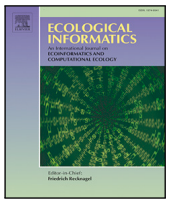
Ecological Informatics 92 (2025) 103514

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Application of association rule mining to assess forest species distribution in Italy considering abiotic and biotic factors

Valeria Aloisi [a], Sergio Noce [b,c], Italo Epicoco [a,c,*], Cristina Cipriano [b,c], Massimo Cafaro [a,c], Giuseppe Brundu [b,d], Lorenzo Arcidiaco [e], Donatella Spano [b,c,d], Giovanni Aloisio [a,c], Simone Mereu [b,c,e]

[a] Department of Engineering for Innovation, University of Salento, Lecce, Italy
[b] National Biodiversity Future Center (NBFC), Palermo, Italy
[c] CMCC Foundation - Euro-Mediterranean Center on Climate Change, Lecce, Italy
[d] Department of agricultural sciences, University of Sassari, Sassari, Italy
[e] Institute of Bioeconomy, National Research Council of Italy (CNR), Sesto Fiorentino (FI), Italy

ARTICLE INFO

Dataset link: https://geodata.ucdavis.edu/climate/worldclim/2_1/base/wc2.1_30s_bio.zip, https://doi.org/10.6084/m9.figshare.7504448.v5, https://data.isric.org/geonetwork/srv/ita/catalog.search#/metadata/14e7c761-6f87-4f4c-9035-adb282439a44, https://code.earthengine.google.com/, langnico.github.io/globalcanopyheight/assets/tile_index.html, https://github.com/CMCC-Foundation/ARM_for_Plants_Distribution

Keywords:
Forest species
Biodiversity
Association rules
Remote sensing
Machine learning

ABSTRACT

Biodiversity monitoring represents a pressing global priority, and assessing forest community composition plays a crucial role due to its influence on ecosystem functions. The spatial distribution of forest species becomes essential for understanding biodiversity dynamics, territorial planning, aiding nature conservation and enhancing ecosystem resilience amid global change. Association Rule Mining, commonly applied to other scientific contexts, is now innovatively adopted in the ecological field to explore the relationships among co-occurring plant species and extract hidden interpretable patterns, also with abiotic and biotic conditions. Multiple heterogeneous data sources were integrated through data preprocessing into a unique dataset, including georeferenced information about 151 plant species monitored within 6,784 plots across Italy and several bioclimatic indices, soil-related factors, and variables from earth observations. The Frequent Pattern Growth algorithm, used for association rule mining, provided interesting and encouraging findings, suggesting ecological rules among plant species and environmental conditions. Indeed, temperature seasonality between 650–700 and precipitation seasonality between 45–50 resulted very correlated with *Picea abies* (confidence = 90.9%, lift = 7.13). Patterns detected for *Picea abies* highlighted its ecological specificity, indicating a strong association with cold, highly seasonal environments, and particular plant communities. Some species appeared acting as community "hubs", frequently co-occurring with other species, suggesting ties to specific environmental or biotic conditions. These findings represent a valuable resource for future research, especially in regions with similar environmental settings and when prior ecological knowledge exists, also underlining the importance of publicly accessible, high-quality ecological data.

1. Introduction

Monitoring biodiversity has become an urgent global priority in light of escalating environmental pressures and accelerating species loss (Díaz et al., 2019). The adoption of the Kunming-Montreal Global Biodiversity Framework and, at the European level, the entry into force of the Nature Restoration Regulation mark a turning point in policy ambition (Ma, 2023; Penca and Tănăsescu, 2025). These frameworks imply the need for accurate, scalable, and timely ecological information to guide restoration efforts, conservation planning, and climate adaptation strategies (Kissling et al., 2018). Forest ecosystems, which host intricate biological communities and deliver essential ecosystem services, are at the forefront of these restoration goals (Hua et al., 2022; Pan et al., 2011).

Building on current global and regional priorities, assessing forest community composition is particularly critical, as it underpins key ecosystem functions and directly influences biodiversity (Bertrand et al., 2011; Brose and Hillebrand, 2016). Understanding the spatial distribution of forest species provides essential information for biodiversity monitoring, territorial planning, and adaptive forest management, thereby supporting nature conservation and ecosystem resilience in the

* Corresponding author at: Department of Engineering for Innovation, University of Salento, Lecce, Italy.
E-mail addresses: valeria.aloisi@unisalento.it (V. Aloisi), sergio.noce@cmcc.it (S. Noce), italo.epicoco@unisalento.it (I. Epicoco), cristina.cipriano@cmcc.it (C. Cipriano), massimo.cafaro@unisalento.it (M. Cafaro), gbrundu@uniss.it (G. Brundu), lorenzo.arcidiaco@cnr.it (L. Arcidiaco), spano@uniss.it (D. Spano), giovanni.aloisio@cmcc.it (G. Aloisio), simone.mereu@cmcc.it (S. Mereu).

https://doi.org/10.1016/j.ecoinf.2025.103514
Received 4 August 2025; Received in revised form 6 November 2025; Accepted 6 November 2025
Available online 8 November 2025

face of global change (Sullivan et al., 2017; Chazdon and Brancalion, 2019).

Understanding and predicting community composition has long been a central goal in ecology, giving rise to the dedicated subfield of community ecology. This discipline focuses on the patterns of species diversity, abundance, and composition within ecological communities, and the underlying processes that shape them. While the conceptual roots of community ecology can be traced back to the pioneering work of Alexander von Humboldt and Frederic Clements in the 19th and early 20th centuries, it was not until the 1950s that quantitative approaches emerged, notably through the development of niche theory and competition models (MacArthur and Levins, 1967).

Species do not exist in isolation; rather, they interact through mechanisms such as competition, neutrality, and facilitation (Thuiller et al., 2009; Elith and Leathwick, 2009). Therefore, predicting community composition cannot rely solely on species' responses to abiotic conditions. These interactions imply that species co-occurrence is not entirely stochastic. Although a general theory of community assembly across space and time is still lacking, ecologists increasingly recognize that community composition results from both deterministic and stochastic processes, often referred to collectively as assembly processes (Gravel et al., 2006). Four primary processes shape these dynamics (Vellend, 2010): selection (differences in survival and reproduction among species), drift (random changes in how common species are), speciation (the emergence of new species), and dispersal (movement across space). Some approaches based on Information Theory models were initially developed by Pàl Juhász-Nagy to analyze the within-community patterns of the number and diversity of species combinations (Juhász-Nagy, 1967, 1976, 1984, 1993; Juhász-Nagy and Podani, 1983). These Juhász-Nagy Pàl models (JNP models) offered a suite of diversity measures (e.g., the Shannon diversity index) designed for representing the fine-scale spatial organization in communities, and they have been thoroughly tested in the past (Tsakalos et al., 2022; Podani et al., 1993; Bartha et al., 1998). Starting from the classical Shannon diversity, a bioinformatic version of it was proposed in Tsakalos et al. (2022) as Compositional Diversity (CD) based on the number and relative abundance of realized (observed) species combinations to characterize the biodiversity and provide information on species coexistence relationships at fine-scale within the sampling unit or within the community. The development of Species Distribution Models (SDMs) in the late 1990s, facilitated by advances in computational capacity and the availability of Geographic Information Systems (GIS), provided tools to quantitatively link species distributions with environmental variables. SDMs, also known as ecological niche models, are now a fundamental component of community ecology. They include a variety of methods, such as Generalized Linear Models (GLMs), Generalized Additive Models (GAMs), climatic envelope models, and more recently, machine learning algorithms (Norberg et al., 2019; Noce et al., 2017, 2019, 2023; Swan et al., 2021; Franklin, 2023; Rathore and Sharma, 2023). A key limitation of SDMs in a community context is that they typically model species independently, assuming that their distribution is solely determined by environmental factors (Guisan and Thuiller, 2005; Watling et al., 2015). However, species interactions are implicitly embedded in presence data, meaning that the resulting models often reflect where species actually live under current conditions (realized niche), rather than all the places they could potentially survive if there were no competitors or other limiting factors (the fundamental niche). Joint Species Distribution Models (JSDMs) represent a major advance by modeling the distribution of multiple species simultaneously. This enables the identification of community-level patterns in species–environment relationships (Ovaskainen et al., 2016). More sophisticated JSDMs, such as the Hierarchical Modeling of Species Communities (HMSC) framework (Ovaskainen et al., 2017), estimate residual species co-occurrence, captured in a residual correlation matrix that accounts for associations unexplained by environmental covariates (Pollock et al., 2014). While Warton et al. (2015) noted that such correlations can also arise from shared unmeasured environmental variables or sampling artifacts, the co-occurrence matrix also offers valuable potential for conditional prediction, that is, estimating the probability of a species' presence given the occurrence of another (Warton et al., 2015). This capacity holds promise for applied ecology, including species detection, monitoring, and management. Most JSDMs are implemented within a Bayesian framework, which is often computationally demanding, restricting their application to small datasets, limited spatial extents, or low-resolution grids. This poses challenges for addressing urgent environmental questions that require high-resolution, large-scale predictions.

To support biodiversity conservation and restoration at scale, it is important to develop modeling frameworks that are not only ecologically comprehensive but also computationally efficient and user-friendly. In recent years, Artificial Intelligence (AI) and Machine Learning (ML) algorithms have been increasingly applied in terrestrial ecology to improve model performance, manage complex datasets, and uncover non-linear patterns that traditional methods may miss (Cipriano et al., 2025).

In light of this, the present study focuses on advancing community-level ecological models that are computationally efficient, transparent, and scalable. Deep Learning methods involving Neural Networks (NNs) are usually adopted for predictive tasks, but these techniques are not preferable when the goal is explainability. Instead, Association Rule Mining (ARM) is a well-established technique within the field of ML, more specifically in unsupervised learning, aimed at discovering hidden and interpretable patterns in large datasets without relying on predefined class labels or target variables (Han et al., 2012). Therefore, an approach based on ARM could enable a clearer interpretation due to its inherent explainability.

ARM has been exploited only on a small number of studies in an ecological context. Specifically, a study explored how plant trait correlations, insularity, and climate may influence plant community assembly on islands, focusing on the Southwest Pacific (Ciarle, 2024). Instead, concerning the assessment of plant species co-occurrence, a relevant study conducted by Silva et al. (2016) exploited ARM techniques, based on the Apriori algorithm, to extract hidden patterns of co-occurrences among 312 forest species monitored on the Barro Colorado Island, Panama. A similar method was then applied in Orozco-Arias et al. (2019) to explore worldwide co-occurrences among 17 species of the genus *Brachypodium*. However, a limitation of the previous two studies is that they did not consider any abiotic or biotic factors that could influence the spatial distribution of the plant species. Another ARM approach was adopted in Pratheepa et al. (2016) to investigate the role of some abiotic factors on the incidence of the insect pest *Helicoverpaarmigera* on cotton crops. The obtained if-then rules revealed that, under specific ranges for temperature and relative humidity, the occurrence of this pest would be high (Pratheepa et al., 2016). Souza et al. (2021) confirmed the feasibility of applying association rule analysis to ecological datasets in tropical forests, highlighting its potential as a tool for detecting patterns of species coexistence and ecosystem functioning. A recent paper (Ghosh et al., 2025) has proposed a method based on association rules to identify threatened, ecologically synergistic mangrove species in India for targeted restoration.

The present study aims to investigate the relationships among species recorded within the same plots, where the biotic factor is intended to encompass both the structural attributes of woody species and the co-occurrence of multiple taxa, together with abiotic conditions. The analysis is based on 151 forest species monitored across 6784 sites in Italy, thereby improving the assessment of forest community distribution. To the best of our knowledge, this is the first study to apply ARM with a more efficient algorithm, the Frequent Pattern Growth (FP-Growth), for automatically extracting hidden patterns of plant co-occurrences, taking into account several climatic, soil-related, and earth observation factors.

2. Methods

2.1. Data sources and database implementation

2.1.1. Plant species data

The Italian National Forest Inventory (INFC), established to meet the United Nations Framework Convention on Climate Change (UNFCCC) requirements, also supports national and international forest reporting, including the Kyoto Protocol and the Food and Agriculture Organization of the United Nations (FAO) assessments. The INFC2015, in accordance with the INFC2005, follows a three-phase design with systematic sampling on a 1 × 1 km grid (Gasparini et al., 2022). Phase 1 involved photo-interpretation of 301,000 points using Coordination of Information on the Environment (CORINE) and FAO classifications. In Phase 2, about 30,000 forest points were field-checked for qualitative data. Phase 3 focused on 7000 plots where quantitative data were collected. These data support statistically sound estimates of about 50 forest variables, mainly for wood and carbon assessments. We derived plot coordinates and basal area from the third phase of the INFC2015. Most of the variables used in this study were calculated within a 13-m diameter subplot (AdS13), which serves as the core sampling area for dendrometric and structural measurements.

The species list with the associated unique identifier (UNICODE) is shown in Supplementary Table 1.

2.1.2. Bioclimatic variables

Data for 20 bioclimatic variables were downloaded from two global datasets, 19 of which are from the WorldClim 2.1 database (Fick and Hijmans, 2017) for the 1970–2000 period and the Aridity index from Version 3 of the Global Aridity index and potential Evapotranspiration database (Zomer et al., 2022). Both datasets have a spatial resolution of 30 arcsec (≈ 1 km). To avoid multicollinearity among the WorldClim variables, a pairwise correlation analysis was conducted across the 19 layers. Based on a correlation threshold of $|r| \geq 0.7$, we excluded highly correlated variables and retained only a subset for subsequent analyses: BIO1 (Annual Mean Temperature), BIO4 (Temperature Seasonality), BIO8 (Mean Temperature of Wettest Quarter), BIO12 (Annual Precipitation), and BIO15 (Precipitation Seasonality). Bioclimatic values were extracted from the above-described sets of raster layers based on the INFC2015 sampling plots, within the ESRI ArcGIS Pro ver 3.4.3 environment, using the Nearest Neighbor (NNb) technique, which assigns to each point the value of the 1-km resolution cell in which the plot falls.

2.1.3. Geopedological variables

Physical and chemical soil properties were extracted from SoilGrids at 250 m resolution (Hengl et al., 2017) released in May 2020 and accessed in March 2025. Variables included in the database were the volume of water content at −1500 kPa $\left(10^{-3}\ \mathrm{cm}^3\ \mathrm{cm}^{-3}\right)$ and pH Water (pH·10) for the 5–15 cm soil depth layer. As in the previous case, values were extracted from raster layers using the INFC2015 sampling plots within the ESRI ArcGIS Pro environment, applying the NNb technique. Lithological information was derived from the recent Italian lithological map provided by Bucci et al. (2022), which offers a standardized classification of rock types across the Italian territory. Lithology data were spatially intersected with sampling plots using the same geoprocessing procedures to ensure consistency with soil variable extraction. A table describing the lithological codes and their corresponding types is provided in the Supplementary Table 2.

2.1.4. Earth observation variables

The quantification of spatiotemporal variations in vegetation dynamics was conducted by extracting key statistical descriptors of the Normalized Difference Vegetation Index (NDVI) using Sentinel-2 surface reflectance level 2 data and the INFC2015 sampling plots.

The analysis was conducted within the Google Earth Engine (GEE) platform (Gorelick et al., 2017), leveraging its scalable cloud-based geospatial processing capabilities to efficiently manage large datasets and perform high-resolution time-series computations. The methodological framework was designed to assess both intra-annual and inter-annual vegetation trends, while minimizing atmospheric noise and accounting for localized spatial variability. The INFC2015 sampling plots dataset was uploaded to GEE as a FeatureCollection, and it was used to extract spectral metrics. To capture seasonal and interannual variations in vegetation activity, the analysis was performed over a seven-year period from 2018 to 2024. For each year, the growing season, typically characterized by peak photosynthetic activity in temperate forest ecosystems, was defined as the period from March to September (both inclusive).

This temporal window was selected to minimize potential confounding effects from winter dormancy or snow cover in mountainous areas. For each year within the specified interval, all available Sentinel-2 Level 2A Surface Reflectance images were acquired from the Copernicus S2_Sr_Harmonized dataset. This harmonized collection ensures radiometric consistency across Sentinel-2A and 2B sensors, making it suitable for multi-year time-series analysis. To mitigate cloud contamination, a cloud-masking procedure was applied based on the Msk_Cldprb band, which provides per-pixel estimates of cloud probability. For each image, pixels with cloud probabilities exceeding 20% were excluded from analysis. This threshold was selected based on a balance between retaining sufficient data for temporal compositing and ensuring the reliability of vegetation index measurements. The cloud mask was applied before NDVI calculation to preserve only high-quality observations. This approach is particularly important in regions with frequent cloud cover, such as northern and mountainous parts of Italy. NDVI was computed for each image using the standard formula:

$$\mathrm{NDVI} = \frac{\mathrm{NIR} - \mathrm{RED}}{\mathrm{NIR} + \mathrm{RED}} \tag{1}$$

where the near-infrared (NIR) reflectance corresponds to band B8 and the red reflectance to band B4 of Sentinel-2 imagery. After cloud-masking, the NDVI images were stacked into yearly collections, filtered by date and spatial extent. For each year, NDVI time series were summarized using four statistical measures: minimum (*NDVI_min*), maximum (*NDVI_max*), median (*NDVI_median*), and standard deviation (*NDVI_stdDev*). These metrics capture key aspects of vegetation behavior: minimum and maximum values represent seasonal extremes, the median provides a measure of central tendency over the growing season, and the standard deviation reflects intra-seasonal variability, which may indicate land cover transitions or phenological dynamics.

In addition to point-based statistics, we sought to understand the local spatial variability in NDVI values surrounding each sampling point. This was achieved through a neighborhood analysis using a 3 × 3 pixel window (corresponding to a spatial extent of 30 m × 30 m at 10 m resolution). Specifically, focal operations were performed on the median NDVI layer to derive local maximum (vic_max), and median (vic_median) values.

Spatial variability was further quantified using a spatial standard deviation (vic_std), computed with a 3 × 3 square kernel through the *reduceNeighborhood GEE function*. This neighborhood-based analysis helps capture spatial heterogeneity, which can be indicative of fragmented vegetation, edge effects, or mixed land cover types within the vicinity of the point. For each year, NDVI summary statistics were extracted at each sampling location using the *reduceRegions* function, which allows batch computation of zonal statistics across multiple points. Pixel-based NDVI values were matched with their corresponding spatial metrics using unique point identifiers. The final feature set for each year included the following attributes: Maximum NDVI 3 × 3 and Standard Deviation of NDVI 3 × 3, representing the highest NDVI

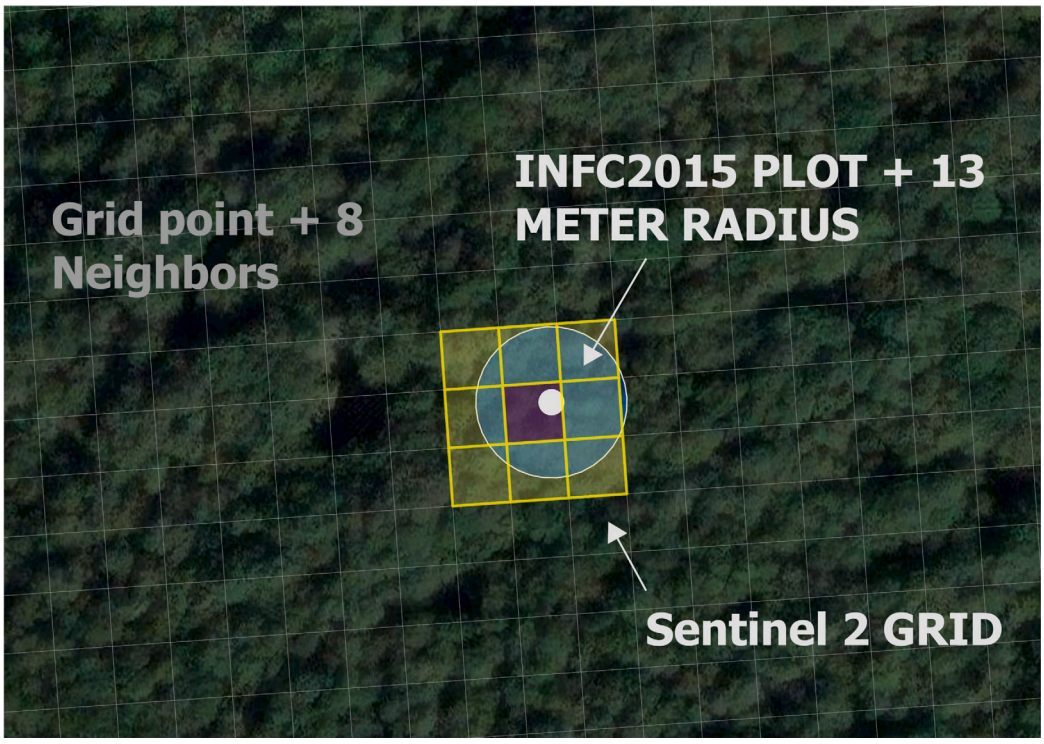


Fig. 1. Earth observation data value calculation.

value and spatial variability within a 3 × 3 pixel neighborhood, respectively. For each sampling point, these values were averaged over the entire temporal series (2017–2024) to obtain a representative metric, which was then integrated into the final database (MDVIME, SDVIME). These metrics offer a spatially explicit view of vegetation conditions, supporting ecological modeling and land condition assessments by capturing local-scale vegetation dynamics influenced by environmental or anthropogenic factors.

Additionally, we used a global high-resolution canopy height dataset to calculate the mean vegetation height and standard deviation within each block unit of the study area. This dataset, developed by Lang et al. (2023), integrates Sentinel-2 imagery with GEDI spaceborne LiDAR data and employs a deep learning model to estimate canopy height globally for the year 2020 at a spatial resolution of 10 m. Given the 13-m radius of the INFC subplot, Earth Observation indicators were calculated by averaging the values within a 3 × 3 cell window centered on the cell corresponding to the INFC point (Fig. 1).

2.1.5. Final database implementation

All INFC2015 plots containing one or more missing values in the bioclimatic or soil variables were excluded. Regarding the Earth Observation data, plots with more than one missing value within the 2017–2024 time series for both the Maximum NDVI and the Standard Deviation NDVI indices were also removed (e.g., if missing data occurred in two or more years, the entire plot was excluded). Starting from a total of 6893 plots in INFC2015, 6784 plots were retained for the final analysis, corresponding to the removal of approximately 1.6% of the original dataset. These selected plots are shown in Fig. 2.

The variable naming convention adopted in this study follows a standardized schema designed to facilitate categorization and interpretation. Each variable code is composed as follows: *X_YYYYYY*, where *X* is a single uppercase letter indicating the thematic category, and *YYYYYY* is a unique identifier for the specific variable. For instance, the prefix *P* refers to plot-level information such as plot ID, geographic coordinates, or structural attributes; *S* identifies soil-related variables including physical and chemical properties; *C* designates climate variables such as temperature and precipitation indices, and *E* is for Earth observation variables.

This structured format ensures consistency across datasets and enhances automated data handling. As examples, the code `P_LATITU` refers to the plot's latitude, `S_SGPHWA` to soil pH, and `C_WC0001` to the annual mean temperature. The implemented final database is described in Supplementary Table 3.

2.2. Association rule mining

Association rule mining is one of the most important and well-researched techniques in data mining (Zhao and Bhowmick, 2003), which represents the core process of the so-called Knowledge Discovery in Database (KDD). The term KDD was initially introduced in a workshop at the International Joint Conference on Artificial Intelligence in Detroit, USA, in 1989 (Zhang and Wu, 2011), then it was defined in 1992 and 1996. According to the widely accepted 1996 definition, KDD represents the *nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data* (Fayyad et al., 1996; Zhang and Wu, 2011).

Association rule mining was first proposed by Agrawal et al. (1993) for market basket analysis (Zhang and Wu, 2011) to identify the sets of items that are frequently bought together at a supermarket by analyzing customer shopping carts (Zaki and Meira, 2020). Then, association analysis found several applications in many other fields such as telecommunication networks, risk management, inventory control, recommendation systems, data classification and clustering, catalog design, healthcare, and loss-leader analysis (Zhao and Bhowmick, 2003; Kumbhare and Chobe, 2014; Shen et al., 2024; Papi et al., 2022; Darrab et al., 2024; Rawat et al., 2023; Versichele et al., 2014; Cagliero et al., 2016).

The goal of ARM is to find interesting relationships, correlations, patterns, and association structures among frequently appearing items in transactional datasets, without implying causality (Zhao and Bhowmick, 2003; Orozco-Arias et al., 2019; Silva et al., 2016; Wu et al., 2008; Tan, 2007).

Fig. 2. Spatial distribution of the 6784 monitoring sites across Italy.

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of elements called *items*. For example, I may represent the collection of all the products (e.g., milk, butter, bread, etc.) sold at a supermarket (Zaki and Meira, 2020). A set $X \subseteq I$ is called *itemset*. A *transaction* corresponds to the set of items in an operation, such as the subset of products purchased by a particular customer in market basket analysis (Silva et al., 2016; Brin et al., 1997). Each transaction is usually associated with a unique identifier called *Transaction IDentifier (TID)* (Zhang and Wu, 2011; Zaki and Meira, 2020). Given a transaction dataset D whose records are transactions over I, each itemset is characterized by a statistical measure called *support*, denoted as *supp*, indicating the number of transactions in D that contain the itemset (Zhang and Wu, 2011; Zaki and Meira, 2020). An itemset X is said to be *frequent* in D if its support is greater than or equal to a *minimum support threshold*, denoted as *minsup* (Zhang and Wu, 2011; Zaki and Meira, 2020). If this threshold is specified as a fraction by users or experts, it must be referred to the so-called *relative support* of an itemset, indicating the fraction of transactions in D that contain the itemset (Zaki and Meira, 2020). Given these premises, an association rule is an implication in the form of $X \rightarrow Y$, where X and Y are disjoint itemsets, that is $X, Y \subseteq I$ and $X \cap Y = \varnothing$ (Zhao and Bhowmick, 2003; Zhang and Wu, 2011; Zaki and Meira, 2020). In this case, X is called *antecedent*, whereas Y is named *consequent* (Zhao and Bhowmick, 2003). Therefore, an association rule detects a relationship between the itemsets X and Y, thus providing information in the form of if-then statements (Zhang and Wu, 2011). For example, in the context of market basket analysis, association rules may detect that 'bread' and 'butter' are frequently brought together, suggesting that if a customer buys bread, then they may also purchase butter (Kumbhare and Chobe, 2014). There are three key metrics for evaluating the quality of the association rules: *support*, *confidence*, and *lift*. Specifically, given an association rule $X \rightarrow Y$ with $X, Y \subseteq I$ and $X \cap Y = \varnothing$, the *support* of the rule is the number of transactions in which both X and Y co-occur as subsets (Zaki and Meira, 2020):

$$supp(X \rightarrow Y) = supp(X \cup Y) \tag{2}$$

Thus, the *relative support* of the rule, denoted as *rsupp*, is the fraction of transactions containing $X \cup Y$ with respect to the total number of transactions in D (Zaki and Meira, 2020):

$$rsupp(X \rightarrow Y) = \frac{supp(X \cup Y)}{total\ number\ of\ transactions} \tag{3}$$

The *confidence* of the rule is defined as the conditional probability that a transaction contains Y given that it contains X (Zaki and Meira, 2020). Hence, it can be computed by dividing the number of transactions that contain $X \cup Y$ by the total number of transactions that contain X (Zhao and Bhowmick, 2003):

$$conf(X \rightarrow Y) = P(Y|X) = \frac{P(X \cap Y)}{P(X)} = \frac{supp(X \cup Y)}{supp(X)} \quad (4)$$

Therefore, the confidence of an association rule provides information about the strength of that rule; for example, if a rule has a confidence of 80%, it means that 80% of the transactions that contain X also contain Y together (Zhao and Bhowmick, 2003).

The *lift*, also known as *interest*, of an association rule can be considered a measure of its importance (Silva et al., 2016), and it is defined as follows (Zaki and Meira, 2020):

$$lift(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X) * P(Y)} = \frac{rsupp(X \cup Y)}{rsupp(X) * rsupp(Y)} = \frac{conf(X \rightarrow Y)}{rsupp(Y)} \quad (5)$$

Specifically, a lift of 1 reveals that the antecedent and the consequent of the corresponding rule are independent (Silva et al., 2016), whereas if the lift is greater than or less than one, this suggests that the antecedent and the consequent itemsets have a positive or a negative correlation, respectively (Silva et al., 2016). Moreover, the lift metric provides a deeper insight into the relationships between association rules by quantifying whether the predictive power of a rule exceeds what would be expected by chance (Shen et al., 2024). Indeed, it allows distinguishing the significant associations (high lift values) from those due to random variations in the data (low lift values).

ARM consists in extracting those association rules that satisfy predefined minimum support and confidence at the same time (Zhang and Wu, 2011; Zhao and Bhowmick, 2003; Agrawal and Srikant, 1994; Kumbhare and Chobe, 2014). This condition is usually known as the *supp-conf framework* (Zhang and Wu, 2011; Agrawal et al., 1993). Consequently, the ARM problem is usually decomposed into two subproblems (Zhao and Bhowmick, 2003): *(i)* detecting the itemsets with a support greater than, or equal to, the predefined minimum support threshold (i.e., identifying all the frequent itemsets) (Zhang and Wu, 2011); *(ii)* given the frequent itemsets, generating those association rules that have a confidence exceeding the minimum confidence threshold (Zhang and Wu, 2011; Zhao and Bhowmick, 2003). In this way, the ARM becomes a two-step process (Zhang and Wu, 2011). The second sub-problem is straightforward once the first one is solved (Zhao and Bhowmick, 2003), hence several approaches were developed in recent years to address it. Starting from the first algorithm, AIS (from Agrawal, Imielinski, and Swami who proposed it Agrawal et al., 1993), the Apriori algorithm was conceived (Agrawal and Srikant, 1994). It is based on a level-wise search for mining the frequent itemsets, where those containing k items are exploited to explore those including $k + 1$ items (Kumbhare and Chobe, 2014). This searching mechanism is known as the *candidate generation process* (Kumbhare and Chobe, 2014). Although the Apriori algorithm represented an improvement with respect to previous approaches, it had two main drawbacks: *(i)* the complexity of the candidate generation process which consumes a lot of resources in terms of time, space, and memory; and *(ii)* the fact that it requires multiple scans of the dataset (Zhao and Bhowmick, 2003; Kumbhare and Chobe, 2014). To overcome these bottlenecks, tree-structured approaches were investigated (Zhao and Bhowmick, 2003). The FP-Growth algorithm was first proposed by Han and Pei (2000). It is based on the construction of the Frequent Pattern Tree (FP-Tree) (Han et al., 2000), a compressed and more efficient representation of the dataset. Hence, the procedure is composed of two steps: the construction of the FP-Tree and the generation of the frequent patterns from the FP-tree (Kumbhare and Chobe, 2014). This algorithm requires only two scans over the dataset for creating the FP-Tree. The first scan is used to compute the F-List, the list of frequent items sorted by frequency in descending order, whereas the second pass is needed to compress the dataset in the FP-Tree (Kumbhare and Chobe, 2014). Consequently, the problem of extracting the most frequent itemsets is converted to searching and constructing trees recursively (Kumbhare and Chobe, 2014) without any candidate generation process. This way, FP-Growth proved to be more efficient than the previous algorithms, being an order of magnitude faster than Apriori (Zhao and Bhowmick, 2003; Kumbhare and Chobe, 2014).

Table 1
Example with three transactions and eight species.

ID plot	Items (plant species)
1	LARDEC, PICABI, PINCEM, PINSYL
2	ALNINC, FRAEXC
3	BETPEN, PICABI, POPTRE

2.3. Data preparation

As described in Section 2.1, several heterogeneous data sources related to various factors that may influence the forest species distribution were integrated through data preprocessing steps into a unified dataset. This dataset included information about the presence or absence of 151 plant species within 6784 monitoring plots across Italy. Each plot was characterized by the geographical coordinates (i.e., latitude and longitude) and a unique identifier, named *ID plot*. More than one species could be detected in a specific plot. To apply the ARM procedure, the dataset was transformed to achieve the transactional format, where each record represents a transaction. Specifically, in this case, the set I of all the items corresponds to the set of all the 151 plant species and a transaction over I can be thought as a subset of plant species detected in a single plot. Therefore, the transaction dataset presented a record for each plot tracking the list of those species found in that plot, as depicted in Table 1.

Moreover, several abiotic and woody species structural factors were extracted, preprocessed, and associated with each plot based on the geographical coordinates. The abiotic variables included five climatic drivers (i.e., annual mean temperature, temperature seasonality, mean temperature of the wettest quarter, annual precipitation, and precipitation seasonality) and four soil-related factors: aridity index, lithology, soil water content (SGWC33), and soil pH (SGPHWA). Instead, the following features derived from Earth observations were considered as biotic components: ETH Global Height, NDVI Max (temporally averaged over 2017–2024), and NDVI Std Dev (temporally averaged over 2018–2024).

Due to the continuous nature of these factors, a categorization procedure was needed to proceed with the association analysis. The criterion used to categorize each variable is reported in Table 2. Then, a one-hot encoding step was performed. Consequently, the final dataset consisted of 6784 records corresponding to plots and 433 boolean columns, of which 151 indicated the presence or absence of each species, whereas the remaining 282 were derived from the categorization of the abiotic and biotic features. Therefore, in this dataset, the transaction described by a record indicates those plant species detected in a plot, together with the abiotic and biotic conditions associated with that plot.

The ARM procedure was performed on this dataset by using the Python package *mlxtend*, and the minimum support threshold set in the FP-Growth algorithm was equal to 0.01. Setting a higher threshold could restrict the analysis to only the predominantly common species (Silva et al., 2016), potentially overlooking ecologically relevant but less common patterns. Conversely, the chosen minimum support value enabled the identification of a higher number of most frequent itemsets from which to extract rarer and more interesting patterns. Moreover, a threshold of 0.07 was set on the confidence metric to filter the resulting association rules, thus ensuring a high level of precision and reliability.

Table 2
Categorization of the continuous abiotic and biotic factors.

Variable	Code	Categorization
Abiotic factors		
Annual mean temperature (°C)	Bio1	40 classes for each 0.5 °C change
Temperature seasonality (standard deviation × 100)	Bio4	7 classes for each 50 units
Mean temperature of the wettest quarter (°C)	Bio8	47 classes for each 0.5 °C change
Annual precipitation (mm)	Bio12	52 classes for each 50 mm change
Precipitation seasonality (Coefficient of variation)	Bio15	14 classes for each 5 units
Aridity index	ARIIND	20 classes in 5% percentiles
Soil WC33 ($10^{-2} \times cm^3 \times cm^{-3}$) × 10	SGWC33	20 classes at evenly spaced intervals
Soil PHWA (pH × 10)	SGPHWA	5 classes at evenly spaced intervals
Lithology (already categorical)		–
Biotic factors		
ETH Global Height	GLH	21 classes for each 2 m change
NDVI Max (temporally averaged over 2017–2024)	MDVIME	20 classes in 5% percentiles
NDVI Std Dev (temporally averaged over 2018–2024)	SDVIME	20 classes in 5% percentiles

3. Results

The FP-Growth algorithm was applied to the transaction dataset and identified 3925 most frequent itemsets with a relative support greater than or equal to 0.01. These itemsets included plant species, abiotic, and biotic categorical classes as items. Then, the ARM procedure produced a total of 15,548 association rules and 2783 out of them had consequent itemsets including only plant species. These resulting 2783 association rules are shown in Fig. 3 with respect to the support, confidence, and lift metrics. Specifically, Panel A presents two 2D scatterplots with each rule plotted by its support or confidence against lift. Panel B provides a 3D scatterplot presenting a comprehensive visualization of the association rules according to the three metrics at once, with rule color indicating lift. As emerges from this figure, higher support values correspond to less interesting association rules (i.e., low lift values), whereas support values close to the threshold result in non-trivial and more interesting association patterns (i.e., very high lift values).

Not all 2783 association rules are strong enough and relevant, hence a confidence threshold equal to 0.7 was set to assure high precision and reliability. Thus, 44 association rules with a confidence greater than or equal to 0.7 resulted and are reported in Table 3.

For example, the first association rule $PHILAT \rightarrow QUIELE$ highlights that 80.4% of the transactions that contain *Phillyrea latifolia* also contain *Quercus ilex* together, thus suggesting an interesting co-occurrence pattern. Moreover, it is noteworthy that the antecedent itemsets of these rules often include not only plant species but also abiotic and biotic categorical factors. Consequently, these association rules provide information concerning the abiotic and biotic conditions under which specific plant species are likely to be present with a certain level of confidence, both individually and in association with other ones. For instance, rule no.3 suggests there exists a probability of nearly 71% that, in the presence of *Arbutus unedo* (*ARBUNE*) and with a soil condition in terms of pH water between 61.2 and 67.8, also *Quercus ilex* (*QUEILE*) will be present.

Additionally, several rules indicate very strong relationships between environmental variables and specific species. For example, Rule no.4, with a confidence of 90.9% and a lift of 7.13, reveals a highly significant ecological pattern: the combination of Bio4 (temperature seasonality between 650.0–700.0) and Bio15 (precipitation seasonality between 45.0–50.0) is very correlated with the presence of *Picea abies* (*PICABI*). More broadly, *PICABI* appears as a consequent in 15 different rules, consistently showing high confidence values (ranging from 72.3% to 90.9%) and lift values above 5.5, confirming strong associations with both climatic variables (e.g., Bio1, Bio4, Bio15) and co-occurring taxa such as *Larix decidua* (*LARDEC*) and soil properties SGPHWA. In contrast, association rules involving *Quercus pubescens* (*QUEPUB*) show moderately strong patterns, with lower lift values (mostly between 2.6 and 2.9), although still above the minimum threshold of interest. These rules tend to include more diverse and complex antecedent itemsets, often involving multiple bioclimatic variables (e.g., Bio4, Bio15), soil properties (e.g., SGPHWA, GLH), and other species such as *Fraxinus ornus* (*FRAORN*) and *Ostrya carpinifolia* (*OSTCAR*).

Focusing on the lift metric, all association rules in Table 3 resulted in being important, showing lift values greater than 1.2, which is generally recognized as a threshold for identifying interesting rules. Specifically, their lift ranges from 2.6 to 8.2, thus highlighting those rules that are deemed valuable.

Figs. 4–6 provide a graph-based visualization of association rules, which is especially useful when exploring moderately-sized rule sets, as it highlights the structure and strength of relationships between items. Specifically, Fig. 4 depicts the association rules where *Picea abies*(*PICABI*) is the consequent species (rules no. 4–18 in Table 3), whereas Figs. 5 and 6 show those that have *Quercus pubescens* (*QUEPUB*) and *Ostrya carpinifolia* (*OSTCAR*) as the consequent item, respectively.

4. Discussion

The present study proposed an innovative application of a traditional data mining technique, usually employed in other scientific contexts, to the ecological and biodiversity monitoring fields for assessing forest community distribution. Indeed, association rule mining was adopted to uncover interpretable patterns among plant species, abiotic conditions, biotic factors, and co-occurrence dynamics, providing insights into the ecological "rules" governing species presence and interactions.

Although association rules should not be considered as causal implications due to their inherent probabilistic nature, the proposed approach provided interesting and encouraging findings, suggesting interpretable "if-then" rules among plant species, biotic and abiotic drivers.

Despite the large number of species present in the dataset, only a limited subset appears as consequents in the association rules with confidence values exceeding the threshold, as if the rules were polarized around these species. Ecologically, this suggests that certain species may act as key "hub" species within the community, exhibiting strong and frequent co-occurrence patterns with multiple antecedents. These species might represent dominant or ecologically influential taxa whose presence is strongly linked to particular environmental conditions or biotic interactions.

Another relevant pattern emerging from the results is the presence of ecologically coherent species groups. The association rules mined highlight a clear tendency for some broadleaf species to co-occur in a structured and ecologically consistent way. Notably, *Quercus pubescens* (QUEPUB) and *Acer opalus* (ACEOPA) are often associated with *Ostrya carpinifolia* (OSTCAR), as shown by the rule QUEPUB, ACEOPA → OSTCAR, suggesting a meaningful ecological relationship between

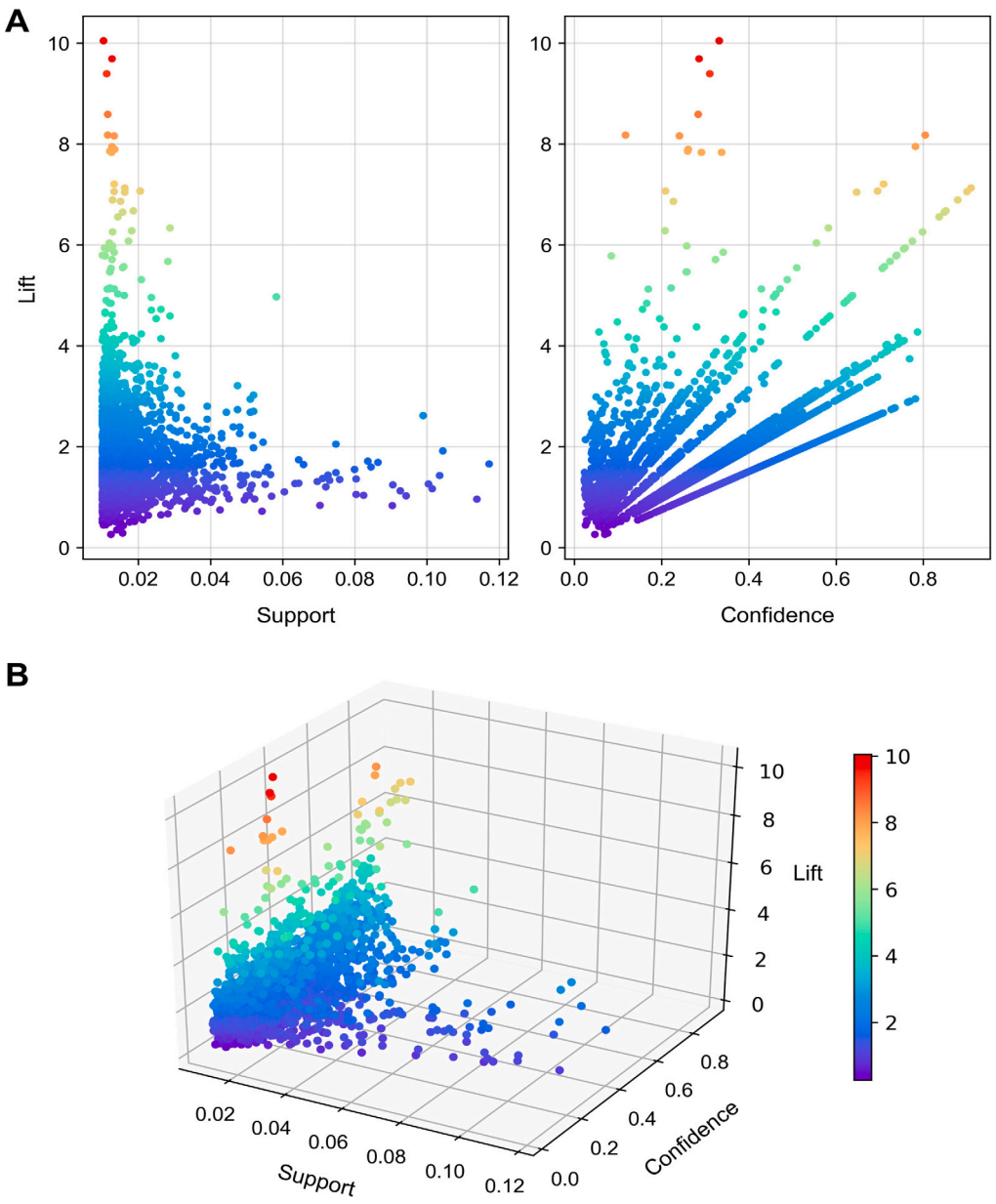


Fig. 3. Distribution of 2783 association rules based on support, confidence, and lift. Panel A provides bidimensional visualizations of the association rules, with the support or confidence values on the *x*-axis and the lift on the *y*-axis. Panel B displays the association rules according to the three metrics simultaneously in a three-dimensional scatterplot.

these species. Similarly, *Fraxinus ornus* (FRAORN) and ACEOPA also tend to co-occur with OSTCAR, as indicated by the rule FRAORN, ACEOPA → OSTCAR. Other relevant combinations, such as FRAORN with specific climatic conditions and lithological settings (carbonate rocks), reinforce the central role of OSTCAR in these associations. Altogether, these findings point to a recurring assemblage of species sharing both ecological preferences and spatial distribution patterns, likely reflecting the structure of temperate deciduous forests in the study area. The patterns identified for *Picea abies* (*PICABI*) emphasize the ecological specificity of this species, which seems closely tied to cold and marked seasonal climatic variation, as well as particular plant communities. The comparatively lower lift may suggest that *Quercus pubescens* (*QUEPUB*) is more ecologically generalist or able to thrive across a wider range of conditions.

Beyond pattern recognition, ARM provides insights into potential ecological mechanisms such as niche overlap, habitat filtering, and facilitative or competitive interactions. For instance, the recurring co-occurrence of mesophilous deciduous species under similar abiotic conditions may reflect habitat filtering, while associations between shade-tolerant and light-demanding species might suggest successional or facilitative dynamics. These interpretations highlight how ARM can support the formulation of ecological hypotheses to be further tested with mechanistic or process-based models.

The present study has several strengths. First, it adopts a more computationally efficient algorithm, FP-Growth, for ARM among plant species compared to previous studies (Zhao and Bhowmick, 2003; Kumbhare and Chobe, 2014). Consequently, hidden patterns (e.g. observed co-occurrence) are automatically extracted in a few seconds from a large amount of ecological data. This rapid processing capability is particularly valuable when dealing with ecosystems or forest communities for which limited prior ecological knowledge is available, allowing the generation of preliminary hypotheses and guiding further ecological investigation. This makes the proposed approach a

Table 3
Results. Association rules with support ≥0.01 and confidence ≥0.7, sorted in descending order by lift values.

No	Antecedent	Consequent	Support	Confidence	Lift
1	PHILAT	QUEILE	0.011	0.804	8.179
2	ARBUNE, bio4 (550.0–600.0)	QUEILE	0.013	0.782	7.952
3	ARBUNE, SGPHWA (61.2–67.8)	QUEILE	0.013	0.709	7.208
4	bio4 (650.0–700.0), bio15 (45.0–50.0)	PICABI	0.016	0.909	7.130
5	SGPHWA (48.0–54.6), LARDEC	PICABI	0.013	0.900	7.058
6	ABIALB, bio4 (650.0–700.0)	PICABI	0.013	0.879	6.892
7	bio4 (650.0–700.0), LARDEC	PICABI	0.019	0.851	6.677
8	LARDEC, bio15 (40.0–45.0)	PICABI	0.016	0.848	6.651
9	bio15 (45.0–50.0), LARDEC	PICABI	0.014	0.836	6.558
10	bio4 (600.0–650.0), SGPHWA (48.0-54.6)	PICABI	0.013	0.798	6.260
11	bio4 (650.0–700.0), bio15 (40.0-45.0)	PICABI	0.017	0.775	6.077
12	bio4 (650.0–700.0), SGPHWA (48.0-54.6)	PICABI	0.011	0.758	5.944
13	bio1 (5.0–5.5) °C	PICABI	0.011	0.755	5.921
14	bio1 (4.5–5.0) °C	PICABI	0.010	0.739	5.797
15	FAGSYL, LARDEC	PICABI	0.011	0.737	5.783
16	SGPHWA (48.0–54.6)	PICABI	0.028	0.723	5.674
17	Lithology class 12, LARDEC	PICABI	0.016	0.711	5.572
18	GLH (32.0–34.0)	PICABI	0.012	0.706	5.536
19	QUEPUB, ACEOPA	OSTCAR	0.010	0.787	4.275
20	bio1 (7.5–8.0) °C, bio4 (600.0–650.0)	FAGSYL	0.010	0.742	4.170
21	FRAORN, bio4 (700.0–750.0), SGPHWA (54.6–61.2)	OSTCAR	0.010	0.756	4.107
22	FRAORN, bio4 (700.0–750.0), Lithology class 12	OSTCAR	0.012	0.752	4.089
23	FRAORN, ACEOPA	OSTCAR	0.012	0.743	4.041
24	FRAORN, bio1 (10.5–11.0) °C	OSTCAR	0.014	0.742	4.033
25	Lithology class 12, bio1 (8.0–8.5) °C	FAGSYL	0.011	0.717	4.030
26	bio4 (600.0–650.0), bio1 (8.0–8.5) °C	FAGSYL	0.013	0.714	4.015
27	FRAORN, SGPHWA (54.6–61.2), Lithology class 12	OSTCAR	0.011	0.730	3.968
28	bio4 (650.0–700.0), ACEOPA	OSTCAR	0.012	0.718	3.903
29	FRAORN, ACEPSE	OSTCAR	0.013	0.713	3.876
30	FRAORN, FAGSYL	OSTCAR	0.013	0.703	3.822
31	bio4 (700.0–750.0), OSTCAR, QUEPUB	FRAORN	0.011	0.768	3.742
32	GLH (10.0–12.0), bio4 (650.0–700.0)	QUEPUB	0.010	0.782	2.949
33	bio15 (20.0–25.0), OSTCAR, bio4 (650.0–700.0), SGPHWA (67.8–74.4)	QUEPUB	0.010	0.769	2.902
34	FRAORN, OSTCAR, bio4 (650.0-700.0), SGPHWA (67.8–74.4)	QUEPUB	0.011	0.765	2.888
35	FRAORN, bio15 (20.0-25.0), OSTCAR, bio4 (650.0–700.0)	QUEPUB	0.012	0.764	2.883
36	GLH (12.0–14.0), bio4 (650.0–700.0), SGPHWA (67.8–74.4)	QUEPUB	0.011	0.735	2.772
37	FRAORN, bio15 (20.0–25.0), bio4 (650.0–700.0), SGPHWA (67.8–74.4)	QUEPUB	0.010	0.732	2.762
38	FRAORN, OSTCAR, SGPHWA (67.8–74.4)	QUEPUB	0.019	0.726	2.740
39	FRAORN, GLH (12.0–14.0)	QUEPUB	0.014	0.708	2.670
40	FRAORN, SGWC33 (355.1–363.8), bio4 (650.0–700.0)	QUEPUB	0.011	0.705	2.659
41	ACEMON	QUEPUB	0.010	0.704	2.657
42	bio15 (20.0–25.0), OSTCAR, SGPHWA (67.8–74.4)	QUEPUB	0.013	0.704	2.656
43	FRAORN, bio15 (20.0–25.0), bio4 (650.0–700.0)	QUEPUB	0.020	0.704	2.655
44	FRAORN, bio15 (20.0–25.0), SGPHWA (67.8–74.4)	QUEPUB	0.016	0.701	2.644

valid alternative or support, in terms of memory and processing time, to more traditional modeling techniques such as species distribution models (SDMs) and joint species distribution models (JSDMs), which are typically used for analyzing individual or community-level species responses to environmental gradients.

Unlike SDM or other approaches, which require model specification, assumptions about species–environment relationships, and often complex fitting procedures, the ARM methodology enables the rapid identification of co-occurrence patterns or associations with ecological drivers, without requiring strong assumptions. However, it is essential to emphasize that this approach is not intended as a replacement for these methods, but rather as a complementary tool in an integrated framework. If used alongside SDMs and JSDMs, ARM could provide useful insights into multispecies interactions and assemblage patterns that may be overlooked in species-by-species analyses.

It is important to note that the primary aim of this study was to test ARM as a novel analytical tool for ecological datasets, rather than to perform a full methodological comparison with other approaches. A comprehensive comparison with frameworks such as SDMs or JSDMs will be addressed in future research. Such a comparison is inherently complex because ARM is purely statistical and descriptive, focusing on uncovering association patterns rather than predicting species occurrences or estimating causal effects. The novelty of ARM lies in its ability to reveal previously undetected ecological relationships and multispecies patterns, offering a complementary perspective to conventional predictive models. Moreover, several environmental and earth observation variables were considered in the mining process as potential abiotic and biotic drivers, allowing the identification of ecological conditions that may favor the presence of a specific plant species assemblage. This contributes to a better understanding of species–environment relationships and can support the extrapolation of ecological patterns to other regions with similar environmental settings, thus helping the design of biodiversity conservation and management strategies. This aspect is especially useful when prior knowledge about the presence of certain species in a similar area is available, hence making it possible to hypothesize other species that are likely to be present as well. Furthermore, the interpretability of association rules and their potential to be transferred across different ecological contexts represent a unique strength.

However, there are also some limitations. Indeed, the present study does not provide causality patterns, but probabilistic association findings strictly related to the transaction dataset on which the ARM procedure is performed. This is due to the fact that the key metrics (i.e., support, confidence, and lift) adopted to evaluate and filter the association rules are by definition conditioned on the number and type of plant species transactions detected across Italy. Consequently, the strength and the relevance of a resulting association rule should not be interpreted in absolute terms, but it is inherently linked to the working transaction dataset. Therefore, this clearly highlights the crucial importance of developing increasingly comprehensive ecological datasets, including large volumes of data with high variability in

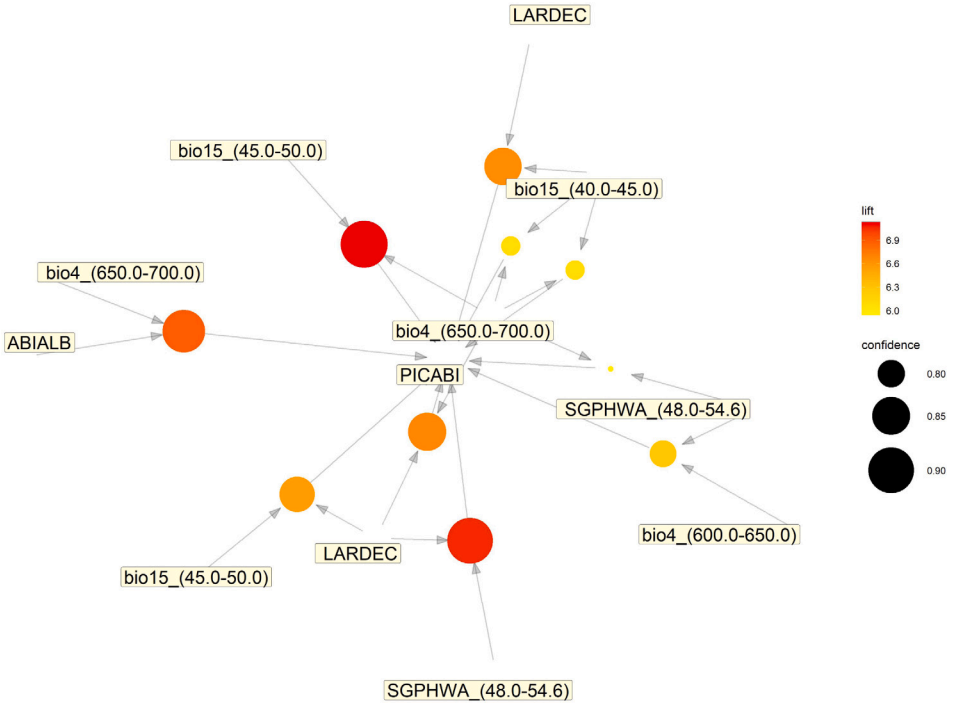


Fig. 4. Graph-based visualization of association rules with confidence ≥0.7 presenting *Picea abies* (PICABI) as the consequent species. Round nodes represent association rules and text nodes correspond to the items. Arrows indicate the direction from antecedent items to the rule node, and from the rule node to the consequent item at the center of the graph. The size and the color of a rule node reflect its confidence and lift, respectively.

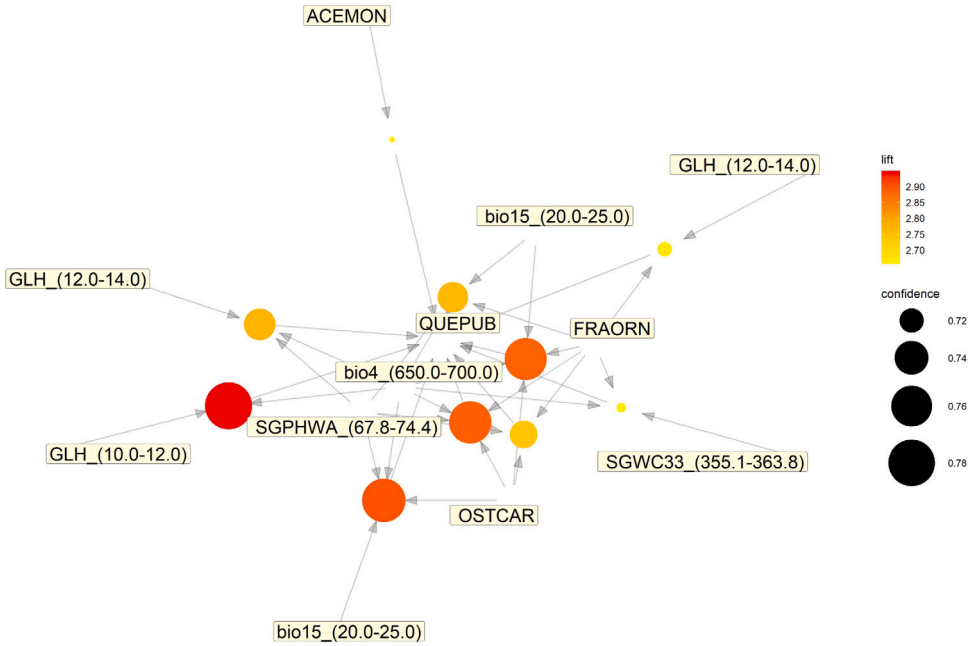


Fig. 5. Graph-based visualization of association rules with confidence ≥0.7 presenting *Quercus pubescens* (QUEPUB) as the consequent species. Round nodes represent association rules and text nodes correspond to the items. Arrows indicate the direction from antecedent items to the rule node, and from the rule node to the consequent item at the center of the graph. The size and the color of a rule node reflect its confidence and lift, respectively.

plant species, to be as representative as possible of the various existing ecosystems and thereby improve the generalizability of the results. Moreover, the proposed approach does not provide a quantitative estimate of the influence of each abiotic or biotic factor on the presence and the abundance of plant species, therefore making it challenging to obtain a numerical measure of its specific contribution to the observed ecological pattern. Furthermore, the application of this approach should always be guided by experts with solid ecological knowledge, as many of the associations identified may be statistically significant yet lack ecological meaning. Expert interpretation plays a crucial role in evaluating the validity and relevance of the outputs, ensuring that the resulting patterns are ecologically sound and contextually appropriate.

In conclusion, this study contributes to a better understanding of plant species distribution and assemblage by offering useful insights

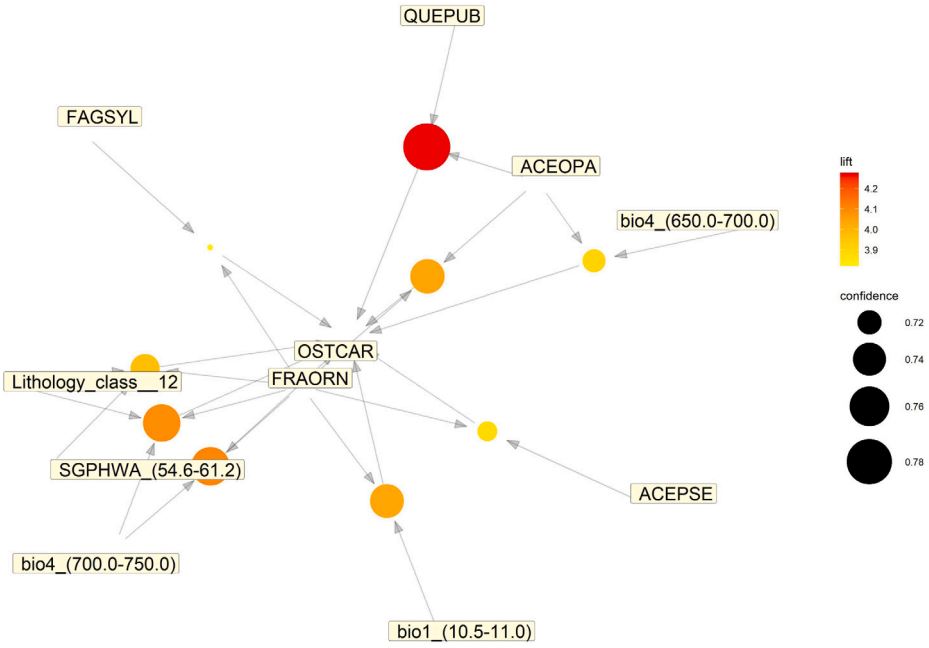


Fig. 6. Graph-based visualization of association rules with confidence ≥0.7 presenting *Ostrya carpinifolia* (OSTCAR) as the consequent species. Round nodes represent association rules and text nodes correspond to the items. Arrows indicate the direction from antecedent items to the rule node, and from the rule node to the consequent item at the center of the graph. The size and the color of a rule node reflect its confidence and lift, respectively.

and identifying meaningful association patterns among plant species and their abiotic and biotic drivers. The proposed approach proves effective in addressing the growing demand for automatic knowledge extraction from large, complex, and often noisy ecological datasets. Notably, its rapid processing capabilities enable the generation of a general overview of forest community structure and biodiversity patterns, highlighting dominant associations and potential ecological hubs.

Although challenges remain, particularly concerning the handling of ecological data complexity and the need for solid measures of rule significance and ecological relevance, the findings presented here represent a useful resource for future research exploring species interactions and community structure, especially in contexts where prior ecological knowledge is available. These results also underscore the importance of publicly accessible, high-quality ecological data repositories for advancing biodiversity research and conservation planning.

Future applications of this approach could extend association rules to identify multi-species assemblages as consequents, moving beyond single-species predictions to capture complex co-occurrence patterns. Moreover, although the current methodology is inherently linked to the spatial sampling design of the underlying forest dataset (INFC2015), there is considerable potential to incorporate spatial scale explicitly. For instance, plots could be aggregated using regular grid frameworks (e.g., fishnet polygons) of varying cell sizes, or proximity-based filters could be applied to account for local neighborhood effects. Additionally, analyzing patterns within distinct biogeographical regions could provide valuable insights into region-specific species interactions and community structures. Such GIS-based and biogeographically informed strategies would enhance the ecological interpretability of association patterns and provide a more comprehensive understanding of biodiversity dynamics across heterogeneous landscapes.

CRediT authorship contribution statement

Valeria Aloisi: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sergio Noce:** Writing – review & editing, Writing – original draft, Visualization, Validation, Formal analysis, Data curation. **Italo Epicoco:** Writing – review & editing, Supervision, Methodology, Investigation, Conceptualization. **Cristina Cipriano:** Writing – review & editing, Validation. **Massimo Cafaro:** Writing – review & editing, Methodology, Conceptualization. **Giuseppe Brundu:** Writing – review & editing, Conceptualization. **Lorenzo Arcidiaco:** Writing – review & editing, Data curation. **Donatella Spano:** Writing – review & editing, Conceptualization. **Giovanni Aloisio:** Writing – review & editing, Conceptualization. **Simone Mereu:** Writing – review & editing, Validation, Conceptualization.

Funding

This research was partially funded under the National Recovery and Resilience Plan (NRRP), Italy, Mission 4 Component 2 Investment 1.4 – Call for tender No 3138 of 16 December 2021, rectified by Decree n.3175 of 18 December 2021 of Italian Ministry of University and Research funded by the European Union – NextGenerationEU. Project code CN_00000033, Concession Decree No 1034 of 17 June 2022 adopted by the Italian Ministry of University and Research, CUP C83C22000550007, Project title "National Biodiversity Future Center – NBFC".

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank the Comando Unità Forestale, Ambientale e Agroalimentare (CUFAA) of the Arma dei Carabinieri for full access to the INFC 2015 data.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ecoinf.2025.103514.

Data availability

INFC 2015 data: access upon consent by the Comando Unità Forestale, Ambientale e Agroalimentare (CUFAA) of the Arma dei Carabinieri.

WorldClim 2.1 database: https://geodata.ucdavis.edu/climate/worldclim/2_1/base/wc2.1_30s_bio.zip

Version 3 of the Global Aridity index and potential Evapotranspiration database: https://doi.org/10.6084/m9.figshare.7504448.v5

Soil-related data: SoilGrids at https://data.isric.org/geonetwork/srv/ita/catalog.search#/metadata/14e7c761-6f87-4f4c-9035-adb282439a44

Lithology information: Italian lithological map provided by Bucci et al. (2022).

NDVI data: derived using a multi-year Sentinel-2 imagery available in the Google Earth Engine GEE archive https://code.earthengine.google.com/, (Gorelick et al., 2017).

Global high-resolution canopy height dataset: langnico.github.io/globalcanopyheight/assets/tile_index.html

Software repository: https://github.com/CMCC-Foundation/ARM_for_Plants_Distribution.

References

Agrawal, R., Imielinski, T., Swami, A., 1993. Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD International Conference on Management of Data. ACM, pp. 207–216.

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB, Morgan Kaufmann, pp. 487–499.

Bartha, S., Podani, J., et al., 1998. Exploring plant community dynamics in abstract coenostate spaces. Abstr. Bot. 22, 49–66.

Bertrand, R., Lenoir, J., Piedallu, C., Riofrío-Dillon, G., De Ruffray, P., Vidal, C., Pierrat, J.-C., Gégout, J.-C., 2011. Changes in plant community composition lag behind climate warming in lowland forests. Nature 479 (7374), 517–520.

Brin, S., Motwani, R., Silverstein, C., 1997. Beyond market baskets: Generalizing association rules to correlations. ACM SIGMOD Rec. 26 (2), 265–276.

Brose, U., Hillebrand, H., 2016. Biodiversity and ecosystem functioning in dynamic landscapes.

Bucci, F., Santangelo, M., Fongo, L., Alvioli, M., Cardinali, M., Melelli, L., Marchesini, I., 2022. A new digital lithological map of Italy at the 1: 100 000 scale for geomechanical modelling. Earth Syst. Sci. Data 14 (9), 4129–4151.

Cagliero, L., Cerquitelli, T., Chiusano, S., Garza, P., Ricupero, G., Xiao, X., 2016. Modeling correlations among air pollution-related data through generalized association rules. In: Proceedings of the 2016 IEEE International Conference on Smart Computing. SMARTCOMP, IEEE, pp. 1–6. http://dx.doi.org/10.1109/SMARTCOMP.2016.7501707.

Chazdon, R., Brancalion, P., 2019. Restoring forests as a means to many ends. Science 365 (6448), 24–25.

Ciarle, R., 2024. Trait correlation and the assembly of island plant communities: Evidence from the southwest Pacific. J. Veg. Sci. 35 (4), e13291. http://dx.doi.org/10.1111/jvs.13291.

Cipriano, C., Noce, S., Mereu, S., Santini, M., 2025. Algorithms going wild–a review of machine learning techniques for terrestrial ecology. Ecol. Model. 506, 111164.

Darrab, S., Broneske, D., Saake, G., 2024. Exploring the predictive factors of heart disease using rare association rule mining. Sci. Rep. 14 (1), 18178. http://dx.doi.org/10.1038/s41598-024-69071-6.

Díaz, S., Settele, J., Brondízio, E.S., Ngo, H.T., Agard, J., Arneth, A., Balvanera, P., Brauman, K.A., Butchart, S.H., Chan, K.M., et al., 2019. Pervasive human-driven decline of life on earth points to the need for transformative change. Science 366 (6471), eaax3100.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40 (1), 677–697.

Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From data mining to knowledge discovery: an overview. In: Advances in Knowledge Discovery and Data Mining. AAAI Press / The MIT Press, pp. 1–34.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37 (12), 4302–4315.

Franklin, J., 2023. Species distribution modelling supports the study of past, present and future biogeographies. J. Biogeogr. 50 (9), 1533–1545.

Gasparini, P., Di Cosmo, L., Floris, A., De Laurentis, D., 2022. Italian national forest inventory—methods and results of the third survey: Inventario nazionale delle foreste e dei serbatoi forestali di carbonio—metodi e risultati della terza indagine. Springer Nature.

Ghosh, M., Mondal, S., Fajriyah, R., Mondal, K.C., Roy, A., 2025. Association of IUCN-threatened Indian mangroves: A novel data-driven rule filtering approach for restoration strategy. Ecol. Informatics 103164.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

Gravel, D., Canham, C.D., Beaudet, M., Messier, C., 2006. Reconciling niche and neutrality: The continuum hypothesis. Ecol. Lett. 9 (4), 399–409. http://dx.doi.org/10.1111/j.1461-0248.2006.00884.x, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-33645353062&doi=10.1111%2fj.1461-0248.2006.00884.x&partnerID=40&md5=d0818206ee9c309c391639aace17ad2a.

Guisan, A., Thuiller, W., 2005. Predicting species distribution: offering more than simple habitat models. Ecol. Lett. 8 (9), 993–1009.

Han, J., Kamber, M., Pei, J., 2012. Data mining: Concepts and techniques, 3rd Morgan Kaufmann Publishers.

Han, J., Pei, J., 2000. Mining frequent patterns by pattern-growth: Methodology and implications. ACM SIGKDD Explor. Newsl. 2 (2), 14–20.

Han, J., Pei, J., Yin, Y., 2000. Mining frequent patterns without candidate generation. In: Chen, W., Naughton, J.F., Bernstein, P.A. (Eds.), Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. ACM Press, pp. 1–12.

Hengl, T., Mendes de Jesus, J., Heuvelink, G.B., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M.N., Geng, X., Bauer-Marschallinger, B., et al., 2017. SoilGrids250m: Global gridded soil information based on machine learning. PLoS One 12 (2), e0169748.

Hua, F., Bruijnzeel, L.A., Meli, P., Martin, P.A., Zhang, J., Nakagawa, S., Miao, X., Wang, W., McEvoy, C., Peña-Arancibia, J.L., et al., 2022. The biodiversity and ecosystem service contributions and trade-offs of forest restoration approaches. Science 376 (6595), 839–844.

Juhász-Nagy, P., 1967. On association among plant populations i. Acta Biologica Debrecina 5, 43–56.

Juhász-Nagy, P., 1976. Spatial dependence of plant populations. Part 1. Equivalence analysis (an outline for a new model). Acta Bot. Acad. Sci. Hung. 22, 61–78.

Juhász-Nagy, P., 1984. Spatial dependence of plant populations. Part 2. a family of new models. Acta Bot. Hung. 30, 363–402.

Juhász-Nagy, P., 1993. Notes on compositional diversity. Hydrobiologia 249, 173–182.

Juhász-Nagy, P., Podani, J., 1983. Information theory methods for the study of spatial processes and succession. Vegetatio 51, 129–140.

Kissling, W.D., Ahumada, J.A., Bowser, A., Fernandez, M., Fernández, N., García, E.A., Guralnick, R.P., Isaac, N.J., Kelling, S., Los, W., et al., 2018. Building essential biodiversity variables (EBV s) of species distribution and abundance at a global scale. Biological Rev. 93 (1), 600–625.

Kumbhare, T.A., Chobe, S.V., 2014. An overview of association rule mining algorithms. Int. J. Comput. Sci. Inf. Technol. (IJCSIT) 5 (1), 927–930.

Lang, N., Jetz, W., Schindler, K., Wegner, J.D., 2023. A high-resolution canopy height model of the earth. Nat. Ecol. Evol. 7 (11), 1778–1789.

Ma, K., 2023. Kunming-montreal global biodiversity framework: An important global agenda for biodiversity conservation. Biodivers. Sci. 31 (4), 23133.

MacArthur, R., Levins, R., 1967. The limiting similarity, convergence, and divergence of coexisting species. Amer. Nat. 101 (921), 377–385. http://dx.doi.org/10.1086/282505.

Noce, S., Caporaso, L., Santini, M., 2019. Climate change and geographic ranges: The implications for Russian forests. Front. Ecol. Evol. 7 (MAR), 57–57.

Noce, S., Cipriano, C., Santini, M., 2023. Altitudinal shifting of major forest tree species in Italian mountains under climate change. Front. For. Glob. Chang. 6, 1250651.

Noce, S., Collalti, A., Santini, M., 2017. Likelihood of changes in forest species suitability, distribution, and diversity under future climate: The case of southern europe. Ecol. Evol. 7 (22), 9358–9375.

Norberg, A., Abrego, N., Blanchet, F.G., Adler, F.R., Anderson, B.J., Anttila, J., Araújo, M.B., Dallas, T., Dunson, D., Elith, J., et al., 2019. A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. Ecol. Monograph. 89 (3), e01370.

Orozco-Arias, S., nez Rincón, A.M.N., Tabares-Soto, R., López-Álvarez, D., 2019. Worldwide co-occurrence analysis of 17 species of the genus brachypodium using data mining. PeerJ 6, e6193. http://dx.doi.org/10.7717/peerj.6193.

Ovaskainen, O., Roy, D.B., Fox, R., Anderson, B.J., 2016. Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. Methods Ecol. Evol. 7 (4), 428–436. http://dx.doi.org/10.1111/2041-210X.12502, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84951752779&doi=10.1111%2f2041-210X.12502&partnerID=40&md5=5ed48c25d33dd8387b82a21de4074591 Cited by: 154; All Open Access, Green Open Access, Hybrid Gold Open Access.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., Abrego, N., 2017. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecol. Lett. 20 (5), 561–576.

Pan, Y., Birdsey, R.A., Fang, J., Houghton, R., Kauppi, P.E., Kurz, W.A., Phillips, O.L., Shvidenko, A., Lewis, S.L., Canadell, J.G., et al., 2011. A large and persistent carbon sink in the world's forests. Science 333 (6045), 988–993.

Papi, R., Attarchi, S., Boloorani, A.D., Samany, N.N., 2022. Knowledge discovery of middle east dust sources using apriori spatial data mining algorithm. Ecol. Informatics 72, 101867. http://dx.doi.org/10.1016/j.ecoinf.2022.101867.

Penca, J., Tănăsescu, M., 2025. The transformative potential of the EU's Nature Restoration Law. Sustain. Sci. 20 (2), 643–647.

Podani, J., Csontos, P., Onda, Z., Bartha, S., 1993. Pattern, area and diversity: the importance of spatial scale in species assemblages. Abstr. Bot. 17, 37–51.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P.A., Mccarthy, M.A., 2014. Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). Methods Ecol. Evol. 5 (5), 397–406. http://dx.doi.org/10.1111/2041-210X.12180, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84900561067&doi=10.1111%2f2041-210X.12180&partnerID=40&md5=74b92176e7a5d6b343e13faefd380fb9, Cited by: 492.

Pratheepa, M., Verghese, A., Bheemanna, H., 2016. Weighted association rule mining for the occurrence of the insect pest helicoverpa armigera (hübner) related with abiotic factors on cotton. In: Proceedings of the 3rd International Conference on Computing for Sustainable Global Development. INDIACom, New Delhi, India, pp. 1884–1887.

Rathore, M.K., Sharma, L.K., 2023. Efficacy of species distribution models (SDMs) for ecological realms to ascertain biological conservation and practices. Biodivers. Conserv. 32 (10), 3053–3087.

Rawat, R., Chakrawarti, R.K., Raj, A.S.A., Mani, G., Chidambarathanu, K., Bhardwaj, R., 2023. Association rule learning for threat analysis using traffic analysis and packet filtering approach. Int. J. Inf. Technol. 15 (6), 3245–3255. http://dx.doi.org/10.1007/s41870-023-01353-0.

Shen, K., Tian, Y., Hu, B., Luo, J., Qi, S., Chen, S., Lin, H., 2024. Association rule mining of air quality through an improved apriori algorithm: A case study in 244 Chinese cities. Trans. GIS 28 (4), 726–745. http://dx.doi.org/10.1111/tgis.13156.

Silva, L.A.E., Siqueira, M.F., dos Santos Pinto, F., Barros, F.S.M., Zimbrão, G., Souza, J.M., 2016. Applying data mining techniques for spatial distribution analysis of plant species co-occurrences. Expert Syst. Appl. 43, 250–260. http://dx.doi.org/10.1016/j.eswa.2015.08.031.

Souza, C.R., Maia, V.A., Aguiar-Campos, N., Farrapo, C.L., Santos, R.M., 2021. Tree species consistent co-occurrence in seasonal tropical forests: an approach through association rules analysis. For. Syst. 30 (2), e006–e006.

Sullivan, M.J., Talbot, J., Lewis, S.L., Phillips, O.L., Qie, L., Begne, S.K., Chave, J., Cuni-Sanchez, A., Hubau, W., Lopez-Gonzalez, G., et al., 2017. Diversity and carbon storage across the tropical forest biome. Sci. Rep. 7 (1), 39102.

Swan, M., Le Pla, M., Di Stefano, J., Pascoe, J., Penman, T.D., 2021. Species distribution models for conservation planning in fire-prone landscapes. Biodivers. Conserv. 30 (4), 1119–1136.

Tan, P.-N., 2007. Introduction to Data Mining, first ed. Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

Thuiller, W., Lafourcade, B., Engler, R., Araújo, M.B., 2009. BIOMOD–a platform for ensemble forecasting of species distributions. Ecography 32 (3), 369–373.

Tsakalos, J.L., Chelli, S., Campetella, G., Canullo, R., Simonetti, E., Bartha, S., 2022. Comspat: an R package to analyze within-community spatial organization using species combinations. Ecography 7, e06216. http://dx.doi.org/10.1111/ecog.06216.

Vellend, M., 2010. Conceptual synthesis in community ecology. Q. Rev. Biol. 85 (2), 183–206. http://dx.doi.org/10.1086/652373, URL: https://www.scopus.com/inward/record.uri?eid=2-s2.0-77952803493&doi=10.1086%2f652373&partnerID=40&md5=ff65fadd5cd3af4022148ed3f3879de5, Cited by: 1934.

Versichele, M., Groote, L.D., Boùùaert, M.C., Neutens, T., Moerman, I., de Weghe, N.V., 2014. Pattern mining in tourist attraction visits through association rule learning on bluetooth tracking data: A case study of ghent, Belgium. Tour. Manag. 44, 67–81. http://dx.doi.org/10.1016/j.tourman.2014.02.009.

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., Hui, F.K., 2015. So many variables: joint modeling in community ecology. Trends Ecol. Evolut. 30 (12), 766–779.

Watling, J.I., Brandt, L.A., Bucklin, D.N., Fujisaki, I., Mazzotti, F.J., Romanach, S.S., Speroterra, C., 2015. Performance metrics and variance partitioning reveal sources of uncertainty in species distribution models. Ecol. Model. 309, 48–59.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., et al., 2008. Top 10 algorithms in data mining. Knowl. Inf. Syst. 14 (1), 1–37.

Zaki, M.J., Meira, Jr., W., 2020. Data Mining and Machine Learning: Fundamental Concepts and Algorithms, second ed. Cambridge University Press, http://dx.doi.org/10.1017/9781108564175.

Zhang, S., Wu, X., 2011. Fundamentals of association rules in data mining and knowledge discovery. WIREs Data Min. Knowl. Discov. 1 (2), 97–116. http://dx.doi.org/10.1002/widm.10.

Zhao, Q., Bhowmick, S.S., 2003. Association Rule Mining: A Survey. Technical Report 135., Nanyang Technological University, Singapore, p. 18.

Zomer, R.J., Xu, J., Trabucco, A., 2022. Version 3 of the global aridity index and potential evapotranspiration database. Sci. Data 9 (1), 409.