Author response to reviews of

Manuscript Paper 147

Urban and socio-economic correlates of property price evolution: Application to Dublin's costal area

submitted to *Special Session-Environmental and Geo-spatial Data Analytics (EnGeoData) of the 7th IEEE International Conference on Data Science and Advanced Analytics (DSAA'2020).*

RC: ***Reviewer Comment*** AR: Author Response ☐ Manuscript text

Dear Antonio, David, Mathieu and Maguelonne,

Thank you to consider the manuscript entitled "Urban and socio-economic correlates of property price evolution: Application to Dublin's costal area" in EnGeoData 2020. We would also like to take this opportunity to express our thanks to the reviewers for the positive feedback and helpful comments for correction or modification.

We believe have resulted in an improved revised manuscript, which you will find uploaded alongside this document. The manuscript has been revised to address the reviewer comments, which are appended alongside our responses to this letter.

1. Reviewer #1

RC: **Urban and socio-economic correlates of property price evolution: Application to Dublin's costal area**

Hypothesis : By taking into account either urban or socio-economic features, it is possible to accurately estimate housing prices and to predict their evolution.

Proposal : By using machine learning algorithms, prediction of house prices has become more and more accurate. In this paper, an alternative machine learning algorithm is used to identify urban and socio-economic correlates of property price evolution in the area of Dublin, Ireland in 2018.

AR: Many thanks for your review and for your comments, we will address each of them as indicated in our answers.

RC: **Page 1 : « 10389 houses » in the Abstract and « 10395 properties » in the Method section and « 10387 Properties » in Discussion section ?**

AR: The figures have been corrected with the appropriate number of observation in our data set (10387 properties). In addition the terms house/s has been replaced by property/ies where applicable.

RC: **Page 2 : « an eXtreme Gradient Boosting (XGBoost) regression model was calculated to examine the extent to which house prices rely on urban and on socio-economic features ». Could you explain wether you worked on the XGBoost algorithm or not ? What was your contribution to develop/adapt/or**

parameterize this algorithm ?

AR: No modifications of the current version of the XGBoost R-package algorithm has been applied to obtain our results. The hyperparameters are all defaults as precised in the manuscript:

> As no specific weight or bias in the models have been hypothesized, parameters of the XGBoost algorithms have been kept as default values.

RC: I suppose that Urban features importance is computed by the XGBoost model (Table I). Could you explain this stage and comment results ? Embassy importance is detected thanks to statistical computations. Is it possible to comment and validate it by qualitative interviews with experts ? Page 4 : Similarly, I couldn't understand how Socio-Economic features importance is computed (Table II). Could you explain this stage and comment results ? Religion importance is detected thanks to statistical computations. Is it possible to comment and validate it by qualitative interviews with experts ?

AR: Indeed the importance is obtained from the XGBoost model. It is a measure of the improvement in accuracy brought by a feature based on its coefficient's absolute magnitude. Compared to the other features included, the importance a specific feature corresponds to the proportion of its contribution to the overall accuracy of the model. This information as been added to the "eXtreme Gradient Boosting" section as follow:

> Next, the most accurate model is used to identify the features that are contributing most to the prediction of property prices. This contribution (also called "importance") is a measure of the improvement in accuracy brought by a feature. It is based on the absolute magnitude of feature's coefficient.

Our explanations on the importance of certain features to the model are interpretative and, therefore, we would like to keep these interpretations in the discussion section. We agree that it would be easier for the reader to find them right after each result section, however, we think that it is important to keep separated objective results and subjective interpretations.

The idea of qualitative interviews is very interesting. However, we would like to keep this submission as a quantitative study. Interviews will be considered in longer manuscripts.

RC: « Embassies are in general located in the most expensive areas of cities ». Are the ambassies installed in expensive areas or are the areas expensive because of their embassies ?

AR: The question of correlation vs. causation is highly relevant. Because linear models are based on correlations, we want to avoid any interpretation related to causation. As a consequence, the following sentence has been added to the document discussion:

> In addition, while some features appear to be more important than others in their contribution to property prices, these results are not causal. Thus, a feature's importance can be either the cause or the consequence of property prices in a neighborhood.

RC: Globally, it is difficult to evaluate the added-value of this research work :

- Is it a proposal of a new regression model ? If yes, you should better explain XGBoost and its originality

- Is it a experiment of an existing regression model applied to housing prices ? If yes, you should comment the results on the 20% of the original dataset : are the price predictions relevant ? Is this approach better than existing ones for price prediction ? Number of parameters ? Computation speed ? etc.

What was not possible before and what is possible now thanks to this work ?

AR: As indicated in the title, the added value of the paper is about identifying urban and socio-economic correlates of property price evolution. These correlated are obtained on the results of the test dataset. The price predictions is relevant to real estate agents in charge of property valuations and to buyers to approximate the real value of a property. This added value has been added in the conclusion as follow:

> These results allow an understanding of why some areas have higher prices than others which is relevant information not only for real estate agents in charge of property valuations but also for buyers in order to estimate the real value of a property.

RC: Typos : Page 1 : « 10395 property »

AR: Many thanks for identifying this typo. It has been corrected in the new submission of the manuscript.

2. Reviewer #2

RC: This paper presents machine learning experiments to estimate price changes in the real estate market.

The main method is the XGBoost algorithm. Perhaps more methods could have been tried, including some simple baseline to put the presented results into perspective.

The focus of the paper in on the contribution of the proposed features.

There in not much novelty, but the experiments are well executed and the results are convincing. Real data is used for training and testing.

AR: Thank you for your review and comments. In order to compare the accuracy of different variations of the XGBoost algorithm, the result of 8 models are now presented table I and III. The models are comparing XGBoost algorithms using tree based or linear regression booster. In addition each booster type is calculated according different learning objective output: squared error, squared log error, gamma, and tweedie.

3. Reviewer #3

RC: The topic of the paper fits well with the special session main topics. Below some suggestions of improvement that can be made to this contribution.

AR: We thank you for your thoughtful and thorough review and we believe your input has greatly improved the manuscript.

RC: Concerning the following statement: "techniques used usually include artificial neural networks due to the volatility of the market", even though it is backed up by bibliographic references, it would be

worth further explaining it. The same applies to the following statement "While these methods are efficient in a non-restricted space, they have limitations when they are used in coastal areas.", explain why.

AR: A precision on the use of artificial neural networks techniques has been added to the manuscript as follow:

> Indeed, artificial neural network techniques can quickly adjust to volatility changes, therefore incorporating sharp increases or decreases in their predictions (Aragonés, Blanco, & Estévez, 2007). However, the model obtained from most artificial neural networks are difficult to interpret (Rudin, 2019).

The limitation of classic models applied to coastal areas is presented as follow:

> This problem is known as "finite area smoothing" and occurs when predictions from a model are approximated across geographical barriers, such as irregular coastal shapes, which can lead to poorly predicted values (Ramsay, 2002).

RC: There are some issues related to the data acquisition process and the possible presence of errors. You noted that errors may occur when the data is being filed, could you provide some calculated information about the quality of the data used based on missing values, incoherent values, etc. Also, property addresses were geocoded relying on OSM geocoder, can you estimate accuracy of the process? The quality of OSM data is usually weak, what are the sources reported on the OSM urban features you extracted (source key: https://wiki.openstreetmap.org/wiki/Key:source)?

AR: In 2018, the Property Price Register references 18665 properties sold in Dublin's area. Among these properties, those corresponding to apartments have been removed from the original dataset in order to evaluate only houses (10.3% of the properties). Finally, some properties have been unsuccessfully geocoded because of mistakes on their address or because they have not been found (38.0% of the properties). These properties have also been removed from the original dataset to obtain a final dataset of 10387 properties.

These information have been added as it in the new version of the manuscript.

Urban features are obtained using the R package osmdata which queries the Overpass API. This information as also been added to the manuscript.

RC: Input data is not available in the github repository given by the authors, can we suppose it is due to restrictions to share the data, if so, this should be said clearly.

AR: The github repository has been update to ensure the full reproducibility of the results. The geocoded dataset has been added as well as the scripts to obtain the batch geocoding of the properties, to access to the Overpass API for the distance calculation with the urban features and the links to download and process the census data.

RC: Add URL to point to the Central Statistics Office and All-Island Research Observatory where data is available.

AR: the URLs for the Central Statistics Office and the All-Island Research Observatory have been added to the manuscript.

The results of Irish 2016 census consultation is accessible through both the Central Statistics Office (Central Statistics Office, 2020) and the All-Island Research Observatory (All-Island Research Observatory, 2020) websites.

RC: explicit acronym "SD".

AR: The acronym SD has been explicitly introduced in the manuscript.

References

All-Island Research Observatory. (2020, June 1). AIRO data store. Retrieved from `http://airo.maynoothuniversity.ie/datastore`

Aragonés, J. R., Blanco, C., & Estévez, P. G. (2007). Neural network volatility forecasts. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *15*(3-4), 107–121.

Central Statistics Office. (2020, June 1). Census 2016 small area population statistics. Retrieved from `https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/`

Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(2), 307–319.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.

All-Island Research Observatory. (2020, June 1). AIRO data store. Retrieved from `http://airo.maynoothuniversity.ie/datastore`

Aragonés, J. R., Blanco, C., & Estévez, P. G. (2007). Neural network volatility forecasts. *Intelligent Systems in Accounting, Finance & Management: International Journal*, *15*(3-4), 107–121.

Central Statistics Office. (2020, June 1). Census 2016 small area population statistics. Retrieved from `https://www.cso.ie/en/census/census2016reports/census2016smallareapopulationstatistics/`

Ramsay, T. (2002). Spline smoothing over difficult regions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(2), 307–319.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.