553/giscience2019_02_s98.

Mete, M. O. and Yomralioglu, T. (2021) 'Implementation of serverless cloud GIS platform for land valuation', *International Journal of Digital Earth*, 14(7), pp. 836–850. doi: 10.1080/17538947.2021.1889056.

Open Data Barometer (2017) *Open Data Barometer 4th Edition*, *World Wide Web Foundation*. Available at: https://opendatabarometer.org/4thedition/ (Accessed: 20 December 2021).

Powell-Smith, A. (2018) *House prices by square metre in England & Wales*. Available at: https://houseprices.anna.ps (Accessed: 12 December 2021).

The European Data Portal (2021) *The Open Data Maturity (ODM) Report 2021*. Luxembourg. Available at: https://data.europa.eu/sites/default/files/landscaping_insight_report_n7_2021.pdf.

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), pp. 435–450. doi: 10.1080/136588197242248.

Yomralioglu, T. (1993) *A Nominal Asset Value-Based Approach For Land Readjustment And Its Implementation Using Geographical Information Systems* (PhD thesis). University of Newcastle upon Tyne, England.

Biographies

Muhammed Oguzhan Mete is PhD student and Research Assistant in Geomatics Engineering Department, Istanbul Technical University, Turkey. His research interests are Geographic Information Systems (GIS), machine learning, big data analytics, land management, and real estate valuation.

Tahsin Yomralioglu is Professor in Geomatics Engineering Department, Istanbul Technical University, Turkey. In 1993, he obtained his PhD from the University of Newcastle upon Tyne, England. His research interests include geographic information technology (GIT), spatial data infrastructure policies, land management, and real estate valuation.