Using mixed-effect random forest models to capture spatial patterns: a case study on urban crime

R. Zurita-Milla*[1], A. Fakhrurrozi[1,2] and O. Kounadi[1]

[1]Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, the Netherlands
[2]Research Center for Geotechnology, Indonesian Institute of Sciences, Bandung, Indonesia
*Email: r.zurita-milla@utwente.nl

Abstract

The increasing access to spatio-temporal datasets, data-driven modelling methods and computational power have transformed the way we do science. Yet most geodata-driven approaches currently disregard the spatial and temporal aspects of the data they are based on. Here we present and evaluate a hybrid machine learning approach that combines statistical mixed effects theory with the power of random forests. This approach, namely mixed-effects random forests or MERF, is used to model monthly crimes in New York City (USA). Our results show that MERF leads to lower prediction errors and to lower spatial autocorrelation in the residuals than a standard random forest model. This shows that there are approaches to mitigate the non-geocomputational nature of machine learning methods.

Keywords: Machine learning, data-driven models, regression, spatial autocorrelation

1. Introduction

An ever-increasing array of spatio-temporal datasets is becoming available to geo-information practitioners and researchers. This relatively new phenomenon, coupled with the emergence of novel analytical approaches, has transformed the way we do science and manage our socio-economic and natural resources. Yet, most of these analytical approaches are not geocomputational by nature. Or put in other words, they do not explicitly consider the presence of spatial and/or spatio-temporal patterns in the data. This despite the fact that such patterns might negatively affect their performance. For instance, if the target variable is spatially autocorrelated, then the assumption of independence in regression models is violated (Lichstein et al., 2002). Moreover, spatial autocorrelation often leads to correlated residuals that indicated structural problems in the selected analytical approaches (Chen, 2016). Spatial autoregressive models have been proposed to handle these autocorrelation problems (Hua et al., 2016). However, these statistical models have difficulties coping with the variety, volume, velocity and high dimensionality of modern spatio-temporal datasets. This difficulty explains the rise and pervasiveness of machine learning-based approaches.

Machine learning-based approaches can solve classification and regression tasks involving small and big data. However, the use of machine learning methods with spatio-temporal datasets requires careful consideration because these methods are not inherently equipped to deal with spatial or spatio-temporal patterns (Santibañez et al., 2015).

Here we present and evaluate a hybrid machine learning method that combines statistical mixed effects theory (to deal with clustered data) with the power of random forests (to model high-dimensional and non-linear problems). Mixed effects random forests (c.f. section 2) have, at least theoretically, a higher potential to properly capture the patterns present in the data. A case study (section 3) based on modelling crime is used to illustrate our work.

2. Mixed effects random forests

Linear mixed effects models are relatively popular because data tends to be clustered (Blood et al., 2010; Zhang et al., 2016). From a geographical perspective this means that data are either clustered hierarchically (i.e. various types of crops inside a generic class "agriculture") or longitudinally (time series available for a set of locations) (Meng, et al., 2012; Ver Hoef et al., 2010). Mathematically linear mixed effects models can be represented as:

$$y_i = X_i\beta + Z_i b_i + \epsilon_i \qquad \text{Equation 1}$$

Where y_i is a vector containing the target variable in cluster i, Xi and Zi are matrices with the fixed and random effects features, β is the unknown vector of fixed effects coefficients, b_i is the unknown vector of random effects coefficients for the cluster *i*, and ϵi is the unknown vector of residual errors.

The mixed-effects random forest (MERF) model was first proposed by Hajjem et al. (2014) who used it to predict the first-week box office revenues of movies. Their results confirmed the superiority of MERF over a standard random forest (RF) model in this non-spatial case study. MERF's mathematical formulation is entirely parallel to that of the linear mixed effect model but replacing the fixed effect term ($X_i\beta$) by a RF model, which can be represented by an unknown function f of the matrix of fixed effects features:

$$y_i = f(X) + Z_i b_i + \epsilon_i \qquad \text{Equation 2}$$

MERF is solved using an iterative approach that relies on the expectation–maximization algorithm (Moon, 1996). Once the algorithm converges, it can predict the value of new observations within a given cluster by adding the population-averaged prediction of the just trained RF model to the corresponding random effects. For more details regarding the MERF algorithm, please refer to Hajjem et al. (2014).

3. Case study

3.1. Objective and experimental set up

The performance of MERF is evaluated using crime and complaints data from New York City (NYC), the most populous city in the USA. A standard RF model is used to benchmark our results.

Both the crime and the complaints datasets were obtained from the NYC open data website (NYC Information Technology & Telecommunications, 2019). The crime dataset refers to all the crimes reported by the NYC police department and contains the date of occurrence and the coordinates of the crime. Reported complaints to the 311 service were used as explanatory variables. This dataset contains various types of complaint, the date when they were reported and the coordinates.

Considering data availability, we concentrated our analysis on eight of the most common complaints found in the 311 dataset and limited our study to the period 2010 to 2017. Further both the crime and complaint datasets were spatially aggregated to the postal zip code level (248 units) and temporally aggregated to a monthly time scale.

One-hot-encoding was used to create machine learning-valid features from the discrete counts found in the complaint dataset. This encoding method was also applied to the zip codes (which were used as clustering feature in the MERF model and as a regular feature in the standard RF model), and to the month feature of the crime dataset. To consider spatial effects, we calculated for each zip code the average number of crimes that occurred in adjacent zip code areas in the previous year (i.e. time offset spatial lag) as well as the local indicators of spatial autocorrelation (also known as LISA's quadrants) of the previous year.

All features were normalized using a robust scaling approach based on the median and interquartile range. This normalization method was selected because it can handle outliers and non-normally distributed features. Group k-fold was used to find the best parameters for both the MERF and RF models. This cross-validation method preserves the temporal structure of our data because full years were included/excluded in the training process.

Finally, we performed numerous experiments to optimize the features that should be used to determine the fixed and random effects. Our model evaluation metrics consisted on the mean absolute error (MAE) of the predicted number of monthly crimes in each zip code, and on the spatial autocorrelation of the residual errors measured by the Moran's I statistic. These metrics were used to find the best MERF model (using data for 2010-2016), and to evaluate the resulting model (for the year 2017).

3.2. Results

The results of our parameterization experiments show that the spatial feature (lagged number of crimes directly around each zip code) consistently gets high feature importance. This confirms the spatio-temporal autocorrelation of the target feature and justifies the use of MERF models. At the same time, it indicates that the selected complaints have less explanatory power to predict crime. Our results also show that there is a positive correlation between the MAE of the prediction and the spatial autocorrelation of the residuals: the lower the errors, the less spatially autocorrelated the residuals are.

Model	MAE of prediction	Moran's I of residual errors
MERF	8.89	0.12
RF	9.72	0.14

Table 1: Mean absolute error of the prediction and spatial autocorrelation of the residuals as provided by the MERF and the standard RF models

The best MERF model had the complaints, the lagged number of crimes and LISA's quadrants as fixed variables, and the month and the lagged LISA's quadrants as random effects features. The zip code was used as cluster variable. A side-by-side comparison between the MERF and RF models (Table 1) shows that the former has both a lower MAE for the predicted crimes, and a lower spatial autocorrelation of the residuals.

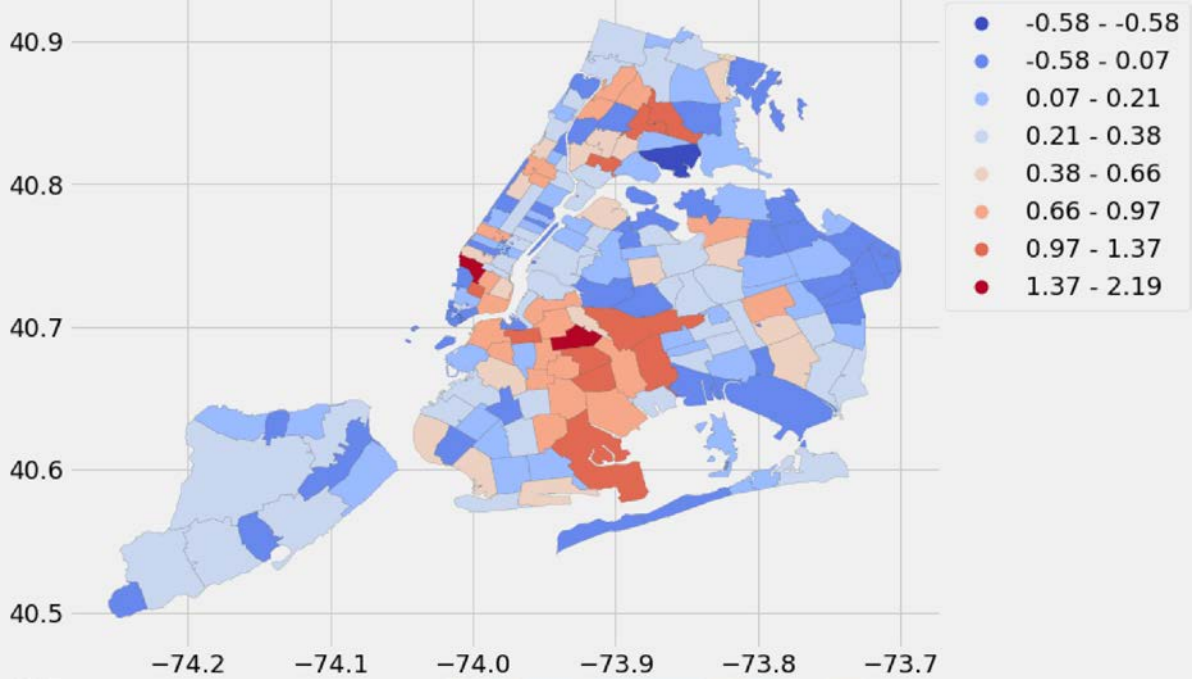

Figure 1: Random effects coefficients for each zip code in NYC

Figure 1 shows the spatial distribution of the random effects' coefficients (i.e. the b_i vector in Equation 2). This map illustrates both the intensity and direction of these effects as well as their spatial clustering in NYC. Finally, figure 2 shows that the MERF model can capture the spatial patterns of crime in NYC, represented here by the LISA's quadrants for April 2017. The corresponding Moran's I for the actual crimes is 0.51 and the predicted one is 0.53.

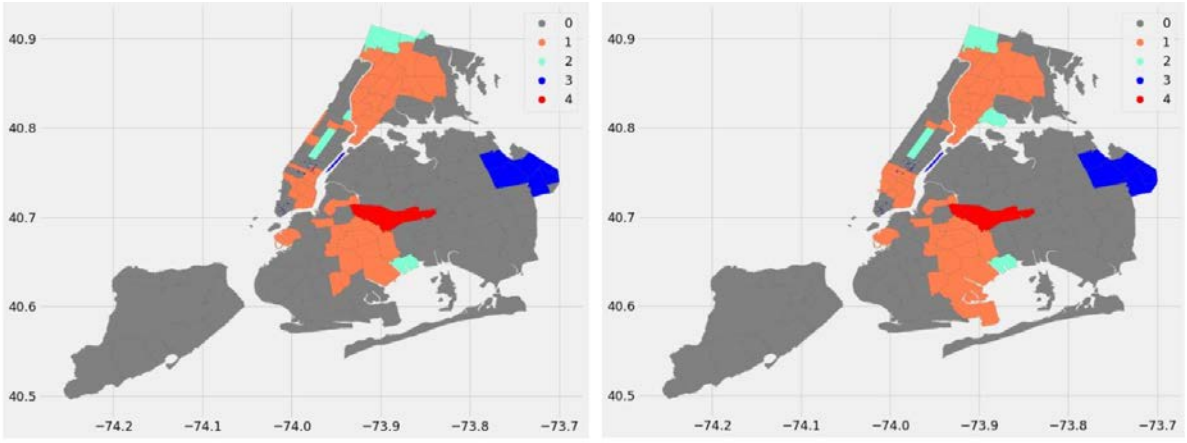

Figure 2: Actual (left) versus predicted (right) distribution of LISA's quadrants for each zip code and for April 2017. Map legend; 0 for not significant, 1 for high-high spatial clusters, 2 for low-high spatial outliers, 3 for low – low spatial cluster, 4 for high – low spatial outliers.

4. Conclusions

The use of machine learning approaches is becoming popular among geo-information practitioners and researchers. Yet, the implementation of these approaches requires careful consideration because spatial data is special. Here we evaluate the capabilities of mixed effects random forests (MERF) to capture spatial patterns of crime in New York City. Our results confirm that MERF models coupled with spatial variables can capture spatial patterns better than standard random forest methods. The MERF models had lower errors and lead to residuals with less spatial autocorrelation. This shows that there are approaches to mitigate the non-geocomputational nature of machine learning methods and that further research is needed to design spatially aware data-driven models.

5. Acknowledgements

A. Fakhrurrozi would like to thank the RISETPro project, funded by the world bank, for allowing him to continue his studies at the Faculty ITC of the University of Twente. Many thanks to the Research Center for Geotechnology of the Indonesian Institute of Sciences (LIPI) for providing access to their high-performance computing facilities.

6. References

Blood, E.A., Cabral, H., Heeren, T. and Cheng, D.M., 2010. Performance of mixed effects models in the analysis of mediated longitudinal data. *BMC medical research methodology*, *10*(1), p.16.

Chen, Y., 2016. Spatial autocorrelation approaches to testing residuals from least squares regression. *PloS one*, 11(1), p.e0146865.

Hajjem, A., Bellavance, F. and Larocque, D., 2014. Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), pp.1313-1328.

Hua, W., Junfeng, Z., Fubao, Z. and Weiwei, Z., 2016. Analysis of spatial pattern of aerosol optical depth and affecting factors using spatial autocorrelation and spatial autoregressive model. *Environmental Earth Sciences*, *75*, pp.1-17.

Lichstein, J.W., Simons, T.R., Shriner, S.A. and Franzreb, K.E., 2002. Spatial autocorrelation and autoregressive models in ecology. *Ecological monographs*, *72*(3), pp.445-463.

Meng, S.X., Huang, S., Vanderschaaf, C.L., Yang, Y. and Trincado, G., 2012. Accounting for serial correlation and its impact on forecasting ability of a fixed-and mixed-effects basal area model: a case study. *European journal of* forest *research*, *131*(3), pp.541-552.

Moon, T.K., 1996. The expectation-maximization algorithm. *IEEE Signal processing magazine*, 13(6), pp.47-60.

NYC Information Technology & Telecommunications, 2019. NYC Open Data. Retrieved February 1, 2019, from https://opendata.cityofnewyork.us/

Santibañez, S. F., Lakes, T., and Kloft, M. 2015. Performance analysis of some machine learning algorithms for regression under varying spatial autocorrelation. In Bacao, F. Yasmina Santos, M. and Painho M. Ed. 18th AGILE International Conference on Geographic Information Science, 9-12 June 2015, Lisbon, Portugal,

Ver Hoef, J.M., London, J.M. and Boveng, P.L., 2010. Fast computing of some generalized linear mixed pseudo-models with temporal autocorrelation. *Computational Statistics*, *25*(1), pp.39-55.

Zhang, D., Sun, J.L. and Pieper, K., 2016. Bivariate mixed effects analysis of clustered data with large cluster sizes. *Statistics in biosciences*, *8*(2), pp.220-233.