Cahiers du CLSL, n° 69, 2025, pp. 127-142

On machine learning from environmental data

Mikhail Kanevski
University of Lausanne
mikhail.kanevski@unil.ch

Abstract

The application of machine learning (ML) algorithms for analyzing, modelling, and visualizing geospatial data has seen remarkable growth in recent years. Many ML techniques have proven to be both effective and efficient in addressing complex challenges within the geosciences. Selected topics on adaptation and application of ML to environmental data are briefly discussed in this research.

1 Introduction

This paper focuses on key aspects of applying machine learning to the analysis and modelling of spatial environmental data. It emphasizes a comprehensive methodology that encompasses from data collection (monitoring network analysis, design and redesign), intelligent exploratory data analysis and visualization, via ML models training and evaluation to understanding and communication of the results for informed decision making (Kanevski & Maignan, 2004; Kanevski et al., 2009). The main components of this methodology are illustrated in figure 1 (p. 132), with further details provided in the subsequent sections.

In practice, ML should be widely utilized across all phases of data-driven modelling. It facilitates comprehensive data exploration, selection of relevant variables, visualization of high-dimensional and large datasets, pattern recognition, adaptive modelling, and result interpretability. Numerous excellent books cover the concepts and theories of ML algorithms, as well as their applications across various domains (see, for example Bishop & Bishop, 2024; Cherkassky & Mulier, 2007;

Hastie et al., 2021; Haykin, 2009). Additionally, an abundance of resources on ML programming and modern open-access packages is now available, further accelerating development and broadening the user community (James et al., 2021; Kuhn & Johnson, 2013).

Recently, a new and rapidly evolving field of research, *geospatial data science*, has emerged as an interdisciplinary domain that integrates geoinformatics, geostatistics, machine learning, network science, and more, see, for example (Gaur et al., 2023; Pebesma & Bivand, 2023). This field encompasses a broad range of applications across diverse domains, including geography and urban planning, earth sciences, environmental science, meteorology and climate science, ecology, and beyond.

Geospatial data present several significant challenges due to their inherent characteristics. Unlike traditional datasets, they often exhibit spatial autocorrelation (Jemeljanova et al., 2024; Kattenborn et al., 2022), meaning that observations are not independently and identically distributed (i.i.d.), which complicates standard statistical and machine learning approaches (Linnenbrink et al., 2023). Additionally, spatial clustering and preferential sampling can introduce biases, affecting model development, evaluation, and the definition of validity domains (Brus, 2022; Meyer & Pebesma, 2021; Schratz et al., 2019).

Moreover, causal analysis in spatial data is particularly challenging due to confounding spatial dependencies and intrinsic uncertainties, making robust inference difficult (Akbari et al., 2023; Gao et al., 2022).

The interpretability of ML-based data analysis and predictions is also becoming increasingly critical in geoscience applications, particularly in studies related to climate, environmental risks, natural hazards, and renewable energy assessments (Jiang et al., 2024).

Addressing these challenges requires the development of specialized methodologies to ensure accurate modelling, reliable predictions, and informed decision-making in geospatial applications.

Prof. F. Bavaud has made important contributions to geospatial data science, particularly in developing innovative algorithms and methodologies (Bavaud, 2009, 2014, 2024; Guex et al., 2023). His work effectively combines theoretical rigour with practical applications, bridging

the gap between fundamental research and real-world challenges.

1.1 Why machine learning?

The application of machine learning in environmental data analysis offers numerous benefits that enhance data processing and modelling while addressing complex environmental problems. Here are several key advantages:

- ML algorithms are powerful *universal non-linear* modelling tools. They can model data with high precision, adapting to a wide range of structures and variability.
- ML exhibit strong generalization capabilities: they make accurate predictions on new unseen (testing) data.
- ML models can handle big and complex data sets from diverse sources.
- ML techniques perform well with high-dimensional data: many algorithms have been specifically developed for non-linear dimensionality reduction and feature selection.
- ML algorithms can be integrated into automatic spatio-temporal environmental data processing and monitoring.
- ML models have already demonstrated their efficiency and usefulness in numerous environmental applications (weather, climate, natural hazards, pollution, renewable resources, ecology, biodiversity, etc.).

Despite their significant success, the use of ML algorithms can encounter various challenges. ML heavily depends on the quality and quantity of data; in real-world applications, training, selecting, and evaluating ML models are non-trivial tasks. Additionally, making predictions and forecasts that account for uncertainties is often difficult. The interpretability and explainability of ML models remain active areas of contemporary research (Jiang et al., 2024). It is also crucial to note that the effective application of ML requires a deep understanding

of algorithms, a grasp of underlying assumptions, and collaboration with domain experts.

1.2 Model-centric and data-centric machine learning

In the field of artificial intelligence (AI), there are two primary approaches to improve the results: model-centric and data-centric, often called MCAI and DCAI. Each focuses on different aspects of ML modelling and has its particular advantages.

Model-centric focuses on model architecture and strengthening the algorithms, in particular on model selection and evaluation, hyperparameters tuning, optimization techniques, combining models, ensemble learning (Bartz et al., 2023; Bishop & Bishop, 2024; Hastie et al., 2021; Montavon et al., 2012). Selection of ML model appropriate to data and objectives can help to achieve state-of-the-art results in many applications. Until recently, model-centric modelling has dominated in developments.

Nowadays, the role of data in ML modelling is promoted by the fast developing concept of data-centric ML (Mahalle et al., 2024). This demonstrates an important shift from model strengthening to data quality and reliability.

In data-centric approach the focus is on *systematic* improvement of data quality and quantity in the process of ML training and evaluation. Data are considered as a dynamic object, while model is usually fixed. Fundamentally there are two possibilities: either improving training dataset using available data or collection/simulation (data augmentation) of additional data.

There are many tools and techniques that are a part of data-centric approach: missing values treatment, outliers/anomalies detection and removal, data validation and correction of errors, feature engineering and data reduction (features and/or instances selection), active learning (guided data selection), interactive visualization. All these techniques improve the quality and quantity of training and testing data sets giving rise to better predictions and reducing errors.

One of the first question in the analysis concerns data representativity, i.e., how available data represent the phenomena under study? In

spatial statistics this is mostly related to data clustering and preferential sampling resulting in biases in estimates and predictions, (Chilès, 2012; Kanevski, 2013). In ML, understanding the topology of the input feature space is critical for quantifying its spatial and dimensional resolution, as well as for defining the validity domain (Kanevski, 2013).

Empirically it was shown that cleaning and refining data can result in better model performance compared to just increasing the model complexity and optimization.

Let us note that training multiple models with different origins for the same task yields valuable insights into the data, improves modelling and enriches the interpretation of results.

2 Methodology

Our experience on application of ML to environmental data has resulted in the development of a generic methodology pointing out the most essential phases of the study, figure 1. First, let us precise some important characteristics of environmental data justifying the use of advanced ML techniques.

2.1 Environmental data

Environmental data are an interesting domain of ML application for several reasons, in particular: 1) quantity (small, large and big) and diverse quality; 2) non-linearity; 3) high spatio-temporal variability; 4) dimensionality (often environmental problems are considered in high dimensional feature spaces); 5) noise and uncertainties; 6) presence of extreme values.

Several fundamental problems commonly encountered in environmental data studies can be efficiently addressed by well-developed basic machine learning models, including: clustering – identifying similar groups in data; classification – analysis and prediction of categorical/discrete data; regression – analysis and prediction of continuous data and probability density function modelling and prediction, which plays a central role in environmental risk assessment.

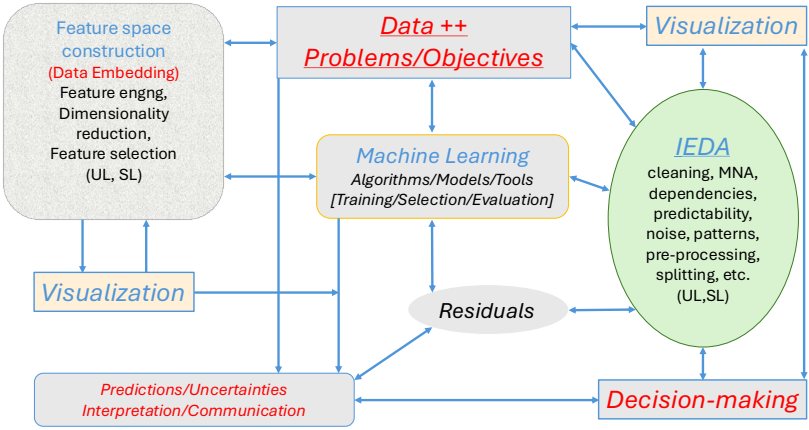

FIGURE 1 – Generic methodology of environmental data ML analysis and modelling. Data++ – raw, reduced and augmented data sets; IEDA – intelligent exploratory data analysis; MNA – monitoring network analysis; UL/SL – unsupervised/supervised learning.

2.2 Data exploration and pre-processing

Exploratory data analysis (EDA) and data pre-processing play a critical role in ML modelling, as the efficacy of data treatment and the results are significantly influenced by the quality and quantity of the input data. They help to better understand the original data and phenomena under study as well as in a proper selection of modelling tools relevant to the objectives of the study. Traditional (EDA) includes techniques to summarize data statistical properties, often using visualization tools, without modelling or making prior hypothesis. It is interesting to note that modern "data science" has its roots in classical EDA (Donoho, 2017).

Contemporary intelligent exploratory data analysis (IEDA) integrates tools and models from statistics, machine learning, and data visualization (Martinez et al., 2022). These tools assist in detecting patterns, quantifying predictability, and constructing the most relevant input space for predictive learning. Efficient and appropriate IEDA is a crucial component of successful data-driven modelling.

Data pre-processing is a key component of IEDA. In addition to

standard techniques such as scaling and normalization, we can highlight several other important methods, including the treatment of missing and extreme values, outlier detection, and feature engineering. Feature engineering involves transforming existing features and creating new ones using domain knowledge to enhance modelling and improve results interpretation. Other processes include data augmentation and splitting the dataset into training, validation, and testing subsets taking into account properties of spatio-temporal data (data clustering, biases, spatial correlations and dependencies) (Kattenborn et al., 2022).

2.3 Data visualization and visual analytics

The importance of visualization in data analysis was well recognized already long time ago by John Tukey (1985): "There is nothing better than a picture for making you think of questions you had forgotten to ask (even mentally)" (Friedman & Stuetzle, 2002, p. 1629). In environmental studies, maps are among the most widely used visualization techniques, effectively summarizing results to support decision-making.

Recently, the field of data visualization has evolved into distinct research domains — visual analytics and visual data mining (Andrienko et al., 2020). Interactive visualization plays a significant role at all stages of environmental data study: from data collection, model construction and learning to the results communication. High dimensional and multivariate data before being visualized usually are processed by applying algorithms and tools for dimensionality reduction, feature selection, projections, etc. (Kuhn & Johnson, 2019; Lee & Verleysen, 2007). Raw high-dimensional data can be visualized and analyzed using popular techniques like parallel coordinates (Inselberg, 2009). Recent advancements in the field of high-dimensional big data visualization have been supported by both algorithmic innovations and hardware improvements, enabling more efficient processing and interpretation of complex datasets.

2.4 Feature selection and dimensionality reduction

The construction of a high-dimensional input feature space relies on three primary sources: 1) expert knowledge of the phenomena and objectives of the study, 2) a critical analysis of existing literature, and

3) feature engineering techniques. However, it is often unclear whether this feature space is comprehensive or contains redundancies. As a result, the input space may include relevant, irrelevant, and redundant features. Therefore, the application of unsupervised and/or supervised feature selection and feature extraction algorithms is essential. These methods not only help reduce dimensionality but also enhance the speed and quality of modeling, while improving interpretability and visualization. (Bolón-Canedo et al., 2015; Guyon & Kacprzyk, 2006; Kuhn & Johnson, 2019; Lee & Verleysen, 2007).

There are three basic approaches to feature selection (FS): 1) *filter methods*, which assess feature relevance independently of machine learning models; 2) *wrapper methods*, which evaluate subsets of features based on model performance; and 3) *embedded methods*, where feature selection is integrated into the training process. The choice of most appropriate FS methods depends on several factors, including data complexity, available computational resources, and the specific model being utilized, since different models may produce varying subsets of features.

Another essential aspect of FS methodology is the concept of intrinsic dimension (ID) of data (Lee & Verleysen, 2007). Numerous methods exist for estimating ID (Camastra & Staiano, 2016). A novel ID estimator based on the multipoint Morisita index introduced initially for spatial data clustering (Morisita, 1959) was proposed in (Golay & Kanevski, 2015). It was demonstrated how Morisita index can be utilized in FS for supervised regression tasks (Golay et al., 2017) and for reducing redundancy in data (Golay & Kanevski, 2017). This research gave an intriguing connection between classical spatial statistics, fractal concepts and contemporary machine learning.

It is important to underline that a well-defined input space in modeling spatial, temporal, or spatio-temporal data can significantly enhance model performance, even when using relatively simple approaches. Conversely, a poorly constructed feature space can lead to suboptimal results, regardless of the complexity or sophistication of the models employed.

3 Machine learning modelling

ML models rely on two types of parameters: hyper-parameters and internal parameters. Hyper-parameters are set by the user before training, while internal parameters are estimated during the training process. For instance, in a multilayer perceptron (MLP), the number of hidden layers and the number of neurons in each layer are hyper-parameters. In contrast, the weights of the connections are computed through optimization algorithms during training.

Fundamentally, learning from data involves two significant steps: model selection and model evaluation. To facilitate these processes, the data are divided into training, validation, and testing subsets. The splitting of geospatial data is a non-trivial task and remains an active area of research (Linnenbrink et al., 2023; Meyer & Pebesma, 2022).

The training subset is used to determine the model's internal parameters, while the validation subset helps identify the optimal combination of hyper-parameters, completing the model selection phase. The test subset is then employed to evaluate the performance of the selected model. The test error provides an estimate of the generalization error, reflecting the model's performance on new unseen data (Bishop & Bishop, 2024; Hastie et al., 2021; Kanevski et al., 2009).

A variety of optimization techniques, e.g., family of gradient descent, are utilized to effectively navigate the solution space and identify optimal or near-optimal parameters for the model. The selection of the cost function is critical, as it influences not only the learning process but also the model's generalization ability to test data.

In more complex scenarios, constraints may be introduced to the optimization problem to account for real-world limitations, for example, expert knowledge, or constraints imposed by physical laws (Karniadakis et al., 2021).

To enhance model performance and prevent overfitting, various effective regularization techniques are also employed (Bishop & Bishop, 2024; Hastie et al., 2021; Haykin, 2009). Regularization helps to constrain model complexity, ensuring that the model not only explains training data but generalizes well.

One of the fundamental questions in data-driven modelling is whether

there exists useful and structured information (i.e., patterns) within the data. Specifically, we seek to determine if the data are predictable within a given input feature space. In the field of geostatistics, this discrimination (pattern – no pattern) can be achieved through the application of variography. When a variogram exhibits a pure nugget effect, it indicates the absence of spatial correlations, signifying the lack of patterns in the data (Chilès, 2012; Kanevski & Maignan, 2004; Kanevsky et al., 1996).

To further assess the presence of patterns, researchers can shuffle the data, resulting in a dataset that maintains the same distribution but destroyed any inherent patterns. Such shuffled datasets serve as control sets, allowing us to evaluate how the model reacts to data lacking of patterns.

A separate and essential question deals with the estimation of the noise in data before modelling. Having this information, we can better perform modelling, avoid overfitting and contribute to the interpretability. There are several non-parametric approaches capable of performing this task (Devroye et al., 2018; Liitiainen et al., 2009).

In summary, data can be represented as: $data = information + noise$ (*unexplained_variability*). The goal of modelling is to extract the information while ensuring that the residuals consist solely of noise. One effective method to achieve this is to shuffle the raw data and the residuals and apply the same analytical approach.

4 Conclusions

Properly applying machine learning in environmental modelling requires deep expertise in both machine learning techniques and the specific data domain. Once the original problem is appropriately reformulated in terms of machine learning, a wide range of models can be employed, including Gaussian processes, random forests, gradient boosting machines, support vector machines, artificial neural networks, deep learning and graph neural networks, among others. These machine learning models are well-developed, optimized, and extensively tested in real-world applications. They are implemented in efficient packages available in R, Python, and other programming languages.

Machine learning is advancing very rapidly, changing fundamentally research domains and practical applications. The integration of openness and reproducibility in modern science makes datasets and codes accessible, which attracts new researchers and accelerates innovation.

To conclude, let us recall some current trends in ML application in environmental studies:

- Nowadays deep learning is an extremely popular approach in environmental applications of machine learning. It drives interest to modern ML and its wide use in applications.

- Physics-informed ML and similar approaches integrate domain expertise and fundamental theories into machine learning models, significantly enhancing their accuracy and reliability in scientific applications.

- Causality. By leveraging ML, researchers have made substantial progress in identifying causal relationships, enabling a deeper understanding of complex systems and improving predictive capabilities (Peters et al., 2017; Runge et al., 2019).

- ML models, in particular deep learning models, are often criticized as "black boxes" due to their complexity and lack of transparency. Recent advancements in Explainable AI (XAI) focus on making these models more interpretable and transparent. XAI aims to bridge the gap between the power of complex ML models and the need for clear, understandable explanations (Molnar, 2018).

- Methodological and practical developments and improvements in data-centric approach are important parts of current innovations in ML.

- Uncertainties quantification and visualization for better environmental risk assessments and intelligent decision making.

And finally, ML is a very useful and stimulating approach in contemporary science worth learning and applying.

Acknowledgements

This paper is dedicated to the memory of Prof. M. Maignan. Collaboration with Michel was always interesting, stimulating and successful. I would also like to thank also to many colleagues and PhD students, for their numerous fruitful discussions on spatial statistics and machine learning, as well as for insights on real data case studies. The improvements in English text were partly made possible thanks to ChatGPT.

References

Akbari, K., Winter, S., & Tomko, M. (2023). Spatial causality: A systematic review on spatial causal inference. *Geographical Analysis*, 55(1):56–89.

Andrienko, N., Andrienko, G., Fuchs, G., Slingsby, A., Turkay, C., & Wrobel, S. (2020). *Visual analytics for data scientists*. Springer, Cham, 1st edition.

Bartz, E., Bartz-Beielstein, T., Zaefferer, M., & Mersmann, O. (2023). *Hyperparameter tuning for machine and deep learning with R: A practical guide*. Springer Nature, Singapore.

Bavaud, F. (2009). Information theory, relative entropy and statistics. In Sommaruga, G. (Ed.), *Formal theories of information: From Shannon to semantic information theory and general concepts of information*, pages 54–78. Springer, Berlin, Heidelberg.

Bavaud, F. (2014). Spatial weights: constructing weight-compatible exchange matrices from proximity matrices. In Duckham, M., Pebesma, E., Stewart, K., & Frank, A. U. (Eds.), *Geographic Information Science*, volume 8728, pages 81–96. Springer International Publishing, Cham.

Bavaud, F. (2024). Measuring and testing multivariate spatial autocorrelation in a weighted setting: A kernel approach. *Geographical Analysis*, pages 573–599.

Bishop, C. M. & Bishop, H. (2024). *Deep learning: Foundations and concepts*. Springer International Publishing, Cham.

Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2015). *Feature selection for high-dimensional data*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer International Publishing, Cham.

Brus, D. J. (2022). *Spatial sampling with R*. Chapman & Hall/CRC the R Series. CRC Press, Boca Raton, 1st edition.

Camastra, F. & Staiano, A. (2016). Intrinsic dimension estimation: Advances and open problems. *Information Sciences*, 328:26–41.

Cherkassky, V. S. & Mulier, F. (2007). *Learning from data: Concepts, theory, and methods*. IEEE Press : Wiley-Interscience, Hoboken, NJ, 2nd edition.

Chilès, J.-P. (2012). *Geostatistics: Modeling patial uncertainty*. Number v.713 in Wiley Series in Probability and Statistics Ser. John Wiley & Sons, Incorporated, Hoboken, NJ, 2nd edition.

Devroye, L., Györfi, L., Lugosi, G., & Walk, H. (2018). A nearest neighbor estimate of the residual variance. *Electronic Journal of Statistics*, 12(1):1752–1778.

Donoho, D. (2017). 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766.

Friedman, J. H. & Stuetzle, W. (2002). John W. Tukey's work on interactive graphics. *The Annals of Statistics*, 30(6):1629–1639.

Gao, B., Wang, J., Stein, A., & Chen, Z. (2022). Causal inference in spatial statistics. *Spatial statistics*, 50:100621.

Gaur, L., Garg, P., & Dey, N. (Eds.) (2023). *Emerging trends, techniques, and applications in geospatial data science*. Advances in Geospatial Technologies. IGI Global, Hershey, PA.

Golay, J. & Kanevski, M. (2015). A new estimator of intrinsic dimension based on the multipoint Morisita index. *Pattern Recognition*, 48(12):4070–4081.

Golay, J. & Kanevski, M. (2017). Unsupervised feature selection based on the Morisita estimator of intrinsic dimension. *Knowledge-Based Systems*, 135:125–134.

Golay, J., Leuenberger, M., & Kanevski, M. (2017). Feature selection for regression problems based on the Morisita estimator of intrinsic dimension. *Pattern Recognition*, 70:126–138.

Guex, G., Loup, R., & Bavaud, F. (2023). Estimation of flow trajectories in a multi-lines transportation network. *Applied Network Science*, 8(1):44.

Guyon, I. & Kacprzyk, J. (2006). *Feature extraction: Foundations and applications*. Number 207 in Studies in Fuzziness and Soft Computing. Springer-Verlag, Berlin.

Hastie, T., Tibshirani, R., & Friedman, J. (2021). *The elements of statistical learning*. Springer Texts in Statistics. Springer, New York, NY, 2nd edition.

Haykin, S. S. (2009). *Neural networks and learning machines*. Prentice-Hall, New York, Munich, 3rd edition.

Inselberg, A. (2009). *Parallel coordinates: Visual multidimensional geometry and its applications*. Springer, Dordrecht, NY.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R*. Springer Texts in Statistics. Springer, New York, NY, 2nd edition.

Jemeljanova, M., Kmoch, A., & Uuemaa, E. (2024). Adapting machine learning for environmental spatial data-a review. *Ecological Informatics*.

Jiang, S., Sweet, L.-b., Blougouras, G., Brenning, A., Li, W., Reichstein, M., Denzler, J., Shangguan, W., Yu, G., Huang, F., et al. (2024). How interpretable machine learning can benefit process understanding in the geosciences. *Earth's Future*, 12(7):e2024EF004540.

Kanevski, M. (2013). *Advanced mapping of environmental data*. Wiley, Somerset.

Kanevski, M. & Maignan, M. (2004). *Analysis and modelling of spatial environmental data*. EPFL press, Lausanne.

Kanevski, M., Pozdnukhov, A., & Timonin, V. (2009). *Machine learning for spatial environmental data. Theory, applications and software*. EPFL Press.

Kanevsky, M., Arutyunyan, R., Bolshov, L., Demyanov, V., & Maignan, M. (1996). Artificial neural networks and spatial estimation of Chernobyl fallout. *Geoinformatics*, 7(1-2):5–11.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., & Yang, L. (2021). Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440.

Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5:100018.

Kuhn, M. & Johnson, K. (2013). *Applied predictive modeling*. Springer, New York, NY.

Kuhn, M. & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman & Hall/CRC Data Science Series. CRC Press, Taylor & Francis Group, Boca Raton, London, New York, first issued in paperback edition.

Lee, J. A. & Verleysen, M. (Eds.) (2007). *Nonlinear dimensionality reduction*. Information Science and Statistics. Springer, New York, NY.

Liitiainen, E., Verleysen, M., Corona, F., & Lendasse, A. (2009). Residual variance estimation in machine learning. *Neurocomputing*, 72(16–18):3692–3703.

Linnenbrink, J., Milà, C., Ludwig, M., & Meyer, H. (2023). kNNDM: k-fold nearest neighbour distance matching cross-validation for map accuracy estimation. *EGUsphere*, 2023:1–16.

Mahalle, P. N., Wasatkar, N. N., & Shinde, G. R. (Eds.) (2024). *Data-centric artificial intelligence for multidisciplinary applications*. Chapman & Hall CRC, London.

Martinez, W. L., Martinez, A. R., & Solka, J. (2022). *Exploratory data analysis with MATLAB*. Chapman & Hall CRC, London, 3rd edition.

Meyer, H. & Pebesma, E. (2021). Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 12(9):1620–1633.

Meyer, H. & Pebesma, E. (2022). Machine learning based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1):2208.

Molnar, C. (2018). *Interpretable machine learning*. Leanpub, British Columbia, 3rd edition.

Montavon, G., Orr, G. B., & Müller, K.-R. (Eds.) (2012). *Neural networks: tricks of the trade*, volume 7700 of *Lecture Notes in Computer Science*. Springer, Berlin, Heidelberg, 2nd edition.

Morisita, M. (1959). Measuring of dispersion of individuals and analysis of the distributional patterns. *Memoires of the Faculty of Science, Kyushu University, Series E. Biology*, 2:215–235.

Pebesma, E. & Bivand, R. (2023). *Spatial data science: With applications in R*. Chapman & Hall CRC, London.

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press, Cambridge, MA, London, England.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J., et al. (2019). Inferring causation from time series in earth system sciences. *Nature communications*, 10(1):2553.

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine learning algorithms using spatial data. *Ecological Modelling*, 406:109–120.