Performance Analysis of Machine Learning Algorithms for Regression of Spatial Variables. A Case Study in the Real Estate Industry

Sebastian F. Santibanez[1], Marius Kloft[2] , Tobia Lakes[3]

[1]Humboldt University of Berlin / Department of Geography | Urban4M; 135 San Lorenzo ave, suite 530, Coral Gables Fl 33143, Telephone: +1 202 604 6524
Email: sebastian.santibanez@geo.hu-berlin.de

[2]Humboldt University of Berlin / Department of Computer Sciences; Unter den Linden 6, 10099, Berlin, Germany
Email: kloft@hu-berlin.de

[3]Humboldt University of Berlin / Department of Geography, Unter den Linden 6, 10099, Berlin, Germany
Email: tobia.lakes@geo.hu-berlin.de

Abstract

Machine learning is a computational technology widely used in regression and classification tasks. One of the drawbacks of its use in the analysis of spatial variables is that machine learning algorithms are in general, not designed to deal with spatially autocorrelated data. This often causes the residuals to exhibit clustering, in clear violation of the condition of independent and identically distributed random variables. In this work we analyze the performance of some well-established machine learning algorithms and one spatial algorithm in the prediction of the average rent price of certain real estate units in the Miami-Fort Lauderdale-West Palm Beach metropolitan area in Florida, USA. We defined "performance" as the goodness of fit achieved by an algorithm in conjunction with the degree of spatial association of the residuals. We identified significant differences between the machine learning algorithms in their sensitivity to spatial autocorrelation and the achieved goodness of fit. We also exposed the superiority of machine learning algorithms over generalized least squares in both goodness of fit and residual spatial autocorrelation. Finally we show preliminary evidence that blending ensemble learning can be used to optimize a regression problem. Our findings can be useful in designing a strategy for regression of spatial variables.

1. Introduction

Machine learning algorithms (MLAs) are widely used in regression and classification tasks. In recent years they have gained interest in spatial applications such as classification of land-cover or regression analysis of complex spatial datasets (Okujeni 2014, Cheong 2014, Cracknell 2014, Berwald 2012, Li 2011, Lakes 2009).

Applications in real estate are often confronted with nonlinearity in the data (Mu 2014), non-stationarity in the response variable (Bork 2015), along with other unwanted effects. Traditional statistical regression is limited in such cases while MLAs may offer benefits. However, one of the drawbacks when using MLAs in the analysis of spatial variables is that MLAs are not designed to deal with spatially autocorrelated (SAC) data. Frequently, this results in clustering of the residuals which is a clear violation of the condition of independent and identically distributed variables.

Despite the growing popularity of MLAs in spatial analysis, there has been an insufficient number of studies comparing the adequateness of ML algorithms for regression of spatial variables. Some studies that do address this type of comparison are Santibanez (2015) and Li (2011).

This study is part of the "Machine Learning and Spatial Analysis" project developed jointly by the Humboldt University of Berlin and Urban4M LLC, which seeks to develop a framework for the use of Machine Learning tools in different aspects of spatial analysis.

2. Materials and Methods

We present a performance comparison of selected well-stablished MLAs and one spatial method for the regression of real estate data.

2.1 Data

Our response variable is the median rent price per zip code of a two bedroom two bathroom apartment in the Miami-Fort Lauderdale-West Palm Beach metropolitan area in Florida, USA. The response variable is summarized as follows: Min ($942), 1st Quartile ($1206), Median ($1346), Mean ($1456), 3rd Quartile ($1596), Max ($3038).

Our explanatory variables is a set of 23 demographic indicators on age, marital status, education level and income extracted from the 3-year estimates of the American Community Survey 2013 (United States Census). These variables were interpolated to zip code level using the methodology proposed in Shepard (1997).

2.2 Algorithms

We tested the following well-known MLAs: Random Forest (Breiman 2001), Neural Network (Ripley 1996), Neural Network with PCA (Ripley 1996), Cubist (Extension on M'5 algorithm by Quinlan 1992), Partial Least Squares (Buphinder 1998), Support Vector Machines with Linear Kernel (Vapnik 1998), Support Vector Machines with Radial Basis Function Kernel (Vapnik 1998) and Gradient Boosting Machine (Freund 1997). Also, we included the spatial algorithm Generalized Least Squares (Browne 1974) for comparison purposes.

2.3 Methods

The ML algorithms were implemented via "Caret" in "R". The parameter tuning was done with respect to the calculated RMSE via standard grid search using repeated 5-fold cross-validation. The estimated response was also obtained via repeated 5-fold cross validation. GLS was implemented via "nlme" in R. The estimated GLS response was calculated via repeated 5-fold cross validation using "cvTools".

We then calculated the Root Mean Square Error, R^2 and Moran's I statistic of the results delivered by each algorithm to gauge their performance.

Finally, as a first step to assess the possibility of optimizing the results of the regression via ensemble learning, we averaged the solution of two of the strongest algorithms in both RMSE and residual SAC.

3. Results

All chosen explanatory variables present multicollinearity. The Variance Inflation Factor of all explanatory variables can be summarized as follow: VIF = Min(5.1), Median(24.5),

Mean(93.4), Max(756). Figure 1 shows graphically the correlation between explanatory variables. Also, all chosen explanatory variables present low to moderate SAC ($p<0.01$, Moran's I statistics = min 0.210, mean 0.278, max 0.333).

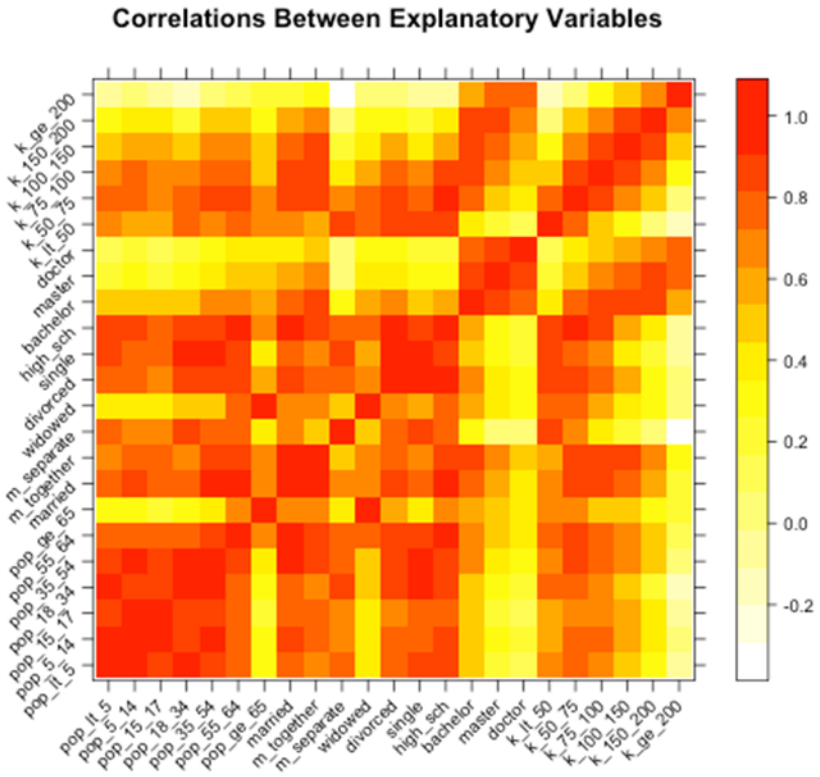

Figure 1. Correlation between explanatory variables

The results of RMSE, R^2, and residual SAC are presented at the end of this document, in figures 2, 3 and 4 respectively.

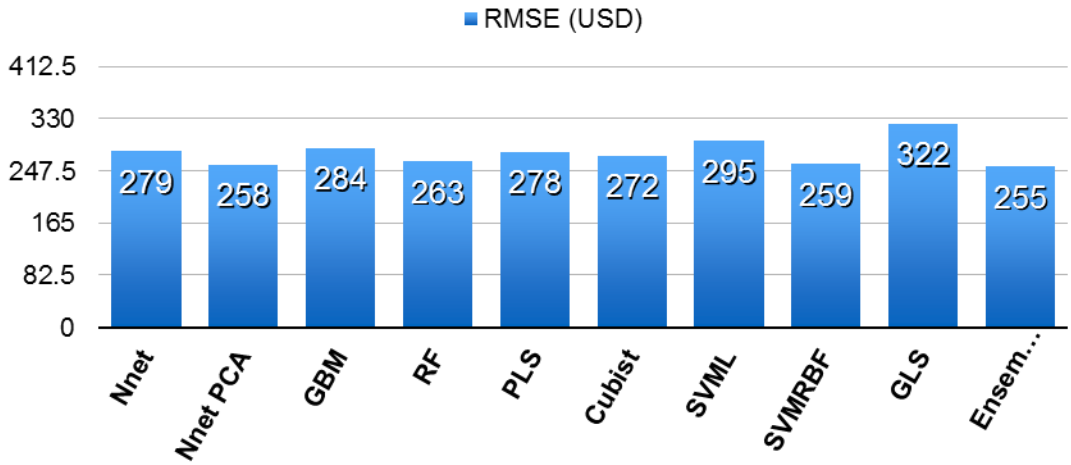

Figure 2: RMSE, all algorithms.

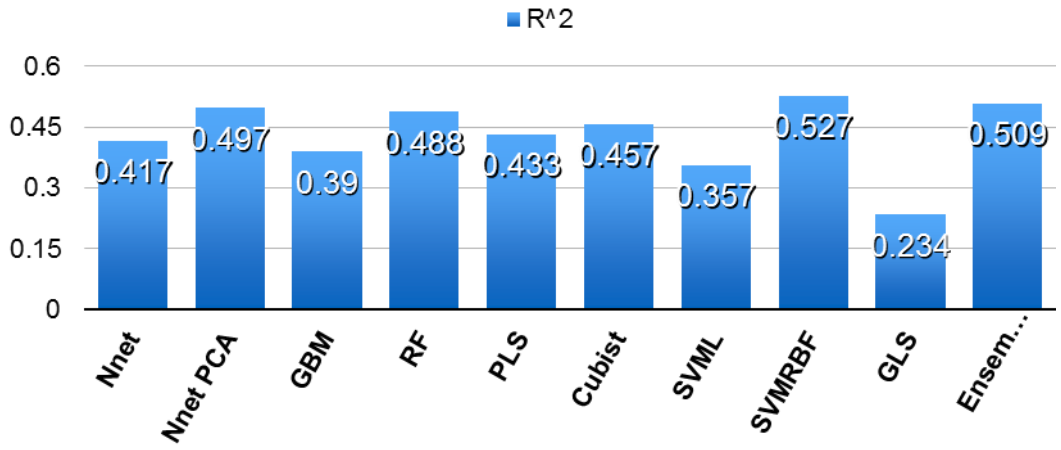

Figure 3: R^2, all algorithms

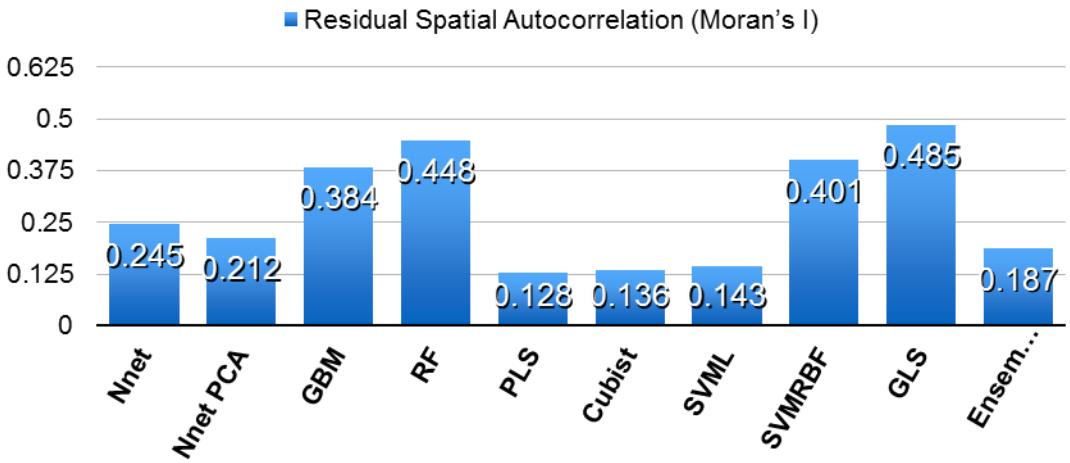

Figure 4: Residual SAC, all algorithms

Our findings show that NNetPCA, SVMRBF and RF are the best performers in RMSE and R^2. The fact that NnetPCA is amongst the top performers can be explained by the suitability of the PCA step for extracting signal from correlated data. RF is naturally strong against multicollinearity and small perturbations in the data since it uses random random subset of features at each tree. The success of SVMRBF can be explained by the regularization step which is effective generating a more parsimonious function and for the natural strength of the Radial Basis Function as kernel.

There is an overall poor performance of all the algorithms in R^2 which can be explained by the exogenous nature of the response variable.

The algorithms PLS, Cubist and SVML showed the best performance in terms of SAC. The explanation of why

We estimate that NNetPCA gives the best standalone results considering our combined criteria of goodness of fit and residual SAC. NNetPCA ranks 1st in RMSE,

2nd in R^2 and 4th in residual SAC. The second best standalone performer is Cubist, which ranks 4th in RMSE, 4th in R^2 and 2nd in residual SAC.

GLS although a spatial method, underperforms in all aspects when compared to the MLAs tested. This might be explained by the fact that GLS only allows the residuals to be spatially autocorrelated, while it requires explanatory variables to be independent. Also GLS, as used in this study does not handle multi collinear data.

The tested ensemble averaged the estimations of NNetPCA with the estimations of Cubist. The combined solution ranks 1st in RMSE, 2nd in R^2 and 4th in residual SAC. This solution offers an improvement on RMSE, R^2 and residual SAC over the best standalone algorithm (NNetPCA). It also offers an improvement on R^2 and RMSE over the second best overall performer (Cubist). It should be noted however, that further improvement in the combined solution is possible if, for instance, a meta learner that maximizes goodness of fit and minimizes residual SAC is used for blending the results. Nonetheless, we consider the combined solution to be the best solution delivered as it balances predictive power while keeping the residual SAC low.

We are aware however, of some limitations in our study regarding the cross validation schema utilized. As standard cross validation does not account for the spatial location of the training and testing records, the results might be showing some degree of overfitting.

4. Conclusions and Outlook

This brief study showed the differences in performance of some well-established ML algorithms for a case study using real estate data, and how they compare against a well-known spatial algorithm. It was proven that the chosen MLAs perform better than GLS in the case study presented. We also showed through an example that ensemble learning can potentially be used to improve the solution delivered by ML algorithms.

Based on our results we suggest a blind approach for regression of spatial variable where no specific ML"Apartment house (5+ units)"A is chosen before hand, but rather, many algorithms are applied to the same problem and the solution that better balances goodness of fit and residual spatial autocorrelation is chosen. We also encourage further research on ways applying ensemble learning to optimize regression of spatial variables as we have identified this as a promising and understudied area of research.

The next step in our research is to compare ML algorithms with a larger suite of spatial algorithms and study the use of meta-learners to blend different solutions under our “performance” criteria. Also, our future work will study the transferability of regression models built with different ML techniques.

5. Acknowledgements

We want to thank Urban4M LLC for facilitating the data used in this study and for the economic support, and NSF for the economic support for attending the Geocomputation 2015 conference.

6. References

Berwald J., et al., Using Machine Learning to Predict Catastrophes in Dynamical Systems. *Journal of Computational and Applied Mathematics,* 236(9):2235-2245, 2012

Bork L., Moller S., Forecasting house prices in the 50 states. using Dynamic Model Averaging and Dynamic Model Selection, *International Journal of Forecasting,* 31:63-78

Breiman L., Random Forests, *Machine Learning*, 45:5-32, 2001

Browne MW., Generalized Least Squares Estimators in Analysis of Covariance Structures, *South African Statistical Journal*, 8(1):1-24, 1974

Buphinder S., et al, Improved PLS Algorithms, *Journal of Chemometrics*, 11(1):73-85, 1998

Cheong Y., et al. Assessment of land use factors associate with dengue cases in Malaysia using Boosted Regression Trees. *Spatial and Spatio-temporal Epidemiology*. 10:75-84, 2014.

Cracknell M. & Reading A. Geological Mapping Using Remote Sensing Data: A Comparison of Five Machine Learning Algorithms, Their Response to Variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences*, 63:22-33, 2014.

Vapnik V.N., Statistical Learning Theory. *John Wiley & Sons, Inc.,*New York, USA p.736.

Freund Y. & Schapire R.E., A decision-theoretic generalization of on-line learning and an application to boosting, *Journal of Computer and System Sciences,* 55(1):119-139, 1997

Lakes T., et al. Cropland change in southern Romania: a comparison of logistic regressions and artificial neural networks. *Landscape Ecology*, 25(9):1195-1206, 2009.

Li J., et al. Advances in Spatially Environmental Modeling: Application of Machine Learning Methods to Spatial Interpolation of Environmental Variables. *Environmental Modelling & Software*, 26(12):1647-1659, 2011.

Mu J., et al., Housing Value Forecasting Based on Machine Learning Methods, *Abstract and Applied Analysis*, 2014, ID 648047, 2014

Okujeni A., et al. A comparison of advanced regression algorithms for quantifying urban land cover. *Remote Sensing*. 6:6324-6346, 2014.

Quinlan J.R., Learning with Continuous Classes, *Proceedings of 5th Australian Joint Conference on Artificial Intelligence, World Scientific, Singapore*, 1992

Ripley B.D., Pattern Recognition and Neural Networks. Cambridge, 1996

Santibanez S., et al., Analysis of the Performance of Some Machine Learning Algorithms for Regression Under Varying Spatial Autocorrelation, *Proceedings of the 18th AGILE Conference on Geographic Information Science*, 2015

Shepard E., et al., Interpolation of Population Related Polygon Data, *Proceedings ESRI user's conference*. 1997

United States Census, American Community Survey, *http://www.census.gov/acs/www/data_documentation/data_main/*. Access: 09/2014