Mass appraisal of urban and rural land values using random forest with spatial restriction

Córdoba M[13*], Monzani F[2], Carranza J[23], Piumetto M[23], Balzarini M[13].

[1]National Research Council of Science and Technology (CONICET), [2] Spatial Data Infrastructure of Córdoba, [3]National University of Córdoba, Córdoba, Argentina.

Introduction

The advancement of computational software and machine learning practice has facilitated enhanced uptake of mass appraisal methodologies for price modelling and prediction of land value. The statistical approach to MA allows the price assessment of a group of properties at a given date using sampling data and standard methods, to predict prices in non-observed properties. Computer-Assisted Mass Appraisal (CAMA) has become popular as an automated valuation models (AVM) (Zhang et al., 2015). An AVM model produces an estimate of the land value based on market analysis of location, market conditions, and real estate characteristics previously collected and used as predictors (IAAO, 2017). The linear regression (LR) model is the most popular statistical method used to develop a predictive model to be employed in AVM (Demetriou, 2017). Alternatively, regression models can be adjusted from regression trees algorithm (Breiman et al., 1984) or an extension of these to enhance prediction from resampling as the Random Forest (RF) (Breiman, 2001) or the Quantile Random Forest (QRF) (Meinshausen, 2006) algorithms. These algorithms were recognized as machine learning technique for real estate mass appraisal. However, they ignore influences (spatial autocorrelation) of neighboring observed data when predicting the value of interest at a given site. To overcome the disadvantage, random forest plus kriging of residuals (sRF) method can be used. Initially, a RF of land values using predictive ancillary variables is carried out in order to model the trend component. In the second step, ordinary kriging is applied to the residuals of RF and a spatial prediction of the residuals is created. The final prediction is an additive combination of both model steps. The aim of this study was to compare performances of RF and quantile QRF both with and without spatial restriction in the prediction of rural and urban land values.

Material and Methods

Data

We use three datasets of 122, 264 and 3718 market data, released between 2017 and 2018. The first two contains data of urban land coming from two cities (denoted Urban 1 and urban 2, respectively) in the Province of Córdoba. The third involves data of rural land value for the whole Province of Córdoba,

[*] *Corresponding author. E-mail address: marianoacba@agro.unc.edu.ar*

Argentina. In addition to the samples of land values, based on the market values of urban vacant lots, a set of variables was generated that can be grouped into three different sets according to their characteristics. a) Variables of distances with respect to point of interest (closed neighborhoods, popular neighborhoods, city center, parks and green areas, bus terminal, university, large commercial areas, industrial parks, railways, main roads, and depreciation areas), b) Services and Infrastructure variables, c) Surroundings Variables (proportion of vacant urban, square meters built, number of real estate transactions carried out during the last year). Indicator variables of land cover and use, use capacity and soil productivity index, variables related to climate, topography, drought hydrology, infrastructure, and data of belonging or environment were used for the modeling of rural land. Special consideration was taken with native forest areas and flooded areas with high recurrence. Each entry in the database contains the market unit value of land (LUV), the value of each of the variables described above and the respective coordinates. The information used in this article was generated within the framework of the Real Estate Territorial Study (ETI) of Province of Córdoba, Argentina (Piumetto et al., 2019).

Statistical process

RF and QRF models were trained using the "caret" library of R software. The number of variables to consider was evaluated for splitting at every node (mtry) using a grid of mtry values. The minimum number of terminal nodes and the number of trees was held constant in 5 and 1000, respectively. The performance of this algorithm under a specific tune parameter was evaluated using a k-fold cross-validation algorithm (k=10). As validation measurements in this step we use the root mean square error (RMSE). Then, residuals from the fitted models were interpolated to prediction grids using ordinary kriging, and the interpolated residuals were added to the RF and QRF prediction results for obtaining the spatial predictions of RF (sRF) and QRF (sQRF). The model performances for making predictions at the unvisited was summarized through the same cross-validation procedure above described to derive prediction error and RMSE. Another way to assess the quality of a model applied on spatial data is that degree of residual spatial autocorrelation (RSA). We employ the commonly used Moran's I Index (MI) (Moran, 1948) to assess the level of residual autocorrelation in incrementally increasing distance ranges from the compared methods.

Results

Model performances of the fitted models do not vary widely in terms of RMSE (Table 1). The models based on RF had better performance. Only in the urban land the methods that incorporate spatial information performed better.

Table 1. Root mean square error (RMSE, %) of two regression tree models used for mass appraisal of urban and rural land values.

Dataset	Quantile Regression Forest	Random Forest
Urban 1	40 (38)[†]	37 (36)
Urban 2	34 (33)	34 (31)
Rural	35 (35)	34 (34)

[†]In parenthesis with and without spatial restriction

Independent of the type of modelling approach, residuals exhibit lower MI values (Fig. 1). In rural dataset the MI is close to zero for all the distances. RF presented lower MI values in all dataset. The residuals from models without spatial restriction (RF and QRF) showed higher RSA than spatially versions of these methods (sRF and sQRF). The greatest difference in RSA is observed between RF and sRF for Urban 2. For this data set, a greater level difference of the RMSE was also found.

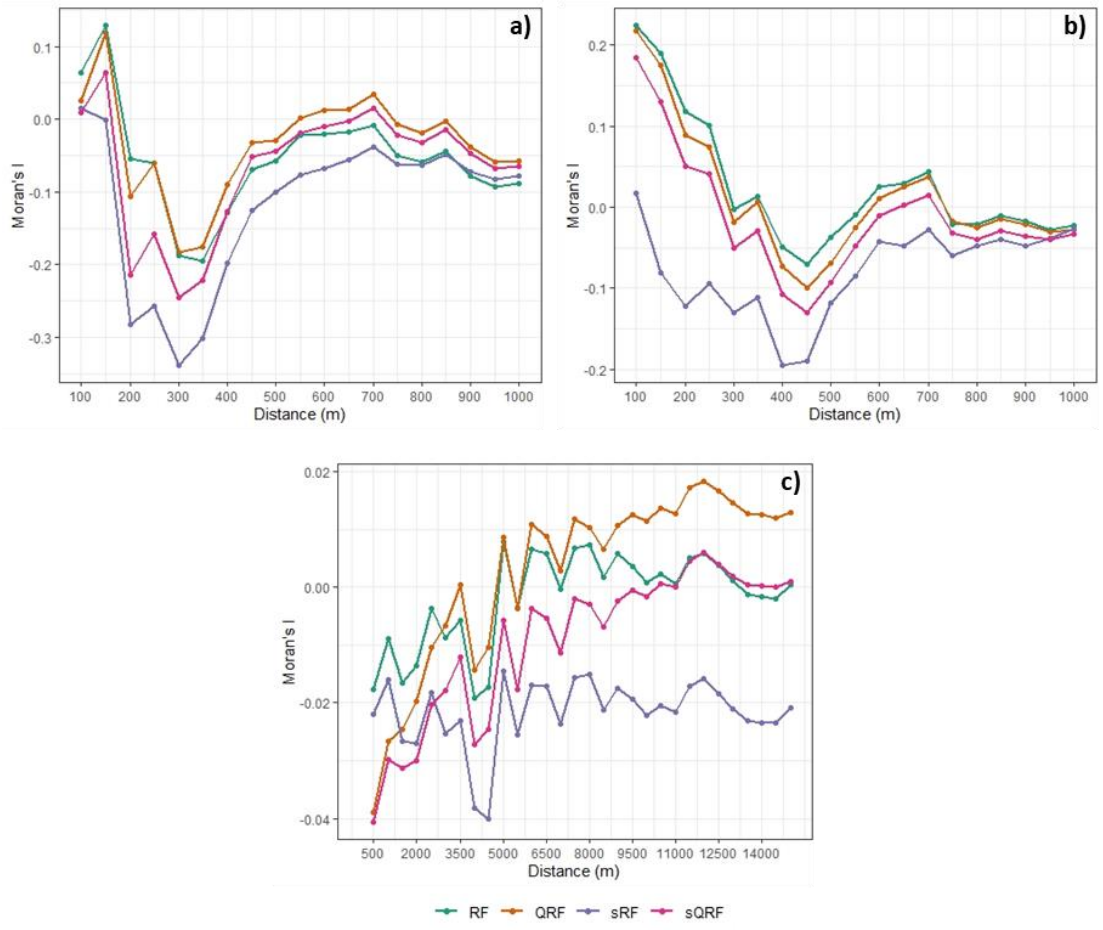

Fig. 1. Moran's I Index at incrementing spatial distances of residuals obtained form random forest and Quantile Regression Forest without (RF and QRF) and with spatial restriction (sRF and sQRF) fitted in thrree dataset, a) Urban 1, b) Urban 2, c) Rural.

Conclusion

The incorporation of spatial restriction slightly improved the performance of regression tree models used for mass appraisal of land values.

References

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and regression trees. CRC press.

Demetriou, D., 2017. A spatially based artificial neural network mass valuation model for land consolidation. Environ. Plan. B Urban Anal. City Sci. 44, 864–883.

IAAO, 2017. Standard on Mass Appraisal of Real Property (SMARP). Kansas.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Piumetto, M., Garcìa, G., Monayar, V., Carranza, J., Morales, H., Nasjleti, T., Menéndez, A., 2019. Técnicas algorítmicas y Machine Learning para la Valuación Masiva de la Tierra de la provincia de Córdoba. Rev. la Fac. Ciencias Exactas, Físicas y Nat. 6, 49–52.

Zhang, R., Du, Q., Geng, J., Liu, B., Huang, Y., 2015. An improved spatial error model for the mass appraisal of commercial real estate based on spatial analysis : Shenzhen as a case study. Habitat Int. 46, 196–205.