





# Investigating Moran's $I$ Properties for Spatial Machine Learning: A Preliminary Analysis

Jakub Nowosad <sup>1,2</sup> and Hanna Meyer <sup>1</sup>

<sup>1</sup>Institute of Landscape Ecology, University of Münster, Heisenbergstraße 2, Münster, 48149, Germany

<sup>2</sup>Institute of Geoecology and Geoinformation, Adam Mickiewicz University, B. Krygowskiego 10, Poznań, 61-680, Poland

Correspondence: Jakub Nowosad ([nowosad.jakub@gmail.com](mailto:nowosad.jakub@gmail.com))

**Abstract.** This study explores the application of Moran's  $I$ , a measure of spatial autocorrelation, in evaluating spatial machine learning models, specifically focusing on random forest (RF) models applied to simulated raster data with varying spatial structures. The research simulates 300 scenarios (raster datasets), each with different spatial autocorrelation ranges (10, 50, and 100). It assesses model performance using root mean square error (RMSE) and Moran's  $I$  values of the residuals across the entire raster, as well as for both training and testing samples. Based on our experimental setup, the results show that Moran's  $I$  of the residuals is affected by the spatial structure of the data, with higher values observed for datasets with greater autocorrelation ranges. A weak correlation is found between RMSE and Moran's  $I$  values, suggesting that Moran's  $I$  can offer valuable supplementary insights beyond RMSE in evaluating the spatial quality of models. However, the study also highlights the sensitivity of Moran's  $I$  to sample size and spatial proximity, which can lead to misleading interpretations of model quality. These findings underscore the potential limitations of relying solely on Moran's  $I$  in spatial machine learning applications and raise critical questions regarding its dependence on sample size and spatial distance. The study calls for further investigation into these factors to enhance model evaluation and improve the accuracy of spatial model assessments.

**Submission Type.** Analysis

**BoK Concepts.** [AM7-4] Global measures of spatial association, [IP3-4-7] Machine learning, [IP4-2-1] Accuracy assessment

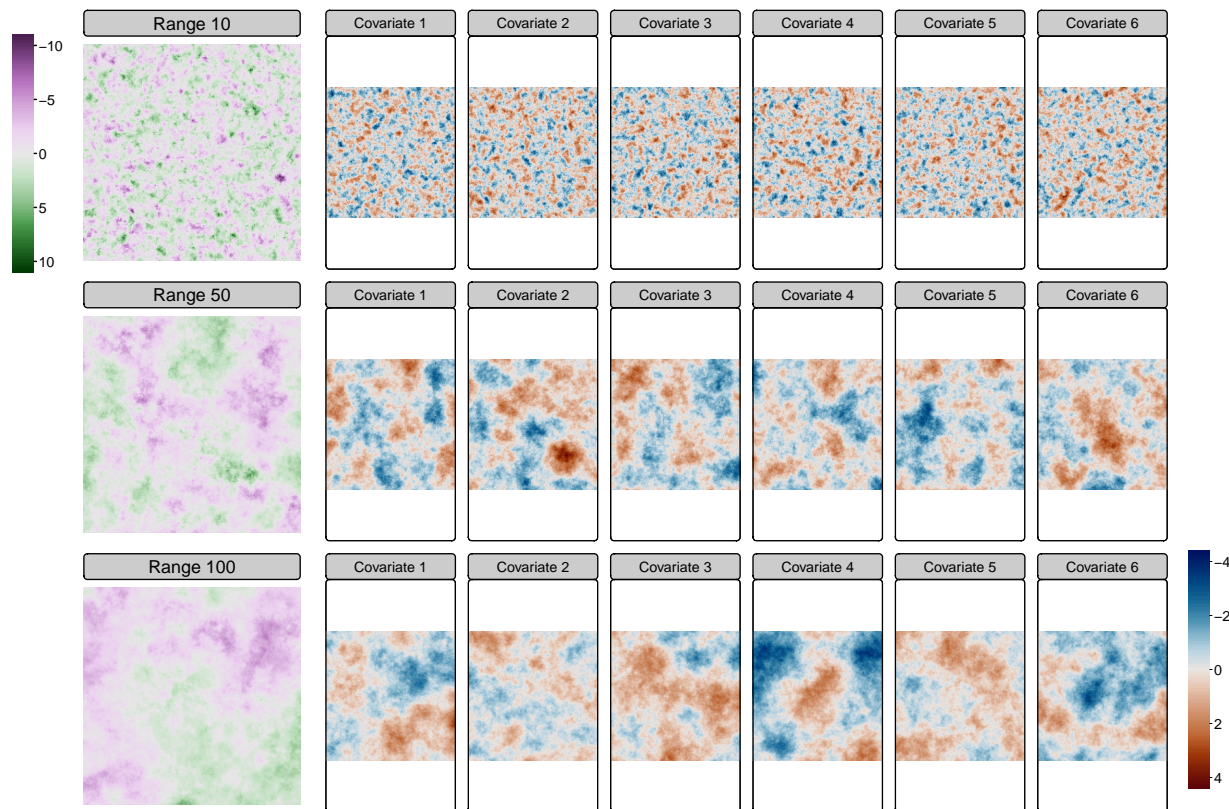
**Keywords.** spatial machine learning, Moran's  $I$ , random forest, spatial autocorrelation, model evaluation

## 1 Introduction

Machine learning methods have been widely applied for prediction and classification of spatial data (henceforth referred to as spatial machine learning) (Nikparvar and Thill, 2021). While traditional models such as support vector machines, random forests (RF), and gradient boosting machines have demonstrated effectiveness in these tasks, they inherently lack spatial awareness. They do not directly account for the spatial structure of the data, such as the spatial relationships between the observations, as their results are based on data in tabular form. If spatial structure in the data is relevant for the prediction, this can lead to models' subpar quality and weak performance in spatial prediction tasks (Meyer et al., 2018; Behrens et al., 2018).

Various approaches have been proposed to incorporate spatial information into machine learning models to improve their performance and better capture the spatial structure of the outcomes (Jemeljanova et al., 2024). One of the most common approaches is to include spatial proxies, such as the coordinates of the observations or Euclidean distances between them, as additional predictors in the model (Behrens et al., 2018). Other approaches include the inclusion of spatial predictors based on a distance matrix among training cases (Dray et al., 2006; Hengl et al., 2018), or applying spatially-aware cross-validation techniques for feature selection and model tuning (Meyer et al., 2019; Schratz et al., 2019). Moreover, variants of machine learning models have been developed to incorporate spatial information directly into the method, such as Geographical RF (Georganos et al., 2021) or RF-GLS (Saha et al., 2023).

A distinct role in spatial machine learning models is played by measures of spatial autocorrelation, such as Moran's  $I$  (Moran, 1950). These measures are used both before the modeling process to understand the spatial structure of the data and after the modeling process to evaluate the model's performance. In the latter case, the



**Figure 1.** Examples of simulated outcomes based on the set of covariates with different ranges (10, 50, 100)

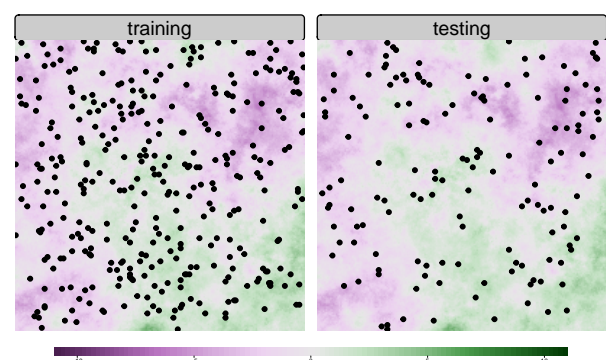
model's residuals are derived, and the spatial autocorrelation of these residuals is calculated: when the residuals are spatially autocorrelated, it indicates that the model is not able to capture the spatial structure of the data well (Besag, 1974; Dormann et al., 2007). In some studies, this information is further used to improve the model by, for example, geostatistical modeling of the residuals and adding the predicted values of this model to the initial model predictions (Hengl et al., 2015; Beguin et al., 2017). Other studies use this information to understand the model's limitations and interpret the results (Mascaro et al., 2014; Kirkwood et al., 2016; Liu et al., 2022; Kim et al., 2023).

Given its widespread use in spatial machine learning, this study explores the strengths and limitations of Moran's  $I$  within this context. Here, we focus on the random forest model and relationships between Moran's  $I$  values for the simulations with different spatial structures. We also investigate the relationship between Moran's  $I$  values of the residuals of the model and the model's quality, as measured by the root mean square error (RMSE).

## 2 Materials and Methods

The study workflow consists of three main parts: data simulation and sampling, modeling, and model evaluation. First, 300 raster outcomes of various spatial structures were simulated, and 350 training and 150 testing sam-

ples were created for each raster. Next, a random forest model was fitted for each raster, and the model's quality was evaluated using the RMSE value. The Moran's  $I$  value was calculated for the residuals of the RF model for the whole raster area, the training samples, and the testing samples. Lastly, the values of the RMSE and Moran's  $I$  were analyzed, and the relationship between the RMSE and Moran's  $I$  values was investigated.



**Figure 2.** Example of a training (350 locations) and testing sample (150 locations) on a simulated map with a range of 100

## 2.1 Simulated Data and Sampling

A raster template with 200 columns and 200 rows was generated as the foundation for constructing both covariates and outcomes. Six covariates were created using the conditional Gaussian simulation based on a spherical variogram model. This process was repeated 100 times for each of the three specified ranges (10, 50, and 100) with a mean of 0 and constant sill of 1 and a nugget value of 0. The covariates were then combined to generate an outcome raster with the formula  $Y = X1 + X2 \cdot X3 + X4 + X5 \cdot X6$ ; we included interactions to reflect potential dependencies between covariates. Thus, a total of 300 outcome rasters were produced, each with six related covariates (Figure 1).

For each outcome raster, a training and testing sample was generated. The locations of the training and testing samples were randomly selected from the raster area, with 300 points for the training samples and 150 points for the testing samples (Figure 2). The goal of this split was to mimic an interpolation problem: predicting the outcome for the locations within the training sample, and evaluating the model's performance using the testing sample found within the same study area.

## 2.2 Modeling

For each outcome raster, we extracted the corresponding values of the covariates and the outcome for the training samples, and fitted a random forest (RF) model. The RF model was fitted with 500 trees with the number of variables to possibly split at in each node (*mtry*) tuned based on the values of 2, 3, 4, 5, and 6. The final model was selected based on the lowest root mean square error (RMSE) value calculated using the out-of-bag (OOB) samples. The final model was used to predict the outcome for the whole raster area, the training samples, and the testing samples.

## 2.3 Model Evaluation Metrics

The quality of each RF model was evaluated using the RMSE value calculated for the whole raster area, the training samples, and the testing samples. Next, the residuals of the RF model (differences between the predicted and observed values) were derived for the whole raster area, the training samples, and the testing samples. These residuals were used to compute Moran's *I* (Moran, 1950) for the whole raster area, the training samples, and the testing samples. Moran's *I* is calculated as:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad (1)$$

where  $n$  is the number of observations,  $x_i$  and  $x_j$  are the values of the observations at locations  $i$  and  $j$ ,  $\bar{x}$  is the mean value of the observations, and  $w_{ij}$  is the spatial weight between the observations at locations  $i$  and  $j$ . Here,

**Table 1.** Average RMSE for different ranges and sample types

Range	Overall	Training	Testing
10	1.349	0.576	1.347
50	1.074	0.458	1.062
100	0.776	0.321	0.761

we used a binary spatial weight matrix based on the eight nearest neighbors, where  $w_{ij} = 1$  if the observations at locations  $i$  and  $j$  are neighbors and  $w_{ij} = 0$  otherwise. Thus, in case of the raster data, the Moran's *I* value was based on the values of eight closest cells, while for training and testing samples, the Moran's *I* value was based on the values of eight closest point samples.

Moran's *I* measures spatial autocorrelation on a scale from -1 to 1, where -1 indicates strong negative spatial autocorrelation (a checkerboard pattern), 0 indicates no spatial autocorrelation, and 1 indicates strong positive spatial autocorrelation with similar values close to each other.

## 2.4 Data and Software Availability

All analyses were conducted in R (R Core Team, 2024). The raster data was processed using the terra package (Hijmans, 2025), while the spatial simulations were conducted using the simsam and the gstat packages (Nowosad, 2025; Pebesma, 2004). Random forest models were fitted using the ranger package (Wright and Ziegler, 2017) and the spatial autocorrelation was calculated using the spdep package (Bivand et al., 2013). Visualizations in the manuscript were made using the ggplot2 (Wickham, 2016) and the tmap packages (Tennekes, 2018).

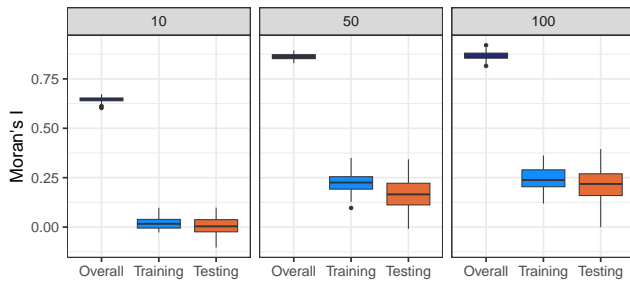
The research code supporting this publication is available at <https://github.com/Nowosad/moran-i-spatial-ml-prelim>.

## 3 Results

In total, 300 RF models were fitted and evaluated, with 100 models for each autocorrelation range. The RMSE values for the whole raster area, the training samples, and the testing samples were calculated for each model, and their average values are presented in Table 1. Two main observations can be made based on the results. First, the models' accuracy, as measured by the RMSE, largely depends on the range of the covariates. Models based on the simulations with a range of 10 have the highest RMSE values, while models based on the simulations with a range of 100 have the lowest RMSE values. Second, while the RMSE values of the training sample are much lower than the overall RMSE values (overly optimistic), the RMSE values of the testing samples are comparable to the overall RMSE values.

Figure 3 shows the Moran's *I* values of the RF models residuals for the whole raster area, the training samples,





**Figure 3.** Moran's  $I$  for different ranges and sample types

and the testing samples. In general, the overall Moran's  $I$  values are much higher than the Moran's  $I$  values of the training and testing samples. Moreover, the residuals of the RF models based on the simulations with a range of 10 have the relatively lowest Moran's  $I$  values, while the residuals of the RF models based on ranges of 50 and 100 have higher Moran's  $I$  values. In all cases the models residuals of the whole rasters are spatially autocorrelated, and thus the RF models were not able to capture the spatial structure of the data well. Training and testing samples for the models of the outcome with the shortest range have the Moran's  $I$  values close to zero, while the Moran's  $I$  values of the training and testing samples for the models of the longest range are higher, but still much lower than the Moran's  $I$  values of the whole raster area. It also seems that the Moran's  $I$  values of the testing samples have smaller median values of Moran's  $I$  than the values of the training samples, while also having a larger variability.

Lastly, we checked the relationship between the RMSE values of the models and Moran's  $I$  values of the RF models residuals (Figure 4). There is a weak positive relationship between the RMSE and Moran's  $I$  values for the whole raster area with a coefficient of determination ( $R^2$ ) of 0.09–0.5. A slight positive correlation was also observed for the training samples for ranges of 50 and 100, while for the testing samples, the relationship is much weaker or even not present.

#### 4 Conclusions and Future Research

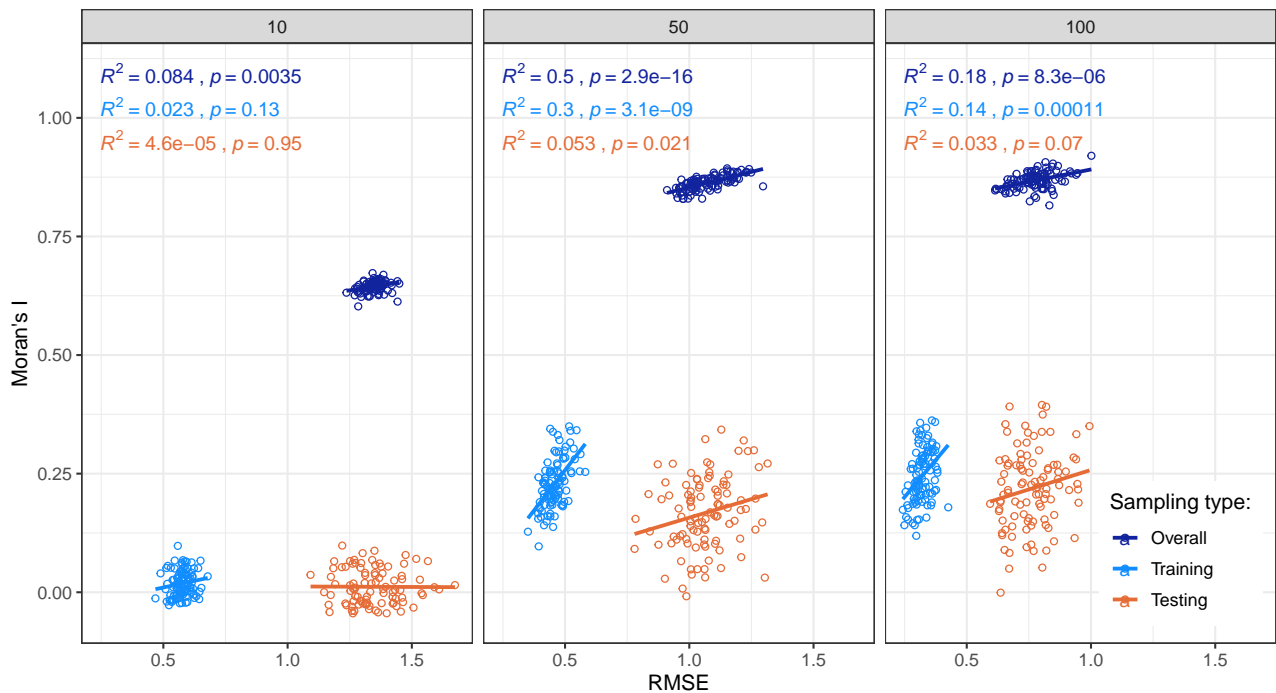
This preliminary work investigates Moran's  $I$  capabilities and limitations in the context of spatial machine learning, using a set of simulated data and random forest models. The results show that the relationship between the RMSE and Moran's  $I$  values of the RF models' residuals is weak for the overall raster area, very weak for the testing samples, and barely present for the training samples. It suggests that Moran's  $I$  values of the residuals provide other information than the RMSE values and can thus be used as an additional measure to evaluate the model's quality.

At the same time, our results suggest that Moran's  $I$ , as measured in this study, exhibits certain properties that may limit its usefulness in spatial machine learning applica-

tions. The Moran's  $I$  values of the residuals of training and testing samples are much lower than the Moran's  $I$  values of the whole raster area. This indicates that the RF models were not able to capture the complete spatial structure of the data well, but Moran's  $I$  values of the residuals of the training and testing samples did not reflect this well. In practice we do not have access to the whole raster area, and thus the Moran's  $I$  values of the residuals of the training and testing samples may not be a good representation of the existence of spatial autocorrelation in the model's residuals. This may create a false sense of confidence in the model's quality, suggesting that the model is able to capture the spatial structure of the complete data well, while in reality, it is not.

Moreover, the relation between Moran's  $I$  values for the whole raster area, the training samples, and the testing samples suggests that, in general, the larger the sample size, the higher Moran's  $I$  values of the residuals are. This indicates that the residuals' Moran's  $I$  values are sensitive to the sampling process and size, or, more specifically, to the distances between the neighboring samples. It makes sense: the larger the random sample size, the closer the neighboring samples are, and thus, their values are more similar, leading to higher Moran's  $I$  values. As stated by Makido and Shortridge (2007), "Moran's  $I$  is a function of spatial resolution," which suggests that Moran's  $I$  values of the residuals may more reflect the sample size than the model's quality. Additionally, this suggests that comparing Moran's  $I$  values between models with different sample sizes may not be meaningful.

The presented results preliminary suggest various properties of Moran's  $I$  in the context of spatial machine learning, while also opening new questions. The issue of the sample size and the distances between the neighboring samples should be further investigated to understand their impact on Moran's  $I$  values of the residuals. For example, is it possible to standardize Moran's  $I$  values of the residuals based on the sample size and possibly sample scheme to make them comparable between models with different sample sizes? Moreover, the study framework could be expanded in a few ways. Simulations without a spatial structure or based on covariates not related to the outcome could be used to understand the relationship between Moran's  $I$  values of the residuals and the model's quality. Including spatial proxies as additional predictors in the RF models could also provide insights into the relationship between Moran's  $I$  values of the residuals and the proxies' properties. Additionally, the expanded framework could help to further clarify and explain the relationship between the RMSE and Moran's  $I$  values of the residuals. In this work, we focused on an interpolation problem, but it would be interesting to check how Moran's  $I$  values of the residuals behave in an extrapolation problem. Lastly, it is worth comparing Moran's  $I$  values of the residuals of the RF models with other measures of spatial autocorrelation, such as Geary's  $C$ , to understand their differences and their usefulness in the context of spatial machine learning.



**Figure 4.** Moran's  $I$  for different ranges and sample types

## Declaration of Generative AI in writing

The authors declare that they have used Generative AI tools in the preparation of this manuscript. The AI tools were utilized for language editing, but not for generating scientific content, research data, or any conclusions. All intellectual and creative work, including the analysis and interpretation of data, is original and has been conducted by the authors without AI assistance.

## Acknowledgments

We sincerely thank the three anonymous reviewers for their valuable feedback and contributions to improving this and future work. This project has received the financial support of the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101147446.

## References

- Beguín, J., Fuglstad, G.-A., Mansuy, N., and Paré, D.: Predicting Soil Properties in the Canadian Boreal Forest with Limited Data: Comparison of Spatial and Non-Spatial Statistical Approaches, *Geoderma*, 306, 195–205, <https://doi.org/10.1016/j.geoderma.2017.06.016>, 2017.
- Behrens, T., Schmidt, K., Viscarra Rossel, R. A., Gries, P., Scholten, T., and MacMillan, R. A.: Spatial Modelling with Euclidean Distance Fields and Machine Learning, *European Journal of Soil Science*, 69, 757–770, <https://doi.org/10.1111/ejss.12687>, 2018.

- Besag, J.: Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 36, 192–225, <https://doi.org/10.1111/j.2517-6161.1974.tb00999.x>, 1974.
- Bivand, R. S., Pebesma, E., and Gómez-Rubio, V.: *Applied Spatial Data Analysis with R*, Second Edition, Springer, NY, <https://asdar-book.org/>, 2013.
- Dormann, C., McPherson, J., Araújo, M., Bivand, R., Bolliger, J., Carl, G., Davies, R., Hirzel, A., Jetz, W., Daniel Kissling, W., Kühn, I., Ohlemüller, R., Peres-Neto, P., Reineking, B., Schröder, B., Schurr, F., and Wilson, R.: Methods to Account for Spatial Autocorrelation in the Analysis of Species Distributional Data: A Review, *Ecography*, 30, 609–628, <https://doi.org/10.1111/j.2007.0906-7590.05171.x>, 2007.
- Dray, S., Legendre, P., and Peres-Neto, P. R.: Spatial Modelling: A Comprehensive Framework for Principal Coordinate Analysis of Neighbour Matrices (PCNM), *Ecological Modelling*, 196, 483–493, <https://doi.org/10.1016/j.ecolmodel.2006.02.015>, 2006.
- Georganos, S., Grippa, T., Niang Gadiaga, A., Linard, C., Lennert, M., Vanhuysse, S., Mboga, N., Wolff, E., and Kalogirou, S.: Geographical Random Forests: A Spatial Extension of the Random Forest Algorithm to Address Spatial Heterogeneity in Remote Sensing and Population Modelling, *Geocarto International*, 36, 121–136, <https://doi.org/10.1080/10106049.2019.1595177>, 2021.
- Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Mendes De Jesus, J., Tamene, L., and Tondoh, J. E.: Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions, *PLOS ONE*, 10, e0125814, <https://doi.org/10.1371/journal.pone.0125814>, 2015.

- Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B., and Gräler, B.: Random Forest as a Generic Framework for Predictive Modeling of Spatial and Spatio-Temporal Variables, *PeerJ*, 6, e5518, <https://doi.org/10.7717/peerj.5518>, 2018.
- Hijmans, R. J.: Terra: Spatial Data Analysis, [https://github.com/rspatial/terra](https://github.com/rsatial/terra), 2025.
- Jemeljanova, M., Knoch, A., and Uemaa, E.: Adapting Machine Learning for Environmental Spatial Data - A Review, *Ecological Informatics*, 81, <https://doi.org/10.1016/j.ecoinf.2024.102634>, 2024.
- Kim, D., Song, I., Miralha, L., Hirmas, D. R., McEwan, R. W., Mueller, T. G., and Šamonil, P.: Consequences of Spatial Structure in Soil-Geomorphic Data on the Results of Machine Learning Models, *Geocarto International*, 38, 2245–381, <https://doi.org/10.1080/10106049.2023.2245381>, 2023.
- Kirkwood, C., Cave, M., Beamish, D., Grebby, S., and Ferreira, A.: A Machine Learning Approach to Geochemical Mapping, *Journal of Geochemical Exploration*, 167, 49–61, <https://doi.org/10.1016/j.gexplo.2016.05.003>, 2016.
- Liu, X., Kounadi, O., and Zurita-Milla, R.: Incorporating Spatial Autocorrelation in Machine Learning Models Using Spatial Lag and Eigenvector Spatial Filtering Features, *ISPRS International Journal of Geo-Information*, 11, 242, <https://doi.org/10.3390/ijgi11040242>, 2022.
- Makido, Y. and Shortridge, A.: Weighting Function Alternatives for a Subpixel Allocation Model, *Photogrammetric Engineering & Remote Sensing*, 73, 1233–1240, <https://doi.org/10.14358/PERS.73.11.1233>, 2007.
- Mascaro, J., Asner, G. P., Knapp, D. E., Kennedy-Bowdoin, T., Martin, R. E., Anderson, C., Higgins, M., and Chadwick, K. D.: A Tale of Two “Forests”: Random Forest Machine Learning Aids Tropical Forest Carbon Mapping, *PLoS ONE*, 9, e85993, <https://doi.org/10.1371/journal.pone.0085993>, 2014.
- Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., and Nauss, T.: Improving Performance of Spatio-Temporal Machine Learning Models Using Forward Feature Selection and Target-Oriented Validation, *Environmental Modelling & Software*, 101, 1–9, <https://doi.org/10.1016/j.envsoft.2017.12.001>, 2018.
- Meyer, H., Reudenbach, C., Wöllauer, S., and Nauss, T.: Importance of Spatial Predictor Variable Selection in Machine Learning Applications – Moving from Data Reproduction to Spatial Prediction, *Ecological Modelling*, 411, 108815, <https://doi.org/10.1016/j.ecolmodel.2019.108815>, 2019.
- Moran, P. A. P.: Notes on Continuous Stochastic Phenomena, *Biometrika*, 37, 1950.
- Nikparvar, B. and Thill, J.-C.: Machine Learning of Spatial Data, *ISPRS International Journal of Geo-Information*, 10, 600, <https://doi.org/10.3390/ijgi10090600>, 2021.
- Nowosad, J.: Simsam: Simulating and Sampling Spatial Data, <https://github.com/nowosad/simsam>, 2025.
- Pebesma, E. J.: Multivariable Geostatistics in S: The Gstat Package, *Computers & Geosciences*, 30, 683–691, 2004.
- R Core Team: R: A Language and Environment for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>, 2024.
- Saha, A., Basu, S., and Datta, A.: Random Forests for Spatially Dependent Data, *Journal of the American Statistical Association*, 118, 665–683, <https://doi.org/10.1080/01621459.2021.1950003>, 2023.
- Schratz, P., Muenchow, J., Iturrutxa, E., Richter, J., and Brenning, A.: Hyperparameter Tuning and Performance Assessment of Statistical and Machine-Learning Algorithms Using Spatial Data, *Ecological Modelling*, 406, 109–120, <https://doi.org/10.1016/j.ecolmodel.2019.06.002>, 2019.
- Tennekes, M.: tmap: Thematic Maps in R, *Journal of Statistical Software*, 84, 1–39, <https://doi.org/10.18637/jss.v084.i06>, 2018.
- Wickham, H.: Ggplot2: Elegant Graphics for Data Analysis, Springer-Verlag New York, <https://ggplot2.tidyverse.org>, 2016.
- Wright, M. N. and Ziegler, A.: ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, *Journal of Statistical Software*, 77, 1–17, <https://doi.org/10.18637/jss.v077.i01>, 2017.