

Sandra De Iaco
Monica Palma
Donato Posa *Editors*

Exploration of Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques

ECOST-DATA, Bari, Italy, July 3-4, 2025

OPEN ACCESS

 Springer

Springer Proceedings in Mathematics & Statistics

Volume 531

This book series features volumes composed of selected contributions from workshops and conferences in all areas of current research in mathematics and statistics, including data science, operations research and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Sandra De Iaco • Monica Palma • Donato Posa
Editors

Exploration of Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques

ECOST-DATA, Bari, Italy, July 3-4, 2025

Editors

Sandra De Iaco 
University of Salento
Lecce, Italy

Monica Palma
University of Salento
Lecce, Italy

Donato Posa
University of Salento
Lecce, Italy



Workshop on Exploration of Spatio-Temporal Environmental Conditions ecost-data ecost-data2025
ECoST-DATA Bari, Italy 2025 07 03 2025 07 04 <https://sites.google.com/unisalento.it/ecost-dataproject/workshop>

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-032-17525-0 ISBN 978-3-032-17526-7 (eBook)
<https://doi.org/10.1007/978-3-032-17526-7>

Mathematics Subject Classification: 60, 62

This work was supported by University of Salento.

© The Editor(s) (if applicable) and The Author(s) 2025. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this book or parts of it.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

This work is subject to copyright. All commercial rights are reserved by the author(s), whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Regarding these commercial rights a non-exclusive license has been granted to the publisher.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

This volume collects the scientific contributions presented at the Workshop “Exploration of Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques (ECoST-DATA)”, held at the University of Bari on July 3-4, 2025, as Satellite event of DUST 2025. It is a key result of the ECoST-DATA Project, funded by European Union – NextGenerationEU with the Cascade Open Calls published by ALMA MATER STUDIORUM – University of Bologna, inside the Project GRINS within the PNRR – Mission 4, Component 2, Investment 1.3 “Partnership extended to Universities, Research Centers, Firms and research projects funding”, D.D. 341 of 15/03/2022, CUP: J33C22002910001.

The contents are dedicated to advanced tools and techniques for analyzing and predicting environmental variables with a spatial or spatio-temporal structure. In particular, they include theoretical reviews on spatio-temporal covariance modelling as well as innovative approaches for assessing environmental quality and its effects on climate change by integrating georeferenced data from multiple sources.

Among the primary objectives of this collection, it is worth mentioning: transfer advanced analytic skills for the modelling, estimation, and prediction of spatial and spatio-temporal environmental variables; provide practical guidance on integrating heterogeneous georeferenced data while accounting for measurement error and spatial misalignment; support risk assessment by equipping practitioners with methods to quantify exposure and uncertainty relevant to human health; and inform sustainability planning through models that identify spatially and temporally targeted mitigation or adaptation strategies.

As a result, the contributions are intended for researchers and graduate students in spatial statistics, environmental science, and geosciences; data scientists and practitioners working on environmental monitoring, climate impacts, and public-health risk assessment; policy analysts interested in data-driven sustainability planning.

Regarding the outcomes, readers will gain both the theoretical grounding and the practical toolset necessary to build, validate, and deploy spatio-temporal models for environmental variables, produce reliable forecasts, and translate model outputs into actionable insights for health and sustainability interventions.

The volume is organized in three parts.

In the first one, *Random fields and Covariance modelling*, Posa clarifies the relationship between covariance functions and positive definite functions, emphasizing that Crum's theorem is the necessary generalization of Bochner's theorem. De Iaco discusses on some families of spatio-temporal complex covariance functions for vectorial data. It is also described how these functions can be generalized to the spatio-temporal setting, as well as how some other complex-valued covariance families can be constructed as positive mixtures of infinite components or as a convolution of the real part. Hristopulos presents the hybrid spectral approach for generating physically inspired, non-separable covariance kernels from stochastic ordinary differential equations. Pere et al. propose the stationary subspace analysis to decompose a random field into independent stationary and non-stationary subspaces.

In Part II, *Environmental control and integration*, Stein focuses on the use of modern deep learning methods for the analysis and management of large data volumes. His study explores the use of satellite imagery and time series to analyze air quality and monitor environmental changes. A further challenging approach is presented in Gómez-Hernández et al. to identify atmospheric dust sources, using the ensemble smoother with multiple data assimilation methodology. Capasso discusses the optimization of pollution control within an economic growth system, presenting a spatially structured model. This model analyzes an optimal control problem designed to maximize intertemporal utility while minimizing pollution, the disutility of taxation, and intervention costs. Additionally, the study by Dutillo and Di Battista explores the functional structure and variation of Italian carbon emissions from electricity using functional data analysis. Daily and monthly profiles are constructed using Fourier basis functions and analyzed at national and electricity market zone levels.

The volume concludes with Part III, *Environmental Analysis*, where Jalilian et al. apply a flexible generalized additive model framework for surface ozone prediction, using spatio-temporally varying coefficients to correct biases between satellite and in situ data. Gomes et al. investigate the geostatistical characterization of climate extremes dynamics, focusing on drought. A Random Forest-based short-term alert system is presented for predicting drought severity classes at a monthly scale, demonstrating strong predictive performance for local areas and offering a flexible tool for managing drought-related risks. For environmental quality assessment, the work by Masoumi et al. addresses a gap in biodiversity modelling in the Apulia region by comparing the performance of various machine learning models (Random Forests, Support Vector Machines, Extreme Gradient Boosting, Decision Trees,

and Linear Regression) for assessing the effects of environmental variables on the multilevel biodiversity index. Furthermore, in Cappello et al., spatio-temporal blind source separation model is reviewed and employed to model and predict ground-level ozone concentration in a complex multivariate spatio-temporal framework.

Lecce, Italy
September 2025

Sandra De Iaco
Donato Posa
Monica Palma

Acknowledgments This research was supported by the European Union-NextGenerationEU with the Cascade Open Calls published by ALMA MATER STUDIORUM - University of Bologna, inside the Project GRINS funded by PNRR - Mission 4, Component 2, Investment 1.3 “Partnership extended to Universities, Research Centers, Firms and research projects funding”, D.D. 341 of 15/03/2022, “ECoST-DATA, Exploring Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques”, CUP: J33C22002910001.

Contents

Part I Random Fields and Covariance Modelling

Covariance Functions	3
Donato Posa	
1 Introduction	3
2 <i>CFs</i> , <i>PD</i> and <i>SPD</i> Functions	4
3 Bochner's Theorem and the Corresponding Generalization	7
4 Conclusions	11
Bibliography	11
Spatiotemporal Complex Covariance Functions for Vectorial Data	15
Sandra De Iaco	
1 Introduction	15
2 Theory of Spatio-temporal Complex Random Fields	16
3 Some Complex Families from Space to Space-Time	19
4 Conclusions	26
Bibliography	26
Non-separable Covariance Kernels for Spatiotemporal Gaussian Processes Based on the Hybrid Spectral Method	29
Dionissios T. Hristopulos	
1 Introduction	29
2 Spatial Covariance Kernels from SPDEs	30
3 Space-Time Covariance Kernels from SPDEs?	34
4 On the Hybrid Spectral Approach	35
5 Conclusions	38
Bibliography	38
Stationary Subspace Analysis for Spatiotemporal Data	41
Jaakko Pere, Sandra De Iaco, and Klaus Nordhausen	
1 Introduction	41
2 Model Formulation	42
3 The stSSA Method for the Mean	44

4	Simulation Study	45
5	Conclusions	50
	Bibliography	51

Part II Environmental Control and Integration

Space Turns to Time: The Advent of Time Series of Remote Sensing Images	55
--	-----------

Alfred Stein

1	Air Quality	55
2	Time Series of Images	57
3	Integrating the Objects	59
4	Ongoing Research	59
5	Looking Forward	60
6	Concluding Remarks	61
	Bibliography	61

Ensemble Smoother with Multiple Data Assimilation for Atmospheric Dust Source Identification: A Generic Framework Approach	63
---	-----------

J. Jaime Gómez-Hernández, Valeria Todaro, Marco D’Oria, and Maria Giovanna Tanda

1	Introduction	64
2	Theoretical Background	65
3	Methodology for Atmospheric Dust Source Identification	66
4	Applications and Advantages	68
5	Future Developments	69
6	Conclusions	69
	Bibliography	70

Optimizing Pollution Control for an Economic Growth System	73
---	-----------

Vincenzo Capasso

1	Introduction	73
2	A Spatially Structured Model	75
3	An Optimal Control Problem	77
	Bibliography	79

Exploring Functional Structure and Variation of Italian Carbon Emissions	81
---	-----------

Pierdomenico Dutillo and Tonio Di Battista

1	Introduction	81
2	Data and Methodology	82
3	Explorative Analysis and FPCA Results	85
	Bibliography	90

Part III Environmental Analysis

Ozone Predictions Through a Generalized Additive Model	95
Abdollah Jalilian, Claudia Cappello, Monica Palma, and Sandra De Iaco	
1 Introduction	96
2 Dataset	96
3 Generalized Additive Model Framework	97
4 Model Evaluation and Results	98
5 Conclusions	101
Bibliography	102
 Characterization and Forecasting of Drought: Geostatistical Approaches	105
Miguel Gomes, Ana Castro, and Amilcar Soares	
1 Objectives	105
2 Study Region and Drought Indices	106
3 Spatial Characterization of the Main Temporal Trends	107
4 An Alert System for Short-Term Prediction of Drought Severity	111
5 Conclusions	114
Bibliography	114
 Using Machine Learning Methods to Explore the Effects of Environmental Variables on Biodiversity	117
Iman Masoumi, Amirhossein Najafabadipour, Sabrina Maggio, and Giuseppina Giungato	
1 Introduction	118
2 Methodology and Data	119
3 Environmental Variables and Biodiversity Prediction	125
4 Conclusions	127
Bibliography	128
 Modeling and Prediction of Ground-Level Ozone Concentrations in a Spatiotemporal Multivariate Context	131
Claudia Cappello, Klaus Nordhausen, and Monica Palma	
1 Introduction	131
2 The ST-BSS Based Approach	133
3 Spatiotemporal Modeling of Ozone Concentrations	133
4 Spatiotemporal Ozone Predictions	137
5 Conclusions	138
Bibliography	138

Using Machine Learning Methods to Explore the Effects of Environmental Variables on Biodiversity



Iman Masoumi, Amirhossein Najafabadipour, Sabrina Maggio,
and Giuseppina Giungato

Abstract Environmental factors, notably air pollutants and meteorological conditions, significantly impact biodiversity distribution and ecosystem dynamics. Among these, PM_{10} is identified as a major stressor, affecting plant physiological responses, reducing species abundance, and modifying soil microbial activity and nutrient cycling. These effects are often compounded when PM_{10} interacts with climatic conditions.

A comprehensive understanding of these combined effects is crucial for regions like Apulia, where biodiversity is shaped by both climate variability and localized human activities. The region remains underexplored in biodiversity modeling, particularly regarding the use of machine learning methods with the Multilevel Biodiversity Index (MBI).

This study aims to address this gap by a comparison of the performance of several Machine Learning (ML) models, including Random Forests (RFs), Support Vector Machines (SVMs), Extreme Gradient Boosting (XGBoost), Decision Trees (DTs), and Linear Regression (LR), for predicting MBI based on PM_{10} concentrations and meteorological variables. Among the models tested, DTs and XGBoost performed best, reliably predicting MBI based on various evaluation metrics.

Keywords Multilevel Biodiversity Index (MBI) · Machine learning (ML) · Fine particulate matter (PM_{10}) · Supervised machine learning algorithms

I. Masoumi

National Biodiversity Future Center, Palermo, Italy

A. Najafabadipour

Faculty of Mining Engineering, University of Jiroft, Jiroft, Iran

S. Maggio (✉) · G. Giungato

Department of Economic Sciences, University of Salento, Lecce, Italy

e-mail: sabrina.maggio@unisalento.it

© The Author(s) 2025

S. De Iaco et al. (eds.), *Exploration of Spatio-Temporal Environmental Conditions: Harmonized Databases and Analytical Techniques*, Springer Proceedings in Mathematics & Statistics 531, https://doi.org/10.1007/978-3-032-17526-7_11

1 Introduction

Biodiversity is crucial for several essential ecological functions, including pollination, water filtration, and climate regulation. However, global biodiversity is currently undergoing a rapid and sustained decline, often referred to as the sixth mass extinction [8, 27, 29]. Localized pressures are responsible for the increase in this decline, especially in biodiversity hotspots such as Italy's Apulia region, which balances significant ecological value with notable vulnerability to both regional and global environmental drivers [4].

Indeed, Apulia hosts diverse ecosystems, including farmlands, forests, urban areas, and protected zones, which are fundamental for unique species and ecological integrity. Despite this richness, Apulia faces increasing threats from pollution, land use changes, and climate variability. Human-driven factors such as urban expansion, intensive agriculture, and diffuse water contamination contribute to ecosystem fragmentation, placing rare and endemic species under additional stress.

Biodiversity decline has broader implications for global sustainability targets, particularly concerning the United Nations Sustainable Development Goals (SDGs), such as Goal 13 *Climate Action* and Goal 15 *Life on Land*. Changes in climate, including rising temperatures, altered precipitation regimes, and increased frequency of extreme weather events, can accelerate biodiversity loss by disrupting ecological stability. Local environmental factors, including land degradation and high levels of air pollutants, like PM_{10} , produce a negative impact on ecosystem resilience.

Abiotic factors like air pollutants and meteorological conditions significantly influence biodiversity patterns. PM_{10} impacts plant processes, species richness, and soil function [24]. These effects are amplified by climate factors, with temperature and precipitation changes affecting habitats and productivity and extreme weather potentially increasing pollutant deposition [24, 26, 29]. In this context, an integrated analysis of multiple drivers is necessary for managing biodiversity in ecologically diverse regions like Apulia [24, 29].

Despite the strong capabilities of machine learning (ML) models like Support Vector Machines (SVMs), Extreme Gradient Boosting (XGBoost), and Decision Tree (DTs) in ecological modeling, limitations persist concerning data quality, spatial heterogeneity, and generalization [9, 15, 25, 32, 35–37, 42]. While remote sensing and citizen science initiatives have improved data coverage, integrating diverse sources into consistent predictive models remains challenging, particularly for citizen science data with uneven spatial effort [28]. Given these limitations, there is a clear need for unified frameworks that can integrate various data types and spatial dimensions, especially in data-constrained regions such as Apulia.

Recent research has shown the power of advanced ML techniques, including ensemble methods like stacking and bagging, to handle complex and nonlinear ecological datasets and enhance predictive accuracy, although these techniques are rarely applied for biodiversity prediction purposes in the Mediterranean area [1, 20], compared to other environmental contexts and sustainability areas [17–19].

This study aims to predict the Multilevel Biodiversity Index (MBI), a composite indicator of species richness and land cover diversity [10] in Apulia region. In particular, a comparison of the performance of four supervised machine learning algorithms such as SVMs, XGBoost, DTs, and LR is discussed. Spatially interpolated environmental variables (PM_{10} , atmospheric temperature, atmospheric pressure, and precipitation) are considered to evaluate model accuracy and identify the best approach for catching spatial variation in biodiversity patterns throughout the Apulia region.

The developed ML framework provides accurate, spatially explicit biodiversity forecasting, thereby fostering scientific comprehension and supporting strategic policy in the study area.

Furthermore, this research improves MBI modeling framework in Mediterranean and data-scarce landscapes by resolving existing methodological issues and increasing predictive reliability through comparative model evaluation.

This chapter is organized into three main sections. Section 2 details the methodology and data covering the study area, data acquisition, the creation of the Multilevel Biodiversity Index (MBI), as well as the machine learning methods used, along with their evaluation metrics. Section 3 presents the biodiversity modeling outcomes, specifically the spatial interpolation of environmental variables in Apulia and the prediction performance of the MBI. Finally, Sect. 4 summarizes the key findings and discusses their implications for biodiversity conservation.

2 Methodology and Data

The methodological framework for predicting the MBI in Apulia has involved several key stages: (1) data collection and preprocessing to correct spatial misalignments and address data quality issues, (2) computation of MBI and interpolation of environmental variables via Ordinary Kriging, (3) fitting of ML models (DTs, XGBoost, SVMs, and LR) to predict MBI, and (4) validation of results.

In particular, the first step has consisted of gathering essential spatial datasets for the Apulia region, including species occurrence records, land use and land cover (LULC) data, and environmental variables such as PM_{10} concentrations, temperature, atmospheric pressure, and precipitation. Then, all datasets have been preprocessed to correct spatial misalignments and address data quality issues.

Following this, meteorological and air quality variables have been interpolated using geostatistical methods implemented in dedicated routines of the GSLib software. This process has allowed for the creation of a uniform prediction grid across the study area, enabling consistent spatial integration of heterogeneous inputs suitable for ML applications. Subsequently, the modeling phase has been based on the comparison of four supervised ML algorithms: XGBoost, SVM, DT, and LR. The dataset has been split into 70% for training and 30% for testing to thoroughly evaluate model performance. Each model has been trained to estimate MBI based on selected environmental predictors. Finally, model performance has been assessed

using various metrics, allowing comprehensive comparative evaluation of predictive accuracy.

2.1 Study Area and Data Acquisition

Apulia, located in Southern Italy, is a pivotal Mediterranean biodiversity hotspot featuring diverse ecosystems like farmlands, forests, urban areas, and protected zones. It is a crossroads of species and hosts vital habitats such as temporary ponds and coastal lagoons, essential for unique species and ecological integrity [7, 10]. However, the region faces growing threats from pollution, land use changes, and climate variability [30]. Human activities, including urban expansion, intensive agriculture, and water contamination, cause ecosystem fragmentation, endangering rare and endemic species. This highlights the need for targeted regional evaluations to comprehend and mitigate biodiversity loss in Apulia's vulnerable environments. For this aim, the analysis has started by gathering the following spatial datasets for the Apulia region:

- *Species Occurrence Data*, obtained from the Global Biodiversity Information Facility (GBIF) (<https://www.gbif.org/>). The dataset has comprised 729 species records from 1,014 georeferenced locations. These records have spanned three taxonomic kingdoms: Animalia (519 species), Plantae (206 species), and Fungi (4 species). Preprocessing for reliability, consistent with GBIF and Darwin Core standards, has been performed, which has included the removal of duplicate records, correction of geospatial inaccuracies, and exclusion of fossil data, enhancing data usability for MBI modeling.
- *Land Use and Land Cover (LULC) Data*, obtained from Italy's National Geoportal (<http://www.pcn.minambiente.it/mattm/servizio-wms/>). This detailed dataset, adhering to CORINE Land Cover standards, has been essential for analyzing biodiversity distribution and mapping key habitats like Mediterranean Temporary Ponds and forests.
- *Environmental Variables* (including PM_{10} concentrations, temperature, atmospheric pressure, and precipitation), collected from ARPA Puglia (Regional Agency for Environmental Protection) monitoring networks (<https://www.arpa.puglia.it>). These data have been collected from a range of monitoring station types, including traffic, industrial, and background sites, located in major urban centers such as Foggia, Bari, Brindisi, Taranto, and Lecce. PM_{10} is known to pose significant anthropogenic pressure, impacting plants and soil. Exposure to fine particulate matter ($PM_{2.5}$ and PM_{10}) negatively affects stomatal function and physiological processes in plants, while also promoting soil acidification, which disrupts microbial communities and nutrient cycling in Mediterranean ecosystems. In Mediterranean regions like Apulia, PM_{10} levels, often affected by African dust intrusions, can significantly influence both soil characteristics and ecosystem functioning. Temperature variations, including long-term warming

trends, contribute to shifts in species distributions, reduced habitat suitability, and increased mortality during extreme heat periods. Elevated temperatures within Mediterranean ecosystems can also affect phenological timing, decrease pollination efficiency, and disrupt plant communities. Biodiversity is influenced by atmospheric pressure through its effects on microclimates, soil moisture retention, and nutrient cycling, particularly in coastal and mountainous areas. The distribution of precipitation is crucial for governing water availability and ecosystem productivity. Reduced rainfall exacerbates drought conditions, negatively affecting soil microbial activity, plant biomass, and water-dependent species.

2.2 *Multilevel Biodiversity Index (MBI) and Spatial Interpolation*

To evaluate spatial biodiversity patterns, the Multilevel Biodiversity Index (MBI) has been computed as an integrated metric combining species richness with land cover diversity, as follows:

$$MBI = (S \times LULC)/A$$

where:

- S represents true species occurrence, reflecting the effective number of existing species.
- $LULC$ denotes the number of land use/land cover diversity classes.
- A is the sampled area in (km^2).

This index standardizes biodiversity values across a spatial grid, making it possible to compare ecologically heterogeneous landscapes. Higher MBI values indicate greater ecological diversity, while lower values suggest limited biodiversity or uniform land cover types. The MBI values were derived from species occurrence data aggregated on a regular grid with dimensions of 15×13 and cell sizes of $19.5 \text{ km} \times 19.5 \text{ km}$. The results have shown elevated MBI values in coastal areas, where species richness and landscape diversity are higher, and reduced values in inland zones characterized by intensive agriculture and industrial activity. Unlike traditional indices, MBI has provided a more holistic view of ecological complexity by integrating land cover diversity and rarity, which has proven especially useful in regions with high environmental variability like Apulia.

For the meteorological and air quality variables, Ordinary Kriging has been used for spatial interpolation. This geostatistical method has been chosen for its ability to account for spatial autocorrelation and produce unbiased, reliable predictions. This process created a uniform prediction grid ($19.5 \text{ km} \times 19.5 \text{ km}$) across the study area, ensuring consistent spatial integration of heterogeneous inputs suitable for

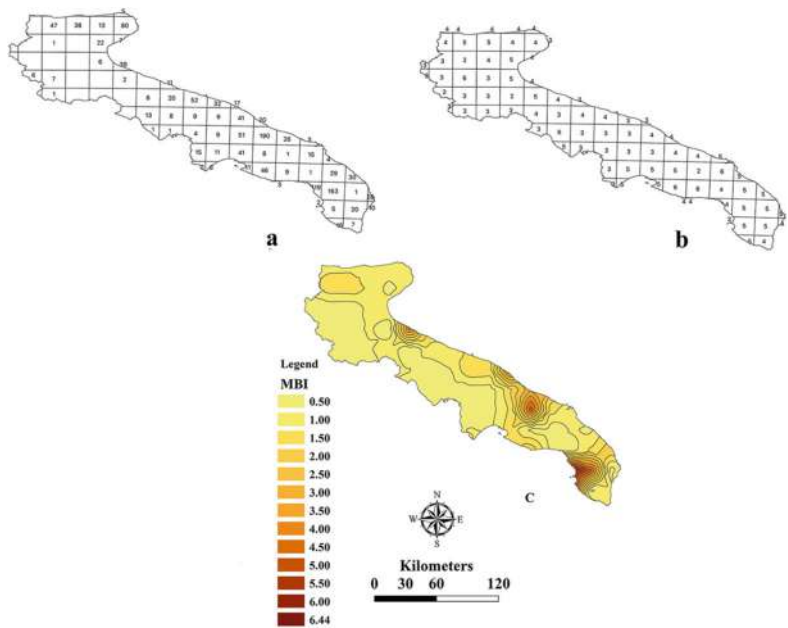


Fig. 1 (a) Number of species occurrence, (b) number of LULC classes, and (c) MBI contour map in the Apulia region

ML applications. Structural analysis, involving variogram estimation and model selection, has been conducted for each variable under study to guarantee spatial continuity and compatibility with the biodiversity modeling.

Figure 1 illustrates the regional spatial patterns of species occurrences, LULC diversity, and the derived MBI.

2.3 Machine Learning Methods

Machine learning methods have increasingly been used in biodiversity studies due to their ability to manage complex ecological datasets, identify nonlinear relationships, and select influential environmental predictors [6, 12]. In this study, four supervised ML algorithms (XGBoost, DT, SVM, and LR) have been applied to model MBI in the Apulia region.

XGBoost is a supervised machine learning algorithm that uses an additive process called gradient boosting to build predictive models from an ensemble of decision trees. It excels at identifying complex, nonlinear relationships and is both scalable and robust to missing data. However, its performance relies heavily on hyperparameter tuning [32], and it can be computationally expensive for very large

datasets [12]. The model has been successfully used in a variety of ecological contexts, such as modeling forest biomass and predicting species richness.

The Decision Tree (DT) algorithm is a nonparametric supervised model that recursively partition data for predictions, valued for interpretability and handling nonlinear interactions [13]. However, DT is prone to overfitting if the tree becomes too complex, which can be mitigated through pruning [22]. In ecology, DTs have been applied to various tasks, from predicting vegetation types to analyzing bird migration.

Support Vector Machine (SVM) is a robust supervised method for classification and regression, effective with high-dimensional, nonlinear data. SVM identifies optimal separating hyperplanes using support vectors and kernel functions [14]. It also has a strong ability to prevent overfitting [43]; however, SVM can be computationally expensive with large datasets and sensitive to parameter tuning. In ecological studies, SVM has been successfully applied to modeling forest diversity, assessing habitat suitability, and predicting species richness [5, 6, 37].

Linear Regression (LR) is a fundamental and efficient statistical method used to model the linear relationship between a dependent variable (like a biodiversity index) and one or more independent variables (like environmental factors) [23]. However, its primary limitation is its inability to effectively model nonlinear relationships and complex ecological patterns, which are very common in biodiversity data. Despite this limitation, LR remains a widely used tool for evaluating how land use impacts species richness or predicting biodiversity changes due to climate variation.

2.4 Model Evaluation Metrics

In this research, the forecasting capability of the machine learning models has been evaluated through the application of several widely used statistical measures. The metrics employed have encompassed the Nash-Sutcliffe Efficiency (*NSE*) [2, 31], the Mean Absolute Percentage Error (*MAPE*) [11, 16, 33, 39], the Root Mean Square Error (*RMSE*) [34], the coefficient of determination (R^2) [3], as well as the Williams plot diagnostics.

Nash-Sutcliffe Efficiency (*NSE*) measures the relative magnitude of residual variance compared to the variance of the observed data. Its values range from $-\infty$ to 1, with 1 denoting a perfect match between observed and predicted values. It is computed as follows:

$$NSE = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2}, \quad (1)$$

where O_i and P_i are observed and predicted values, respectively, n is the number of observations, and $\bar{O}$ is the mean of observed values.

Mean Absolute Percentage Error ($MAPE$) determines the average magnitude of a model's prediction errors as a percentage of the observed values. This provides a clear and consistent way to interpret a model's precision, even when comparing forecasts for different metrics. It is defined as follows:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{O_i - P_i}{O_i} \right| \cdot 100, \quad (2)$$

where n is the number of observations.

Root Mean Square Error ($RMSE$) is a common metric used to measure the average difference between a model's predicted values and the actual observed values. It is often used to quantify a model's accuracy, with lower values indicating a better fit. It is given by

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (O_i - P_i)^2}. \quad (3)$$

Coefficient of determination (R^2) is a statistical measure that evaluates the proportion of variance in observed data explained by the model. This metric is mathematically similar to the NSE . While both are powerful tools for evaluating model performance, they are used in different contexts. A key distinction is that while R^2 is typically used for assessing how well a linear model fits the original data, NSE is more commonly used to evaluate the predictive power of a model on new or simulated data, especially in fields like hydrology. It is expressed as

$$R^2 = 1 - \frac{\sum_{i=1}^n (O_i - P_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2}. \quad (4)$$

R^2 values range from 0 to 1, with higher values indicating better model performance and explanatory power.

Williams plot diagnostics employ a combination of leverage and standardized residuals to evaluate a model's validity and identify outliers. The values for leverage (h_i) and standardized residuals (r_i) are computed as follows:

$$h_i = x_i^T (X^T X)^{-1} x_i, \\ r_i = \frac{\hat{O}_i - P_i}{\sigma},$$

where x_i^T represents the predictor vector for observation i , X is the matrix of predictors, and σ denotes the standard deviation of residuals.

Collectively, these diagnostic tools have offered a complete appraisal of the model's predictive performance, the error magnitude, and its overall reliability.

3 Environmental Variables and Biodiversity Prediction

This section details the spatial interpolation results for environmental variables and the MBI's predictive performance.

3.1 Geostatistical Interpolation of Environmental Indicators

The spatial mapping of several key environmental factors across the Apulia region has been conducted. This has been achieved using Ordinary Kriging, a geostatistical technique widely recognized for its ability to address spatial continuity and provide unbiased estimations in a variety of environmental contexts [38, 40]. In particular, the structural analysis, characterized by the variogram estimation and the model selection, has firstly been carried out for the environmental indicators. According to the spatial modeling results and to ensure spatial consistency with the MBI dataset, Kriging interpolation has been performed on a uniform $19.5 \text{ km} \times 19.5 \text{ km}$ grid.

The resulting interpolated maps (Fig. 2) have highlighted the distinct spatial diversity within the region and have also facilitated the evaluation of environmental drivers that have influenced biodiversity characteristics throughout Apulia.

The generated colormaps are used as inputs for biodiversity modeling as they capture the landscape's essential abiotic patterns. These interpolated outputs provide a continuous spatial representation of key environmental variables (like PM_{10} , atmospheric temperature, atmospheric pressure, and precipitation), which effectively show the region's environmental variability. By applying a uniform grid, the datasets create a consistent foundation for machine learning applications, ensuring all environmental predictors are aligned within a coherent spatial framework. As

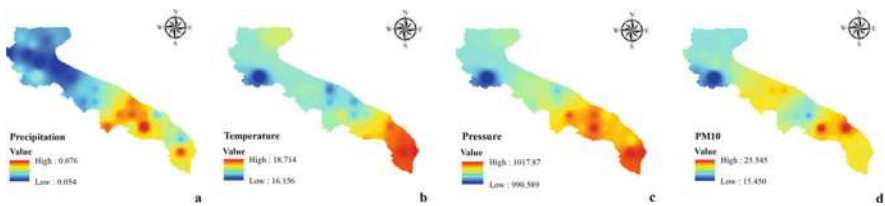


Fig. 2 Colormaps illustrating (a) Precipitation, (b) Temperature, (c) Atmospheric Pressure, and (d) PM_{10} concentrations in Apulia, obtained by applying Ordinary Kriging

shown in Fig. 2, these maps highlight spatial differences and are fundamental to evaluate the influence of environmental factors on biodiversity patterns in Apulia.

Elevated PM_{10} concentrations are found in urban and industrial hubs (e.g., Taranto, Brindisi, and Bari). These areas are correlated with diminished MBI values, suggesting a potential link between atmospheric pollution and ecological stress. Temperature shows both latitudinal and coastal-to-inland variations, with warmer conditions prevalent in the southeastern peninsula and coastal zones, aligning with established Mediterranean climate tendencies and species range shifts. Atmospheric pressure tends to be lower over elevated terrains and more stable across coastal lowlands, reflecting the influence of topography on atmospheric mass. Precipitation exhibits a north-to-south gradient, with higher amounts in the northern Apennine foothills and reduced rainfall in the drier southeastern plains, a critical factor for water-dependent species and ecosystems. These interpolated maps define abiotic gradients directly informing biodiversity modeling. Indeed, by comparing the interpolation results with MBI data, it has been highlighted that regions characterized by elevated particulate matter levels, scarce precipitation, or heightened temperatures have generally displayed reduced biodiversity, underscoring the predictive significance of these factors in spatial ecological risk assessments.

3.2 Model Accuracy and Validation for MBI Prediction

The forecasting accuracy of four machine learning algorithms (XGBoost, DT, SVM, and LR) for MBI estimation in Apulia has been assessed using a multi-metric validation methodology. These models have been trained with fine particulate matter levels, temperature, atmospheric pressure, and precipitation as input features. Model efficacy has been gauged using the NSE and MAPE, complemented by diagnostic tools such as scatterplots comparing predicted versus actual values, cumulative distribution functions of RMSE, Williams plots, and a Taylor diagram [41].

With the highest predictive accuracy (NSE: 0.87, MAPE: 1.08%), the Decision Tree (DT) model has demonstrated strong concordance between observed and predicted MBI values.

XGBoost has also produced commendable performance (NSE: 0.81, MAPE: 3.71%). Conversely, SVM (NSE: 0.54, MAPE: 4.58%) and LR (NSE: 0.16, MAPE: 6.51%) models have exhibited inferior predictive capabilities, indicative of their limited effectiveness in capturing the intricate spatial and nonlinear characteristics of Mediterranean biodiversity.

Williams plot diagnostics for all models have indicated that the majority of standardized residuals fell within the ± 3 range (± 3 times the standard deviation) and within the leverage boundaries, suggesting reliable and generalizable predictions. Outliers might reflect ecotonal regions (e.g., transitions between urban and rural areas or coastal buffer zones) where ecological variability complicates prediction accuracy. A Taylor diagram comparison has shown the DT model achieved the highest correlation coefficient (above 0.95) and a standard deviation

closely aligned with observed data, while XGBoost also performed strongly. SVM and LR models were positioned farther from the reference point, indicating lower predictive accuracy.

Finally, relevancy factor analysis has quantified the contribution of each input variable to the model predictions [21]. Spatial coordinates were identified as the most influential predictors, suggesting the significant influence of underlying spatial gradients and latent spatial heterogeneity not entirely explained by environmental variables alone. Among the atmospheric inputs, pressure and temperature have shown the strongest impact, followed by precipitation and PM_{10} .

4 Conclusions

This study highlights the crucial role of algorithm selection in ecological modeling, particularly for MBI prediction over the Apulia region. A comparative analysis of four supervised machine learning models (XGBoost, DT, SVM, and LR) consistently demonstrated that tree-based methods, specifically Decision Tree (DT) and XGBoost, outperformed linear and kernel-based models across all validation metrics.

The DT model proved to be the most reliable and accurate, exhibiting excellent concordance between observed and predicted MBI values (NSE: 0.87 and MAPE: 1.08%). Its strong performance, supported by diagnostic plots, can be attributed to its capacity to effectively model the complex, nonlinear interactions vital for spatially heterogeneous ecosystems like those found in Apulia. EXGBoost also delivered a strong performance (NSE: 0.81 and MAPE: 3.71%). Conversely, the SVM and LR models showed limited predictive capability, struggling to handle the intricate spatial and environmental variability characteristic of Mediterranean landscapes.

From a conservation standpoint, the accuracy of DT and XGBoost in producing spatially explicit biodiversity maps is paramount for informing adaptive management strategies in sensitive areas. Their ability to integrate complex abiotic gradients, including PM_{10} concentrations and precipitation, highlights their utility in addressing real-world ecological challenges, as these factors are known to influence biodiversity.

In conclusion, this research confirms that carefully selected and validated machine learning methods, particularly tree-based ones, offer powerful tools for MBI prediction and provide a transferable framework for biodiversity modeling in other Mediterranean and data-scarce regions. Future investigations should broaden this approach by incorporating additional environmental descriptors (e.g., vegetation indices, land use shifts, and soil characteristics) and exploring hybrid modeling architectures or temporal data integration to refine predictive accuracy and ecological insight, thereby supporting targeted, evidence-based conservation strategies.

Bibliography

1. Abid, M., González, J.A., de Rivera, Ó.R., Moraga, P.: Mapping the spatio-temporal distribution of burned areas in the Amazon from 2001 to 2020: an ensemble modeling approach. *Environ. Ecol. Stat.* **32**(2), 707–734 (2025)
2. Ahmad, A., Sulaiman, M., Alhindi, A., Aljohani, A.J.: Analysis of temperature profiles in longitudinal fin designs by a novel neuroevolutionary approach. *IEEE Access* **8**, 113285–113308 (2020)
3. Al-Shboul, K.F., Almasabha, G., Shehadeh, A., Alshboul, O.: Exploring the efficacy of machine learning models for predicting soil radon exhalation rates. *Stochastic Environ. Res. Risk Assess.* **37**, 4307–4321 (2023)
4. Antonelli, A., Dhanjal-Adams, K.L., Silvestro, D.: Integrating machine learning, remote sensing and citizen science to create an early warning system for biodiversity. *Plants People Planet* **5**, 307–316 (2023)
5. Asadi, H., Jalilvand, H., Tafazolli, M., Hosseini, S.F.: Modeling habitat suitability of *Quercus castaneifolia* in the Hyrcanian forest: a comprehensive integration of environmental factors for conservation insights. *Biodivers. Conserv.* **34**(1), 315–334 (2025)
6. Bayat, M., Burkhart, H., Namiranian, M., Hamidi, S.K., Heidari, S., Hassani, M.: Assessing biotic and abiotic effects on biodiversity index using machine learning. *Forests* **12**(4), 461 (2021)
7. Beccarisi, L., Zuccarello, V., Accogli, R., Belmonte, G.: Biodiversity and possible bio-indicators of Mediterranean temporary ponds in southern Apulia, Italy. *Diversity* **16**(9), 559 (2024)
8. Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., Courchamp, F.: Impacts of climate change on the future of biodiversity. *Ecol. Lett.* **15**, 365–377 (2012)
9. Cai, R., Xie, S., Wang, B., Yang, R., Xu, D., He, Y.: Wind speed forecasting based on Extreme Gradient Boosting. *IEEE Access* **8**, 175063–175069 (2020)
10. Cazzolla Gatti, R., Notarnicola, C.: A novel multilevel biodiversity index (MBI) for combined field and satellite imagery surveys. *Global Ecol. Conserv.* **13**, e00361 (2018)
11. Chaibi, M., Ben Ghoulam, E.M., Khallouk, N., Tarik, L., El Yousfi, Y., El Hmaidi, A., Berrada, M., Mabrouki, J.: A novel fuzzy-multi-criteria-GIS-machine learning approach for onshore wind power plant site selection. *Euro-Mediterr. J. Environ. Integr.* **10**, 1025–1045 (2025)
12. Chang, G.J.: Biodiversity estimation by environment drivers using machine/deep learning for ecological management. *Eco. Inform.* **78**(10), 102319 (2023)
13. Coops, N.C., Waring, R.H., Beier, C., Roy-Jauvin, R., Wang, T.: Modeling the occurrence of 15 coniferous tree species throughout the Pacific Northwest of North America using a hybrid approach of a generic process-based growth model and decision tree analysis. *Appl. Veg. Sci.* **14**(3), 402–414 (2011)
14. Cortes, C., Vapnik, V.: Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995)
15. De Iaco, S., Hristopulos, D.T., Lin, G.: Special issue: geostatistics and machine learning. *Math. Geosci.* **54**(3), 459–465 (2022)
16. de Myttenaere, A., Golden, B., Le Grand, B., Rossi, F.: Mean absolute percentage error for regression models. *Neurocomputing* **192**, 38–48 (2016)
17. Distefano, V., Palma, M., De Iaco, S.: Multi-class random forest model to classify wastewater treatment imbalanced data. *Soc.-Econ. Plan. Sci.* **95**, 102021 (2024)
18. Distefano, V., De Iaco, S., Masoumi, I.: Artificial neural network optimization to estimate radon in soil. *Stat. Anal. Data Mining ASA Data Sci. J.* **18**(5), e70036 (2025)
19. Distefano, V., Gentile, V., Cucurachi, P.A., De Iaco, S.: Sustainable finance disclosure versus performance: a clustering approach. *Bus. Strateg. Environ.* **34**, 7837–7850 (2025)
20. Farzipour, A., Elmi, R., Nasiri, H.: Detection of monkeypox cases based on symptoms using XGBoost and Shapley Additive Explanations Methods. *Diagnostics* **13**(14), 2391 (2023)

21. Fathinasab, M., Shahbazi, K.: A new correlation for estimation of minimum miscibility pressure (MMP) during hydrocarbon gas injection, *J. Pet. Explor. Prod. Technol.* **10**(6), 2349–2356 (2020)
22. Felix, F.C., Kratz, D., Ribeiro, R., Nogueira, A.C.: Machine learning in the identification of native species from seed image analysis. *J. Seed Sci.* **46**(1), e202446002 (2024)
23. Gómez-Tolosa, M., González-Soriano, E., Mendoza-Cuenca, L.F., Pérez-Mungúa, R.M., Rioja-Paradela, T.M., Espinoza-Medinilla, E.E., Ortega-Salas, H., Rivera-Velázquez, G., Penagos-García, F.E., López, S.: The use of highly diverse clades as a surrogate for habitat integrity analysis: *Argia* damselflies as a practical tool for rapid assessments. *Environ. Sci. Pollut. Res.* **29**, 24334–24347 (2022)
24. Grantz, D.A., Garner, J.H.B., Johnson, D.W.: Ecological effects of particulate matter. *Environ. Int.* **29**(2–3), 213–239 (2003)
25. Grillo, M., Schiaparelli, S., Durazzano, T., Guglielmo, L., Granata, A., Huettmann, F.: Machine learning applied to species occurrence and interactions: the missing link in biodiversity assessment and modelling of Antarctic plankton distribution. *Ecol. Proces.* **13**(1), 56 (2024)
26. Habibullah, M.S., Din, B.H., Tan, S.H., Zahid, H.: Impact of climate change on biodiversity loss: global evidence. *Environ. Sci. Poll. Res.* **29**, 1073–1086 (2022)
27. López Rodríguez, S., van Bussel, L.G.J., Alkemade, R.: Classification of agricultural land management systems for global modeling of biodiversity and ecosystem services. *Agric. Ecosyst. Environ.* **360**, 108795 (2024)
28. Luiselli, L., Pacini, N.: Remote sensing that makes sense in ecological research—from pixels to conservation. *Afr. J. Ecol.* **63**(1), e70002 (2025)
29. Mantyka-pringle, C.S., Martin, T.G., Rhodes, J.R.: Interactions between climate and habitat loss effects on biodiversity: a systematic review and meta-analysis. *Glob. Chang. Biol.* **18**(4), 1239–1252 (2012)
30. Masoumi, I., De Iaco, S., Maggio, S.: Exploring land use and land cover changes in Apulia, Italy: random forest approach utilizing remote sensing data. In: Pollice, A., Mariani, P. (eds.) *Methodological and Applied Statistics and Demography I. SIS 2024*, pp. 259–264. Italian Statistical Society Series on Advances in Statistics. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-64346-0_44
31. McCuen, R.H., Knight, Z., Cutter, A.G.: Evaluation of the Nash–Sutcliffe efficiency index. *J. Hydraul. Eng.* **11**(6), 597–602 (2006)
32. Militino, A.F., Goyena, H., Pérez-Goya, U., Ugarte, M.D.: Logistic regression versus XGBoost for detecting burned areas using satellite images. *Environ. Ecol. Stat.* **31**, 57–77 (2024)
33. Nabillah, I., Ranggadara, I.: Mean Absolute Percentage Error untuk Evaluasi Hasil Prediksi Komoditas Laut. *J. Inf. Syst.* **5**(2), 250–255 (2020)
34. Najafabadipour, A., Kamali, G.R., Nezamabadi-pour, H.: The innovative combination of time series analysis methods for the forecasting of groundwater fluctuations. *Water Resour.* **49**(2), 283–291 (2022)
35. Nel, H., Mishra, A.K., Schonken, F.: EcoSonicML: Harnessing machine learning for biodiversity monitoring in South African Wetlands. *SN Comput. Sci.* **5**(5), 513 (2024)
36. Novi, L., Bracco, A.: Machine learning prediction of connectivity, biodiversity and resilience in the coral triangle. *Commun. Biol.* **5**(1), 1359 (2022)
37. Park, D.S., Willis, C.G., Xi, Z., Kartesz, J.T., Davis, C.C., Worthington, S.: Machine learning predicts large scale declines in native plant phylogenetic diversity. *New Phytol.* **227**, 1544–1556 (2020)
38. Pellicone, G., Caloiero, T., Modica, G., Guagliardi, I.: Application of several spatial interpolation techniques to monthly rainfall data in the Calabria region (Southern Italy). *Int. J. Climatol.* **38**(9), 3651–3666 (2018)
39. Prayudani, S., Hizriadi, A., Lase, Y.Y., Fatmi, Y., Al-Khowarizmi, A.: Analysis accuracy of forecasting measurement technique on random k-nearest neighbor (RKNN) using MAPE and MSE. *J. Phys. Conf. Ser.* **1361**(1), 012089 (2019)
40. Susanto, F., De Souza, P., He, J.: Spatiotemporal interpolation for environmental modelling. *Sensors* **16**(8), 1245 (2016)

41. Taylor, K.E.: Summarizing multiple aspects of model performance in a single diagram. *J. Geophys. Res. Atmos.* **106**(D7), 7183–7192 (2001)
42. Tian, H., Huang, L., Hu, S., Wu, W.: A modified machine learning algorithm for multi-collinearity environmental data. *Environ. Ecol. Stat.* **31**(4), 1063–1083 (2024)
43. Wang, D., Wang, M., Qiao, X.: Support vector machines regression and modeling of green-house environment. *Comput. Electron. Agric.* **66**(1), 46–52 (2009)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits any noncommercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if you modified the licensed material. You do not have permission under this license to share adapted material derived from this chapter or parts of it.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

