Ecological Indicators xxx (xxxx) xxx

Contents lists available at ScienceDirect

Ecological Indicators

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

Estimation of missing Ellenberg Indicator Values for tree species in South-eastern Europe: a comparison of methods

Letizia Leccese [a], Giuliano Fanelli [b], Vito Emanuele Cambria [b], Marco Massimi [b], Fabio Attorre [b], Marco Alfò [a], Svetlana Aćić [c], Erwin Bergmeier [d], Andraž Čarni [e], Mirjana Cuk [f], Renata Custerevska [g], Panayotis Dimopoulos [h], Petrit Hoda [i], Alfred Mullaj [i], Urban Šilc [e], Zeljko Skvorc [j], Zvjezdana Stancic [k], Zora Dajic Stevanovic [c], Rossen Tzonev [l], Kiril Vassilev [m], Luca Malatesta [b,*], Michele De Sanctis [b]

[a] Department of Statistical Science, Sapienza University of Rome, Italy
[b] Department of Environmental Biology, Sapienza University of Rome, Italy
[c] Faculty of Agriculture, University of Belgrade, Serbia
[d] Department of Vegetation Analysis & Phytodiversity, University of Göttingen, Germany
[e] Research Centre of the Slovenian Academy of Science and Arts, Slovenia
[f] Faculty of Sciences, University of Novi Sad, Serbia
[g] Institute of Biology, Faculty of Sciences, Ss. Cyril and Methodius University in Skopje, Macedonia
[h] Laboratory of Botany, Department of Biology, University of Patras University Campus, 26504 Rio, Greece
[i] University of Tirana, Faculty of Natural Sciences, Albania
[j] Faculty of Forestry, University of Zagreb, Croatia
[k] Faculty of Geotechnical Engineering, University of Zagreb, Croatia
[l] Department of Ecology and Environmental Protection, Sofia University "St. Kliment Ohridski", Bulgaria
[m] Institute of Biodiversity and Ecosystem Research, Bulgarian Academy of Sciences, Bulgaria

ARTICLE INFO

Keywords:
Vegetation ecology
Plant indicators
Vegetation databases
Biodiversity informatics
Bioindication
Missing values

ABSTRACT

Ellenberg indicator values (EIV) are widely used in vegetation ecology, but the values for many species in Southeastern Europe are not available due to incomplete knowledge of their ecology: it is therefore of paramount importance to estimate missing values in existing databases. The entire EIV set for a single species can be missing or a single EIV can be missing for species for which other indicator values are available. Our aim here is to provide a simple method to impute missing values for species who have missing data in a single or multiple EIV. For this purpose, we adopt a multiple imputation procedure and compare a number of imputation methods on the basis of two datasets: i) "indices", the set of 9 Ellenberg indicators taken from literature, available for 10,824 species and ii) "vegetation", a set describing the physical and climatic characteristics (Light, Temperature, Continentality, Soil moisture, Nitrogen, Soil pH, Hemeroby index, Humidity, Organic_matter) of 29,935 relevés from Southeastern Europe where at least one tree species is present. The imputation methods we considered are: *k-Nearest Neighbour, multiple linear regression (with or without collinearity correction), Reprediction Algorithm, Weighted Averaging (WA) and Weighted Averaging Partial Least Squares (WAPLS) regression.* The different methods of imputation were compared by looking at the output produced and its deviation from the "true" observed values for a set of species with known EIVs. We have considered a set of species with known EIVs and proceeded to multiple imputation using the methods above; as a measure of performance we adopted the mean squared error (MSE) estimate, and expert judgement of ecological consistency. Models based on Regression and k-Nearest

* Corresponding author.
E-mail addresses: letizia.leccese@hotmail.it (L. Leccese), giuliano.fanelli@uniroma1.it (G. Fanelli), vitoemanuele.cambria@uniroma1.it (V.E. Cambria), marco.massimi@uniroma1.it (M. Massimi), fabio.attorre@uniroma1.it (F. Attorre), marco.alfo@uniroma1.it (M. Alfò), acic@agrif.bg.ac.rs (S. Aćić), erwin.bergmeier@bio.uni-goettingen.de (E. Bergmeier), carni@zrc-sazu.si (A. Čarni), mirjana.cuk@dbe.uns.ac.rs (M. Cuk), pdimopoulos@upatras.gr (P. Dimopoulos), urban@zrc-sazu.si, urban@zrc-sazu.si (U. Šilc), skvorc@sumfak.hr (Z. Skvorc), zvjezdana.stancic@gfv.hr, zvjezdana.stancic@kr.t-com.hr (Z. Stancic), zsatovic@agr.hr (Z. Dajic Stevanovic), rossentzonev@abv.bg (R. Tzonev), kiril5914@abv.bg (K. Vassilev), luca.malatesta@uniroma1.it (L. Malatesta), michele.desanctis@uniroma1.it (M. De Sanctis).

https://doi.org/10.1016/j.ecolind.2024.111851
Received 23 October 2023; Received in revised form 12 February 2024; Accepted 4 March 2024

Neighbour seem to outperform the others. On the contrary, Reprediction algorithm in its different forms: produced less satisfactory results.

Imputation of missing values is generally based on expert knowledge or on some variant of weighted averaging (also known as Hill's method). Here we show that other methods may be more effective and should be appropriately considered by vegetation scientists, since those may allow the application of EIVs in other biogeographic regions.

1. Introduction

Indicator-based classification is one of the most common approaches to concisely and effectively explicit ecosystems' complexity (Diekmann, 2003; Pignatti et al., 2005; Berg et al., 2017). The advantage of indicator values for plant ecologists is that plants can be seen as 'environmental sentinels' in biomonitoring studies, thus assessing the main factors shaping the ecosystem, the interrelationship among species, and the weight and significance of different threats to the environment and biodiversity (Pignatti et al., 2001; Müller and Burkhard, 2012). One of the most widespread bioindication methods in plant ecology is the Ellenberg indicator values system (EIV, Ellenberg et al., 2001; Bertelheimer and Poschlod, 2015). In this system, each plant is characterized by seven indicators (light, temperature, nitrogen or soil fertility, soil moisture, soil reaction, salt, climatic continentality) with values ranging from 1 to 9 (Ellenberg, 1979); indicators identify species-specific score along seven gradients that are considered fundamental for the life of plants. In some extensions of the system, especially in mediterranean environments (Pignatti et al., 2005), the values of some indicators range from 1 to 12. Although certain limitations exist (Zelený and Schaffers, 2012), the system is quick and reliable. Limitations mainly concern the lack of information on the range width of species under investigation and the potential presence of interpretation errors in identifying ecological factors.

The original system of indicators comprised the seven above-mentioned indicators. Other indexes have been added successively, among others the index of hemeroby, measuring the degree of human disturbance on ecosystems according to a ten-point scale (Kowarik, 1990; Fanelli and De Lillis, 2004). Kowarik (1990) introduced this index by calculating the presence and abundance of species in different types of environments. The lowest value on the scale (0) represents pristine environments that hardly exist today in Europe, the highest value (9) represents completely altered artificial habitats. Other indexes have been proposed to complete the original set; for instance, in the French database CATMINAT http://philippe.julve.pagesperso-orange.fr/catminat.htm air humidity (in addition to soil humidity) and soil organic matter were added. In the Landolt system aeration of soil (Durchlüftung) was added. It must be stressed that the indicators represent gradients derived on the field by observation of the species distribution, and that they can correspond to very different ecosystem properties. For instance, Schaffers and Sýkora (2000) and Schaffers and Sýkora (2002) have shown that the nitrogen index is not related to the content of mineral nitrogen in the soil.

Unfortunately, in most databases, the EIVs are not available for many species: for instance, among the 169 tree species in Southeastern Europe, 26 are not referenced to in Dengler et al. 2023.

Efforts have been made over the years to extend the system to other European countries mainly by means of expert judgement (e.g. Pignatti et al., 2005; Dengler et al., 2023; Domina et al., 2018). In a few cases, the problem has been approached by statistical techniques, such as the so-called '*reprediction algorithm*' proposed by Hill et al. (2000). This approach consists in estimating missing data, associated to species that are not present in Ellenberg's original list, by means of values obtained by weighted averages of EIV from species present in a given vegetation or floristic database with known EIVs. Fanelli et al. (2006) applied a variation of Hill's method to extend the Ellenberg indices towards southern Europe and Tichý et al. (2023) used a variation of the Hill algorithm to estimate EIVs for several European species. However, these works have only scratched the complexity of missing data estimation, a subject that has received strong attention in the statistical literature in other contexts (Little and Rubin, 2022; Nugroho et al., 2021; Tsai et al., 2018). Indeed, although multiple imputation is well established in the medical and social fields (e.g. Austin et al., 2021; Hughes et al., 2019; Van Buuren, 2018), its application in ecology is still limited (Nakagawa and Freckleton 2008). Missing data can influence and bias the results of an analysis, especially when their 'presence' is overlooked. Facing missing data by improving existing methods or developing new ones is one of the most debated topics in the recent statistical literature (Austin et al., 2021; Chemolli and Pasini, 2008; Schafer and Graham, 2002). While, as noted by Molenberghs et al. (2008), we cannot make inference on the mixing data mechanism, if not in the very simple MCAR (Missing Completely At Random) framework, and while the multiple imputation approach is based on a MAR (Missing At Random) hypothesis, the use of a precise statistical method for imputation may be appropriate for conducting a post-estimation sensitivity analysis showing how much the obtained estimates are function of specific working hypotheses.

The aim of this study is twofold: 1) estimating indicators for tree species of southeastern Europe and 2) discuss the comparison of a set of imputation methods for estimating missing EIVs.

We employed the test statistics proposed in literature (Kamshidian et al., 2014; Little, 1988) to test for the MCAR framework and employed different imputation techniques and algorithms to discuss their applicability in the context of vegetation science. Our research question is therefore if algorithms frequently used in vegetation ecology are sound and consistent or if they can be replaced by more effective (and general-purpose) approaches. In addition to filling the gaps in existing EIVs datasets, the results of this study may be useful to support the application of approaches based on EIVs outside Europe, where such indicators have never been proposed.

2. Materials and methods

The workflow followed in this research included these steps:

a) We built a database of relevés comprising tree species of southeastern Europe (see section 2.1). EIVs were present for a large majority of the species included in the relevés, while few were absent.
b) We calculated the missing values with different imputation methods (see section 2.2).
c) We assessed the effectiveness of the imputation techniques using the Mean Square Error, computed by comparison of the values obtained from imputation with known EIVs from our initial dataset (see section 2.3).
d) We compared the resulting EIVs dataset with the values presented in recent literature (Dengler et al. 2023, see section 2.4).

2.1. Dataset

We considered only tree species since their ecology is well known and easier to interpret. Two datasets were used for the following analyses. The first ('indices') contains a list of species and the corresponding EIV obtained from two sources: the Prototype of Ecological Flora by Species of Central-Southern Italy (Fanelli et al., 2006), where EIVs for

several Mediterranean species are reported. This dataset includes 10,824 species and 9 indicators. The second source is "Flore et végétation de la France et du Monde: CATMINAT (baseflor)" (http://philippe.julve.pagesperso-orange.fr/catminat.htm), a continuously updated site on the french flora and vegetation with about 6000 records. Both Fanelli et al. (2006) and CATMINAT include the original EIVs (Light (L), Temperature (T), Continentality (K), Soil Moisture (F), Soil pH (R), Nitrogen (N); Salinity was excluded). In CATMINAT two further IVs, namely Air Humidity (Hu), Soil Organic Matter (SOM) were added. In Fanelli et al. (2006) the index of Hemeroby (H) (Kowarik, 1990; Fanelli and Testi 2008) is also included. We considered the reduced dataset obtained by considering only tree species ('indices reduced'), with a size of 206 by 9 (species by indicators). Despite the value for Continentality index can be obtained from distribution maps (Berg et al. 2017), we still included it in our analysis to test the effectiveness of employed imputation methods. For the second ('vegetation') dataset, we considered a vegetation relevés database covering Southeast Europe (Italy, Slovenia, Croatia, Bosnia, Serbia, Macedonia, Montenegro, Greece) and derived from the European Vegetation Archive (EVA; Chytrý et al., 2016). Only the relevés with at least one tree species were included with a final size 206 × 30,797 (species × relevés). The imputation techniques were evaluated considering only those tree species with all the EIVs known, so that a displacement between estimated and observed values can be calculated. Therefore, for the purpose of comparison, we considered a speciesby index matrix of size 97x9, while the number of relevés is reduced to 29,935 after removing relevés where only the excluded species were present.

2.2. Imputation methods

The multiple imputation procedure was based on the techniques reported in Table 1, which also refers to the libraries of the R software (R Core Team, 2018) used for each method. For data manipulation and preprocessing we used the dplyr library (Wickham et al., 2023). To test the null hypothesis that the mechanism generating the missing data is MCAR, the LittleMCAR function from the BaylorEdPsych package (Little, 1988) and the testMCARNormality function from the MissMech package (Jamshidian and Jalal, 2010) were employed.

For further details on the imputation methods see below.

Table 1
Synthesis of adopted imputation techniques.

Method	Acronym	Summary and R packages
K-nearest neighbour	KNNplot, kNNreg	Nonparametric supervised learning method which uses proximity to classification or predictions around an individual data point (Zhang, 2016), from the caret library (Kuhn, 2008)
Multiple linear regression	Reg.lin	Represents an extension of simple linear regression (Van Buuren, 2018), from thecaret library (Kuhn, 2008)
Ridge regression	Reg.ridge	Often referred to as shrinkage method, by adding a regularisation term it solves collinearity problems (Van Buuren, 2018), from the glmnet library (Friedman et al., (2010)
Reprediction algorithm	Hillt.o, Hills.o	Developed by Hill et al. (2000), it combines van der Maarel (1993) and Ter Braak and Gremmen (1987), programmed in R by one Author
Fanelli et al. (2006)	Fanelli et al. ta, Fanelli et al. ho	Reprediction algorithm (Hill et al, 2000) with a different rescaling method, programmed in R by one Author
Weighted averaging and Weighted averaging partial least square regression	WA, WAPLS	Uses regression and weighted mean calibration (Ter Braak et al., 1993), from the rioja library (Juggins, 2023)

2.2.1. The k-Nearest Neighbour method

This technique involves imputing missing values using an average of the values associated to the nearest k units (Zhang, 2016), once a distance function has been chosen. Given the presence/absence of tree species in the different relevés, it was decided to use two different approaches based on k-NN. In the first case (k-NNreg), the Euclidean distance was used without explicitly considering the co-presence of species in the relevés. Formally, the Euclidean distance for a species pair (i,h) with respect to the j-th Ellenberg indicator is defined by:

$$d(i,h) = \sqrt{\left(x_{ij} - x_{hj}\right)^2}$$

where x_{ij} and x_{jh} represent the coordinates of the species in the j-th dimension. Having only one dimension in which the species is measured we have $i = 1$.

In the second case (k-NNplot), we considered the distance function proposed by Jaccard (1901), which measures co-presence (absence) only rather than abundance values. The (Euclidean or Jaccard) distance was considered during model training to select the k nearest neighbours for a species. The k-NN technique was indeed applied using both distance functions.

For the choice of the number of neighbours (k) to be considered, we selected the number that produces the smallest estimate of the average prediction error; for this purpose, we chose to proceed, as the number k varies (from 1 to 20), with a technique known as M-Fold Cross Validation (M-FCV; Monsteller and Tukey, 1968) to evaluate the performance of the technique on the "indices reduced" dataset. This technique involves choosing the number M of folds and the number k of neighbours and it is based on repeating the following procedure for B = 1000 samples: a) dividing the dataset in M equally sized folds; b) training the model on each fold and using the others to validate it by estimating the corresponding Mean Squared Error (MSE); c) compute the average of estimated MSE for the folds. This returns an estimate for the mean squared error for each of the B samples; thus, an approximation is obtained from the distribution of this quantity across the samples. At the end of the procedure, the (CV) optimal k-value was chosen based on the prediction error estimate. This technique was applied to all Ellenberg indices.

2.2.2. Multiple linear regression and ridge estimator

Imputation by multiple linear regression involves estimating a regression equation for each dependent variable (a specific EIV), using the other EIVs and characteristic features as predictors (independent variables). The missing value is imputed using the value predicted by the equation. If we denote by Y_L the Ellenberg index to be predicted and by (X_L, X_T, X_K, X_F, X_N, X_R, X_H, X_{humi}, X_{org}) the remaining indices (predictor variables), with the caution that whatever is response, it should be eliminated from the predictors, the regression equation is defined by:

$$E(Y_L|X) = b_0 + b_1X_L + \cdots + b_9X_{org}$$

In the case of Multiple Linear Regression, the use of several, potentially interrelated, independent variables (regressors) may pose the problem of collinearity, which could lead to unreliable estimates. A solution is provided by ridge regression, which 'constrains' the coefficient estimates, pushing them towards zero. This leads to a reduced set of coefficient estimates. This procedure usually leads to an increase in bias.

For both multiple linear regression and ridge regression, the M-FCV technique (Monsteller and Tukey, 1968) was applied for model validation, repeating the procedure we have described before. As with the two k-NN techniques, the techniques are evaluated by Cross Validation, subdividing the initial dataset in M = 4 folds and replicating the

procedure B = 1000 times. Ridge regression is based on the penalised loss function

$$Q(\beta) = \|y - x\beta\|^2 + \lambda||\beta|\,|^2, \lambda > 0$$

The first term represents the sum of squared residuals (SSR) which is the objective function for the standard linear regression while the λ term denotes the penalty. The second term reduces the variability of the coefficient estimates, shrinking them to zero. When the λ term is zero, the ridge regression reduces to classical linear Regression. The more λ increases, the greater is the weight of the penalty associated to the norm of the regression coefficients, which will tend to zero (even if no coefficient is set precisely equal to zero). This reduces the complexity of the model without having to delete any variable. In terms of Mean Squared Error (MSE), for the two models, we have the following:

- OLS estimator:

$$MSE(\widehat{\beta}|X) = E\left[\|\widehat{\beta}-\beta\|^2 \middle| X\right] = \tau r(Cov[\widehat{\beta}|X])$$

- ridge estimator

$$MSE(\widehat{\beta}_\lambda|X) = E\left[\|\widehat{\beta}_\lambda-\beta\|^2 \middle| X\right] = tr(Cov[\widehat{\beta}_\lambda|X]) + \|bias(\widehat{\beta}_\lambda|X)\|^2$$

The difference between the two is:

$$tr(Cov[\widehat{\beta}|X] - Cov[\widehat{\beta}_\lambda|X]) - \|bias(\widehat{\beta}_\lambda|X)\|^2$$

The ridge estimator usually has a lower variance than the OLS estimator, and generally, the MSE estimate obtained for the ridge estimator is lower. Thus, although the Gauss–Markov theorem states that the OLS estimator has the lowest variance in the class of unbiased estimators, there exists a biased estimator (ridge) whose MSE may be lower than that of the corresponding OLS estimator. The search for λ is achieved using a Leave- One-Out technique (Vehtari et al., 2017) aimed at finding the optimal λ-value (in the set of possible lambda values, λ_p) in the sense that it minimises the corresponding MSE (Mean Squared Error). Even we used two techniques, M-fold cross validation and leave one out, the interpretation of the following results in terms of MSE does not change. It is in fact known (see Efron and Tibshirani, 1997; Berrar, 2019; Hastie et al., 2009), that the Leave-One-Out technique leads to a lower bias and higher variability in the MSE estimates compared to the M-fold CV technique. Therefore it is reasonable to think that if the M-fold technique were applied to Ridge Regression less variability would be obtained (more flattened boxplots) but still high mean and median values.

2.2.3. The Reprediction algorithm

This algorithm (described by Hill et al. (2000) has been implemented in four different versions, depending on certain conditions. The first two, $Hill_{t.o}$ (Hill total offset) and $Hill_{s.o}$ (Hill semi offset), are defined by two steps.

Let us denote by $\mathbf{A} = [a_{ij}]$ the data matrix of the presence (1) / absence (0) of species ($j = 1,\ldots,p$) in the relevés ($i = 1,\ldots,n$). The row totals r_j represent the number of relevés containing a given species, while the column totals correspond to the number of species present in the i-th relevé: $r_j = \sum_i a_{ij} \quad c_i = \sum_j a_{ij}$

We denote by X_j the value of the Ellenberg index considered for species j, $j = 1,\ldots,p$, the following average values can be calculated:

$$mX_i = \frac{\sum_j a_{ij}X_j}{\sum_j a_{ij}} \qquad \left(\text{mean value of X in relevé } i\right)$$

$$mmX_j = \frac{\sum_i a_{ij}mX_i}{\sum_i a_{ij}} \qquad \left(\text{mean value of } mX_i \text{ for species j}\right)$$

In the first step, the average (across species present in a relevé) value is calculated in each relevé mX_i.

Then, the average value of such relevé -specific values is attributed to the species with missing data considering only the relevés where the species is present mmX_j.

The predicted values are then modified by adding an additional offset. This makes it possible to adjust the obtained values to avoid narrowing the distribution towards mean or extreme values.

The $Hill_{t.o}$ and $Hill_{s.o}$ methods differ in the offset calculation. For the $Hill_{s.o}$ method we have:

$$Offset_j = mean_{(S)}\left(X_j - mmX_j\right)$$

where, X_j is the observed Ellenberg index for species j, while mmX_j indicates the value predicted by the model. In this case, not all species are considered for the offset calculation, but only the set S including the 20 species with the smallest distance between the observed value and the value predicted by the model. The final predicted value will be:

$$\widehat{mmX}_j = mmX_j + Offset_j$$

For the $Hill_{t.o}$ method, all species are considered.

The third (*Fanelli et al.*$_{t.m}$) and fourth (*Fanelli et al.*$_{h.o}$) methods are based on Fanelli et al. (2006), a technique based on the guidelines by Hill et al. (2000), with steps 1 and 2 described above. *Fanelli et al.*$_{t.m}$ differs from Hill et al. (2000) in the calculation of rescaling.

$$mmX_j - \left(\overline{A} - mmX_j\right)^{1/3}$$

where$\overline{A}$is the mean value of the dataset.

In *Fanelli et al.*$_{h.o,}$ the offset of Hill et al. (2000) was reintroduced in the rescaling of Fanelli et al. (2006).

2.2.4. Weighted Averaging (WA) and Weighted Averaging Partial least Squares (WAPLS) regression

The Weighted Averaging (WA) and Weighted Averaging Partial Least Squares (WAPLS) techniques, proposed by Ter Braak and Juggins (1993) and Ter Braak et al. (1993), are based on weighted averages to estimate missing indicators' values. The WAPLS technique (Ter Braak and Juggins, 1993), represents an improved version of WA, as it is based on the use of several principal components, as many as necessary to have a good prediction. Thus, the number of components that returns the lowest prediction error is identified through a Leave-One-Out Cross-Validation (LOOCV) (Vehtari et al., 2017) technique applied for each sample (B = 1000) and each variable. For each variable, the most frequent number of components was selected, and the estimated value of Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) were finally calculated. We used the LOO CV approach, since it is the default implemented in the rioja R library (Juggins, 2023).

As for the ridge regression, the LOO technique was also used for WA and WAPLS and therefore the same considerations made on the comparison with the K-fold CV apply.

2.3. Comparison of imputation techniques

For each imputation technique and for each Ellenberg indicator b = 1,..,B (B = 1000) repeated samples have been drawn from the tree dataset with known EIVs. For each sample, except for Ridge Regression, WA and WA-PLS, the M-Fold cross Validation technique Monsteller and Tukey, 1968) has been applied:

1. the sample is subdivided into m = 1,..,M sub-samples
2. the technique has been applied to each of these samples (test) and used to make predictions for the remaining (M−1) folds, considered as test. The MSE associated with the sample is estimated by the sample mean $MSEm = \frac{1}{n_{test}}\sum_{i\in test}(Y_i - \widehat{Y}_i)^2$ where Y_i and $\widehat{\gamma}_i$ denote the observed and the imputed values. This term estimates the mean squared discrepancy between observed and estimated data values; therefore, a low value of this term is to be considered good.
3. repeat 1–2 for each fold. At the end, the estimated MSE for each sample is obtained by

$$MSEb = \frac{1}{M}\sum_{m\in M} MSEm$$

The estimates for MSE in each sample are combined to provide a synthetic measure of the stability of the model. In the context of multiple imputation [2–7;13] 5 values were imputed for each missing value, each obtained by adding a component of a random nature (normal variate with zero mean and very small variance (0.5)).

For each method the following parameters were calculated: a) mean position; b) lower estimated values of the Mean Squared Error (MSE); c) dispersion/variability of the MSE estimates via the interquartile range (IQR); d) position and variability preferring method with the lowest median value when IQr values are similar. Each Ellenberg Indicator is reported in a figure to identify the best method among those considered in the Cross Validation experiment: a boxplot of the distribution of the estimated MSE for each imputation method is reported for each index (see "Results" section).

An expert judgement of estimated EIVs was carried out for each species to double-check the consistency of the results looking at their altitudinal, geographical distribution and ecological behaviour.

2.4. Comparison of imputed values with values from the recent literature

To check our results against the state-of-the-art knowledge about EIVs, we ran correlation tests and computed linear regressions to relate EIVs from Dengler et al. (2023) and imputed values obtained via the most effective method for each of the imputed indicators. Most of the species considered in this research are also present in the dataset considered by Dengler et al. (2023), but the latter lacks information on 26 species in our dataset. These species have therefore been excluded from this comparison.

3. Results

3.1. The missing data generating mechanism

The missing data mechanism can be MCAR, MAR or MNAR (see Little and Rubin, 2022). In the present context, the missing data concern species not included in the original list by Ellenberg (1979) and the values for indicators H, humidity and org_matt. Leaving apart the MCAR meachanism, which is quite simple, we may turn to consider the MAR one; if such a mechanism holds, multiple imputation may provide consistent estimates of the quantities of interest. Considering that the MCAR mechanism is the only one that can be verified by statistical tests, Table 2 shows the results obtained by applying two tests to verify the null hypothesis that the mechanism generating the missing data is MCAR.

While the MAR mechanism could be plausible, in the sense that missingness may depend on unknown quantities (EIVs) only through the selection of a specific geographical area, that is via its specific characteristics, we may not rule out a potential MNAR mechanism. However, the use of a well-defined statistical model to estimate missing EIVs may be a good starting point to build, post-estimation, a sensitivity analysis by varying the weight of each relevé, each index, as in Jansen et al. (2003), and observing how the predicted response changes according to these perturbations. While this is beyond the scope of this paper, it may be simply replicated by those interested in analyzing stability of the results. Below, we report the boxplotsdescribing, for each EIV, the distribution across the samples of the MSE estimate.

Table 2
Tests to verify the MCAR hypothesis for the missing data-generating mechanism.

	p-normality (Howkins)	p-valcomb (Nonparametric)
TestMCARNormality	1.57E-25	<0.001
	p-value	Chi-squared
LittleMCAR	<0.001	11082.67

3.2. Comparing the imputation techniques

Mean Square Error (MSE) represents the expected value of the square difference between the observed and the predicted values obtained by a specific method, while the RMSE represents its square root. If the imputed values are close to the observed values, a small MSE is obtained; the greater the difference between the observed and the imputed values, the more the error will tend to grow. Thus, whether estimates of the MSE or RMSE are considered, the technique with the smaller value tends to better predict the observed value. The graphs in Fig. 1 for each index are based on a summary of the values estimated in the Cross Validation exercise considering B = 1000 samples. For further details see SM4. The imputation technique which results to be the best for most indices is applied to the overall dataset of tree species of size 206 × 9 (see SM5).

3.2.1. Light (L)

On average, the predicted values for light (L) (Fig. 1) obtained by the methods based on weighted averages (WA), i.e. excluding the two regressions (Reg.lin and Reg.ridge) and the two kNN (kNNreg and kNNplot), are higher than those by the other methods. The two regressions have a much lower MSE- *Reg.lin* and *Reg.ridge* have a lower box height (IQR) than the others, so a lower dispersion in the 50 % central part of the estimated MSE distribution. For the other methods, on the other hand, we have that the estimates are further away from the observed values, in particular for $Hill_{t.o}$,$Hill_{s.o}$ and $Fanelli_{t.a}$. In addition, the two regressions report an extremely low median estimated MSE value. The Ridge Regression (Reg.ridge) and Linear Regression (Reg.lin) methods have less interquartile variability of the estimates and tend to focus on small error values. The Ridge Regression (Reg.ridge) method is slightly better than the Reg.lin, under this point of view. Furthermore, the Ridge Regression (Reg.ridge) method has no outliers.

3.2.2. Temperature (T)

The techniques with lower median of estimated MSE for the temperature index (T) are *Reg.lin*, *Reg.ridge* and *kNNreg*. The methods with lower variability of the estimates are *Reg.lin*, *Reg.ridge* and WA, with a lower IQR with respect to the others. So, for 50 % of the samples, there is lower dispersion in the MSE estimates. For the other methods, however, the MSE estimates are greater and, therefore, the predicted index values are more further away from the observed, particularly for the two Hill methods. The Reg.ridge and Reg.lin methods perform better, and the latter should be considered optimal if the species has more missing values.

3.2.3. Continentality (K)

Again, the models with the lowest median of estimated MSE are the *Reg.lin*, *Reg.ridge* and *kNNreg* techniques. The methods with the lowest dispersion of the MSE estimates are *Reg.lin*, *Reg.ridge*,* WAPLS and *kNNreg*, with a lower IQR when compared to the others. For the other methods, in fact, the estimates are far from the observed value, in particular for the two methods of *Hill* and *Fanelli et al*. The WAPLS and kNNreg methods show a lower variability in the MSE estimates, but they have a higher median value when compared to *Reg.lin* and *Reg.ridge*. Again, the 'best' method is Reg.lin, for which no outliers are observed.

3.2.4. Soil moisture (F)

On average, the estimated values of MSE for the soil moisture index (F) for *Reg.lin, Reg.ridge* and *kNNreg* are the lowest. The values for WA, WAPLS and *Fanelli et al.*$_{ta}$ are particularly high. The *Reg.lin, Reg.ridge, kNNreg* and WA methods have a lower box height (IQR) than others, so for 50 % of the samples there is a lower dispersion in the MSE estimates. The other methods, on the other hand, show greater variability. The *Reg.ridge* and *Reg.lin* methods show less interquantile variability of estimates and tend to produce small MSE estimates. The *Reg.ridge* method is preferred, as it has slightly less IQR variability in the estimated MSE than *Reg.lin* with more frequent low values.

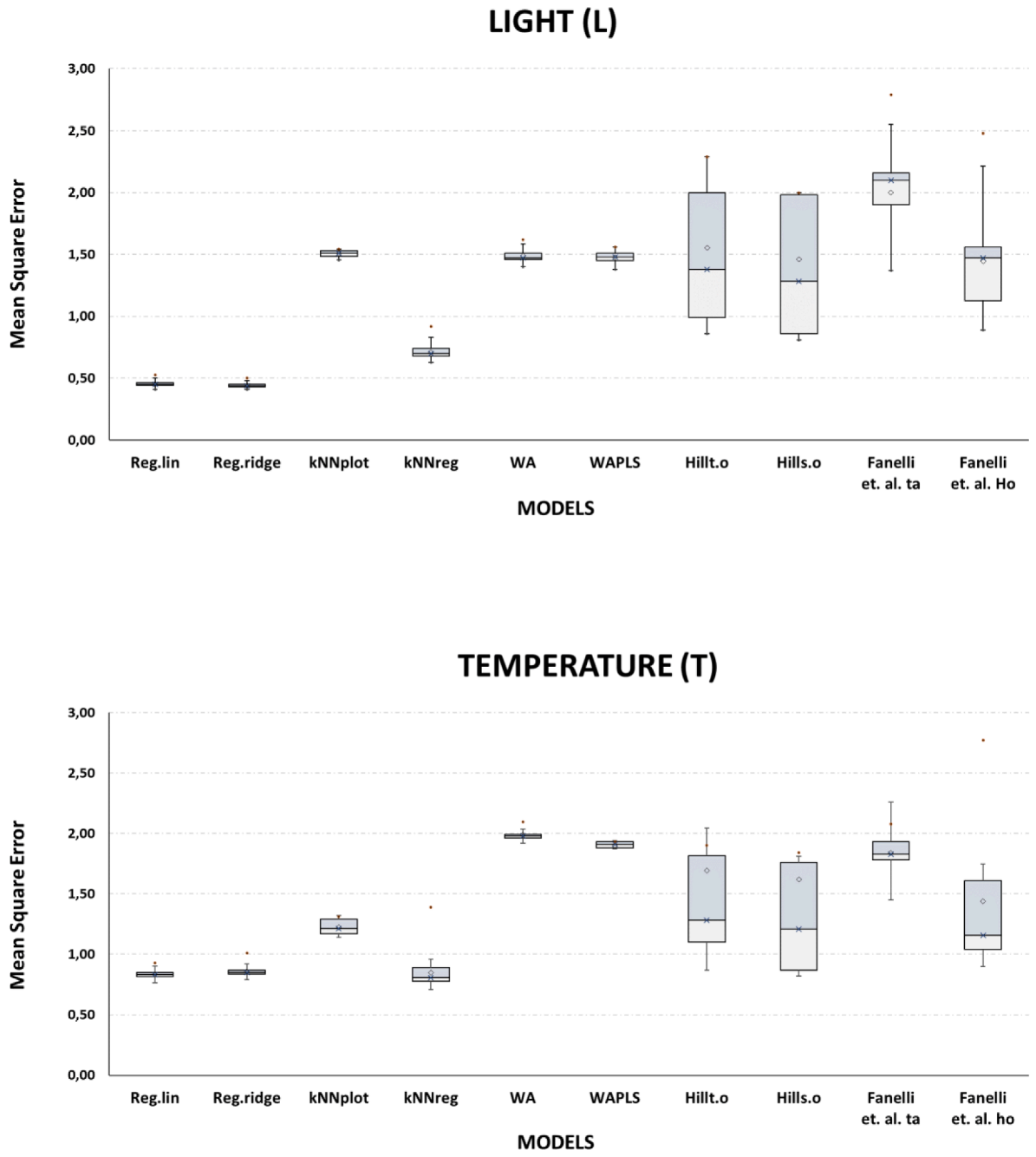


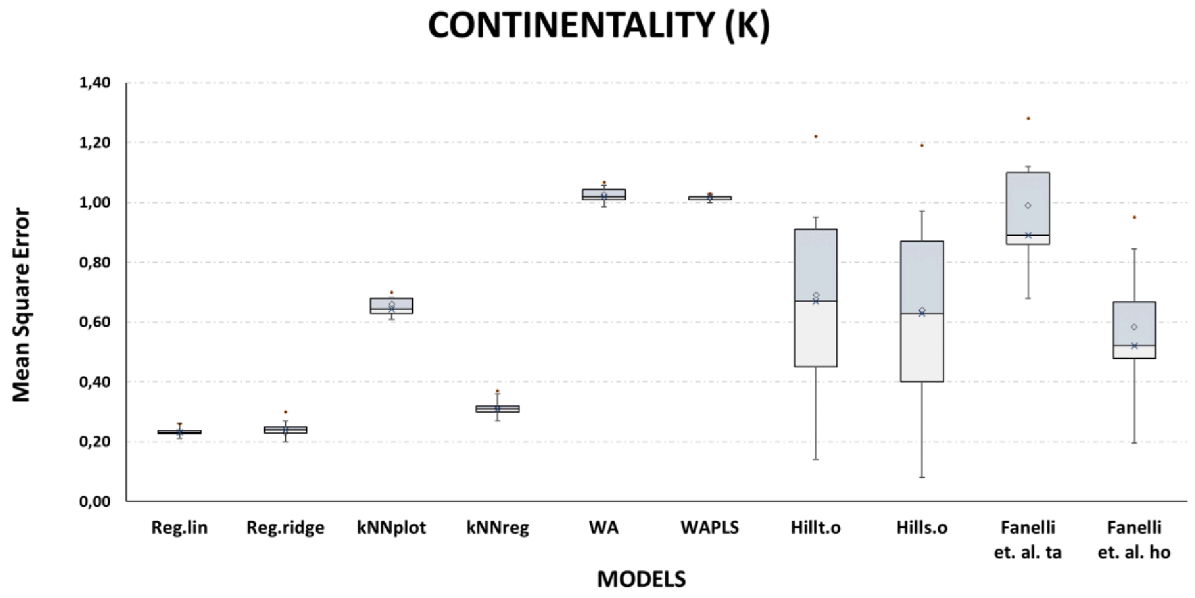
Fig. 1. Boxplots reporting the distributions of the estimated MSE values for the Ellenberg indicators and the analysed methods. For each Ellenberg Indicator we have the distribution of the MSE (y-axis) for each applied imputation technique (x-axis). The indicators of Ellenberg are: L: (Light) T (temperature) K (continentality); F (soil moisture); N (nitrogen); R (soil pH); H (hemeroby index); humidity (air humidity); org_matt (organic matter). Imputation techniques are: Reg.lin: Linear Regression; Reg.ridge: Ridge Regression; kNNplot: k-Nearest Neighbour plot; kNNreg: k-Nearest Neighbour regression; WA: weighted average; WAPLS: Weighted Averaging Partial Least Squares; Hillt.o: Reprediction Algorithm of Hill total offset; Hills.o: Reprediction Algorithm di Hill semi-offset; Fanelli et al ta: revision of prediction algorithm of Fanelli et al; Fanelli et al. ho: revision of prediction algorithm of Fanelli et al. with Hill Offset.

3.2.5. Nitrogen (N)

The estimated MSE values for Nitrogen (N) by regression and *kNN* are on average lower when compared to the other methods. On the contrary, the two methods WA and WAPLS have a higher mean value. Except for the two methods of *Hill* and Fanelli et al., all methods have reduced box heights (IQR), so there is less dispersion in the central distribution of the estimated MSE. When comparing the medians, the *Reg.ridge, Reg.lin* and *kNNreg* methods are all suitable methods, with reduced squared error values. The *Reg.ridge* has a slightly reduced IQR and a slightly lower median value.

3.2.6. Soil pH (R)

On average, the estimated MSE values for soil pH (R) using the regression and *kNN* methods are lower when compared to the other methods. The method by *Fanelli et al.*$_{ta}$ has the highest average MSE estimate. Except for the *Hill*$_{t.o}$ and *Hill*$_{h.o}$ methods, all methods have reduced box heights (IQR), especially *kNNplot* and the two regressions. R*eg.ridge, Reg.lin* and *kNNreg* methods are good due to the low variability in MSE estimates. *kNNreg* method compares favorably with the others with a narrower IQR and a lower median value for all empirical situations where several indicators are missing.



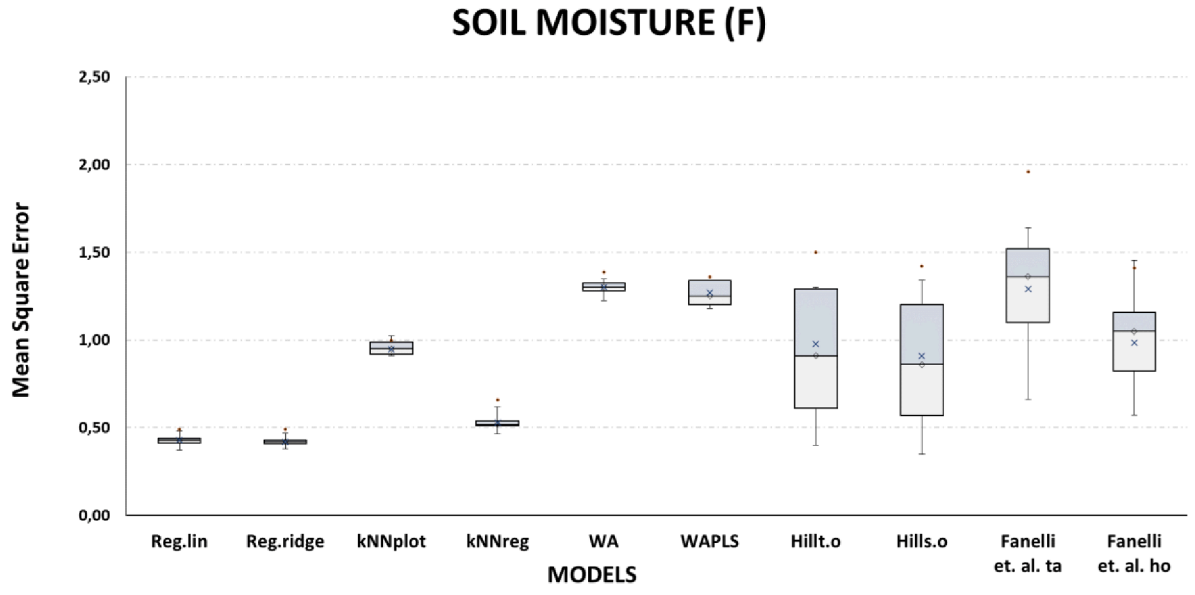


Fig. 1. (*continued*).

3.2.7. Hemeroby (H)

The regression techniques and *kNNreg* have mean MSE estimates, while for the other methods the value is similar and is on the high side. Apart from $Hill_{t.o}$ and $Hill_{h.o}$, all methods have low box heights (IQR), especially *kNNreg* and ridge regression. The *Reg.ridge* and *kNNreg* methods are similar in IQR, but *kNNreg* method has a lower median value and is recommended, particularly for all situations where several indices are missing.

3.2.8. Air humidity (humidity)

All techniques have a higher mean MSE estimate when compared to the other indices. The one with a slightly lower mean value is *kNNreg*. Except for the Hill and Fanelli et al. methods, all methods have reduced box heights (IQR), especially WA and the ridge regression. The two techniques, Regression and *kNN*, are very similar to each other in terms of the variability of MSE estimates. The *kNNreg* and *Reg.ridge* methods are similar in terms of variability and median, so they are both valid for estimating the humidity index. When looking at variability and thus to the smaller box size, the *Reg.ridge* method is preferred, while the *kNNreg* technique is preferred for those empirical situations where several indices are missing.

3.2.9. Organic matter (org_matt)

Similar to the *humidity* index, all models have a high mean value of MSE estimates. The one with a slightly lower mean value is still *kNNreg*. Leaving aside the methods by *Hill* and *Fanelli et al.*, the others have higher IQRs than those seen so far. The kNNreg method seems to be the best in terms of variability, mean value and in all those empirical situations where several indices are missing.

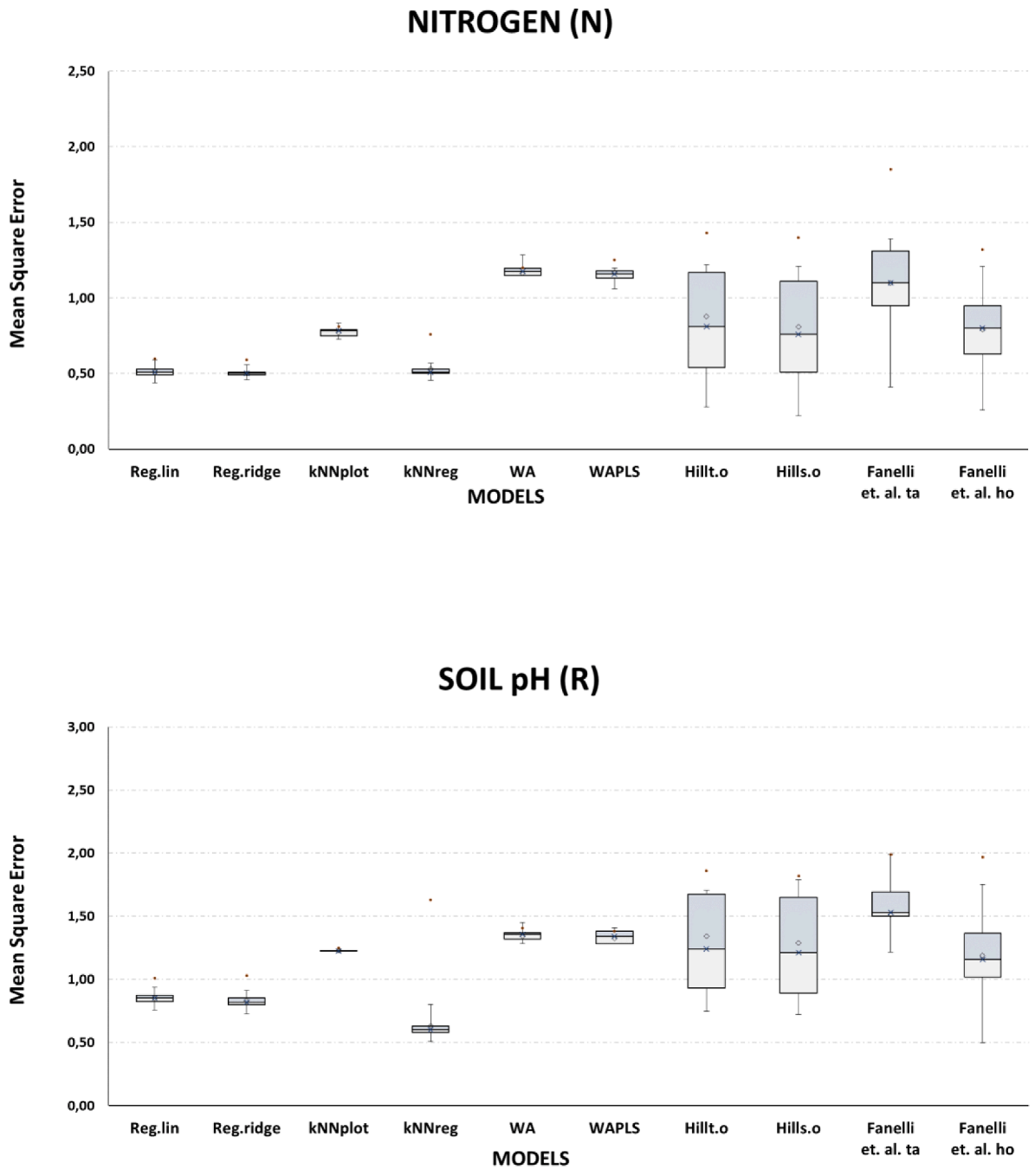


Fig. 1. (*continued*).

3.3. Expert evaluation of the Ellenberg's indicators

3.3.1. Light (L)

With the *Fanelli et al.* methods, light values are low for most species; *kNNreg* gives the most plausible values, while WA and WAPLS produce some outliers in the MSE distribution.

3.3.2. Temperature (T)

Fanelli et al.$_{ta}$ and *Hill* give very high and unreliable values for some temperatures, e.g. *Quercus frainetto* and *Quercus pubescens*. WA and WAPLS give anomalous values for *Picea abies* and other species.

3.3.3. Continentality (K)

As in other cases, the kNN regression method produces the most reliable results. Nevertheless, the results produced by the other techniques are not very different.

3.3.4. Soil humidity (F)

Fanelli et al. and *Hill* give high values for distinctly xerophilous species such as *Quercus suber* and *Quercus ilex, which is unsatisfactory*. *kNNreg* gives the most plausible values for this indicator.

3.3.5. Nitrogen (N)

kNN gives high, unplausible nutrient values for many species. *Fanelli et al.* seem to give more plausible values and *Hill* seems the best.

3.3.6. pH (R)

Hill$_{t.o}$ gives an anomalous value for *Picea abies* and other species regarding R. *Reg* and *kNN* give good results.

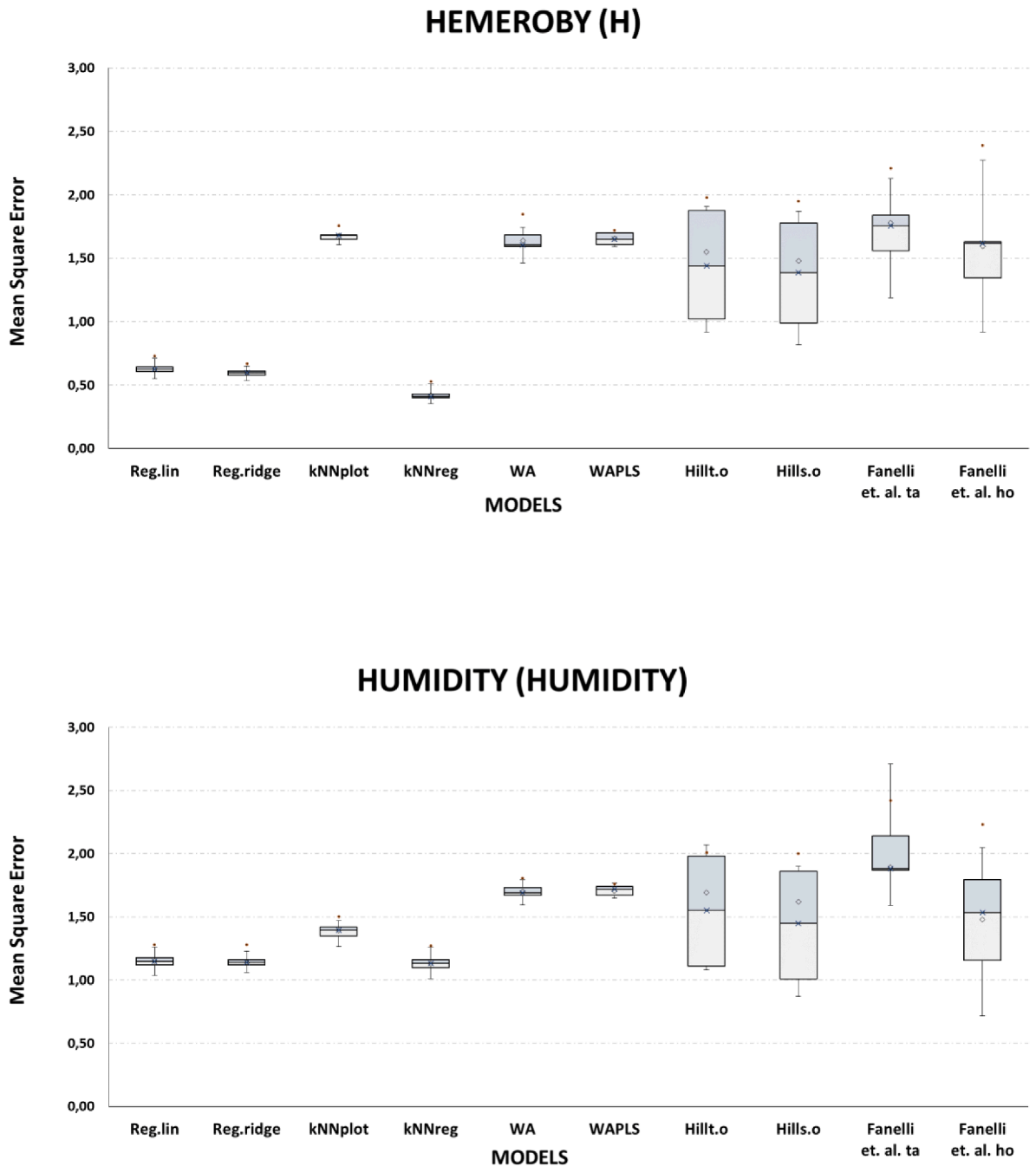


Fig. 1. (*continued*).

3.3.7. Hemeroby index (H)

Fanelli et al. give very squashed values. $HIll_{t.o}$,$Hill_{s.o}$ also present values squashed on the first three levels. kNN gives relatively wide values for hemeroby that seem implausible for these species. *Regr* gives too high a value for *Pinus cembra*, but otherwise, the values seem pretty reliable.

3.3.8. Air humidity (humidity)

All imputation methods give similar, plausible values for air humidity.

3.3.9. Organic matter (org_matter)

Hill and *Fanelli et al.* give plausible and evenly distributed values of humus. Reg squeezes the values to a few levels. *kNN* gives values similar to *Fanelli* et al. and *Hill*.

3.4. Comparison of imputed values with values from recent literature

All correlation tests ran between EIVs obtained by Dengler et al. (2023) and those imputed by following the proposed procedures research resulted highly significant (Table 3). Correlation coefficients obtained by Spearman's method were not very high in some cases. In fact, the scatterplots in Fig. 2 highlight a high dispersion in the data cloud for some EIVs, and the adjusted R-squared values obtained for the regression models are never very high. Nevertheless, all regressions are highly significant as highlighted by the p-value obtained from F-tests.

4. Discussion and conclusion

In plant ecology, EIV are widely used for monitoring the state and evolution of ecological communities (Hedwall et al., 2019; Jonsson

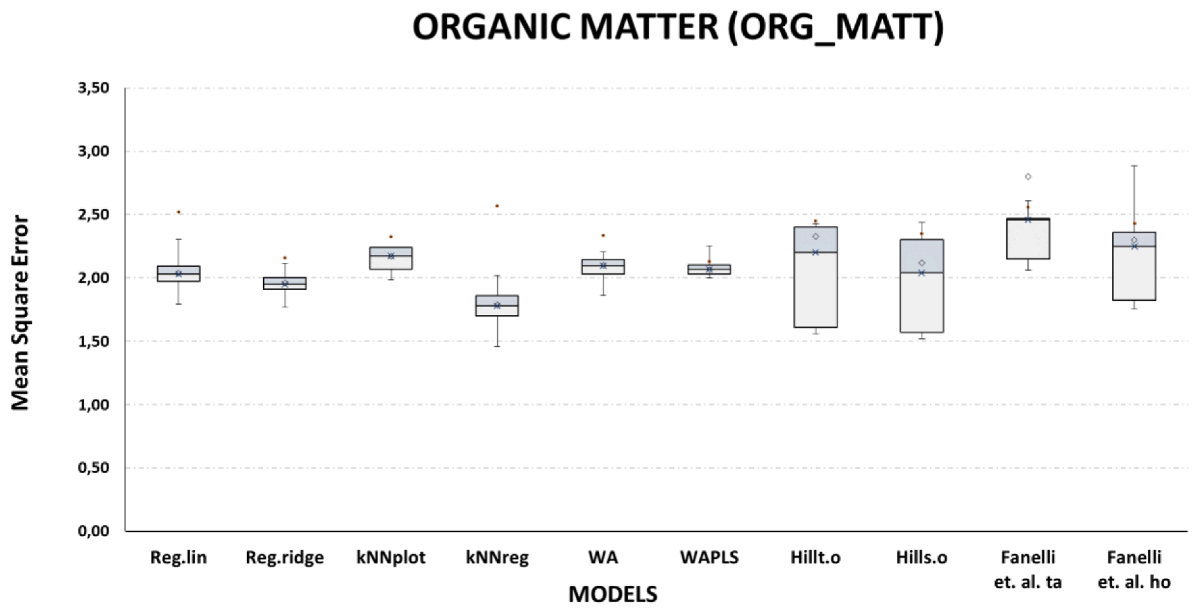


Fig. 1. (*continued*).

Table 3
Spearman's rho (ρ) values obtained from correlation tests between EIVs from Dengler et al. (2023) and imputed values obtained in this study (available in SM).

Ellenberg Indicators					
	M	**N**	**R**	**L**	**T**
Spearman's ρ	0.420	0.437	0.379	0.370	0.525
Significance level	***	***	***	***	***

et al., 2021). There has been recently a resurgence of interest about such indicators, with the publication of lists of EIV for multiple species based on a massive amount of data, in particular the huge number of relevés stored in the EVA database (Dengler et al., 2023; Tichý et al., 2023).

Under ideal conditions, EIV are derived from a wide range of experimental measurements of the environmental factors that determine the realized niche of a species (Fanelli et al., 2007). However, this ideal situation is rarely realized. In particular, detailed information may be available for some species while lacking for many others. In this more frequent situation EIV are derived from expert judgment on the basis of profound experience of the distribution along different gradients. This often leads to contradictions, due in part to the fact that different geographical areas can present a different ecology of the same species and thus different EIV for that species, and more importantly by disagreement among experts (Dengler et al., 2023). In such cases, imputation methods can represent a valid starting point to reconstruct the missing data. In vegetation ecology, the most common methods for handling missing data include 'Ter Braak' and 'Hill' (e.g., Tichý et al., 2023). These are good estimators of indicators for missing species, but further imputation methods frequently used in statistical disciplines (e.g., Onkelinx et al., 2017) have not been adequately considered in the vegetation ecology.

This paper tested some imputation techniques to estimate EIV for South East European tree species. It aimed to assess whether these methods could represent an effective alternative to current methods. The results are encouraging: it appears that such methods, that are rarely used in ecology, appear more appropriate for the estimation of missing data than other more widely used by ecologists.

For the comparison of techniques we focused on MSE estimates. Therefore, for each EIV, a boxplot (Fig. 1) has been obtained in a CV exercise for each method, representing the distribution of the MSE obtained by that method. To determine the best method for each index, the position (distribution of estimated MSE on high or low values) and the variability (interquantile range) have been considered. The method which is distributed towards low error values and which has less variability is to be preferred.

According to this cross-validation exercise, multiple linear regression and k-Nearest Neighbor are the most promising methods. The boxplots in Fig. 1 show, for such methods, less variability in the corresponding MSE estimates ("stronger compression"), which tend to spread over small values and produce a low median (and average) value. Conversely, *Fanelli et al.*$_{t.m}$, *Fanelli et al.*$_{h.o}$, *Hill*$_{s.o}$ and *Hill*$_{t.o}$ are the techniques with the least satisfactory results. These methods show a significant variability in MSE estimates (resulting in an elongated boxplot) with high mean and median values, indicating a poor predictive ability under cross validation. Ridge Regression showed good performance compared to the other methods, especially when data were missing only for one EIV. Ridge regression and linear regression performed well compared to other methods, especially when data was missing for only one EIV, but needed ancillary information (values from other indicators), which sometimes could be difficult to retrieve. Each imputation technique has its merits and flaws and a different way to be approached. For example, some methods, such as *Fanelli et al.*$_{h.o}$ and kNNplot, need to be modified in advance by the operator. Thus, stating that one method is better than the other may lead to misunderstandings or even to errors. It can be said that the ridge regression method performs well in terms of prediction, compared to the other methods applied, but mainly when only one indicator is missing. In the presence of several missing indices, the kNNreg method is to be preferred.

We did not discuss some of the several methods presented in literature, for instance those utilizing the class center-based data imputation (Nugroho et al., 2018, 2021; Tsai et al., 2018). Testing the effectiveness of these additional techniques might represent an interesting opportunity for further research.

The results of the statistical comparison are also consistent from the point of view of the ecology. In general, *kNNreg* gave the result closest to what an expert botanist would consider the correct value for most EIV and most species, whereas *Fanelli et al*$_a$ and *Hill* give often less reliable values, but with notable exceptions, for instance in the case of N indicator, which is more reliable if imputed via the method by *Fanelli et al.*$_{ta.}$ The other methods give values satisfactory but inferior to kNNreg.

Our data are highly comparable with known data from literature

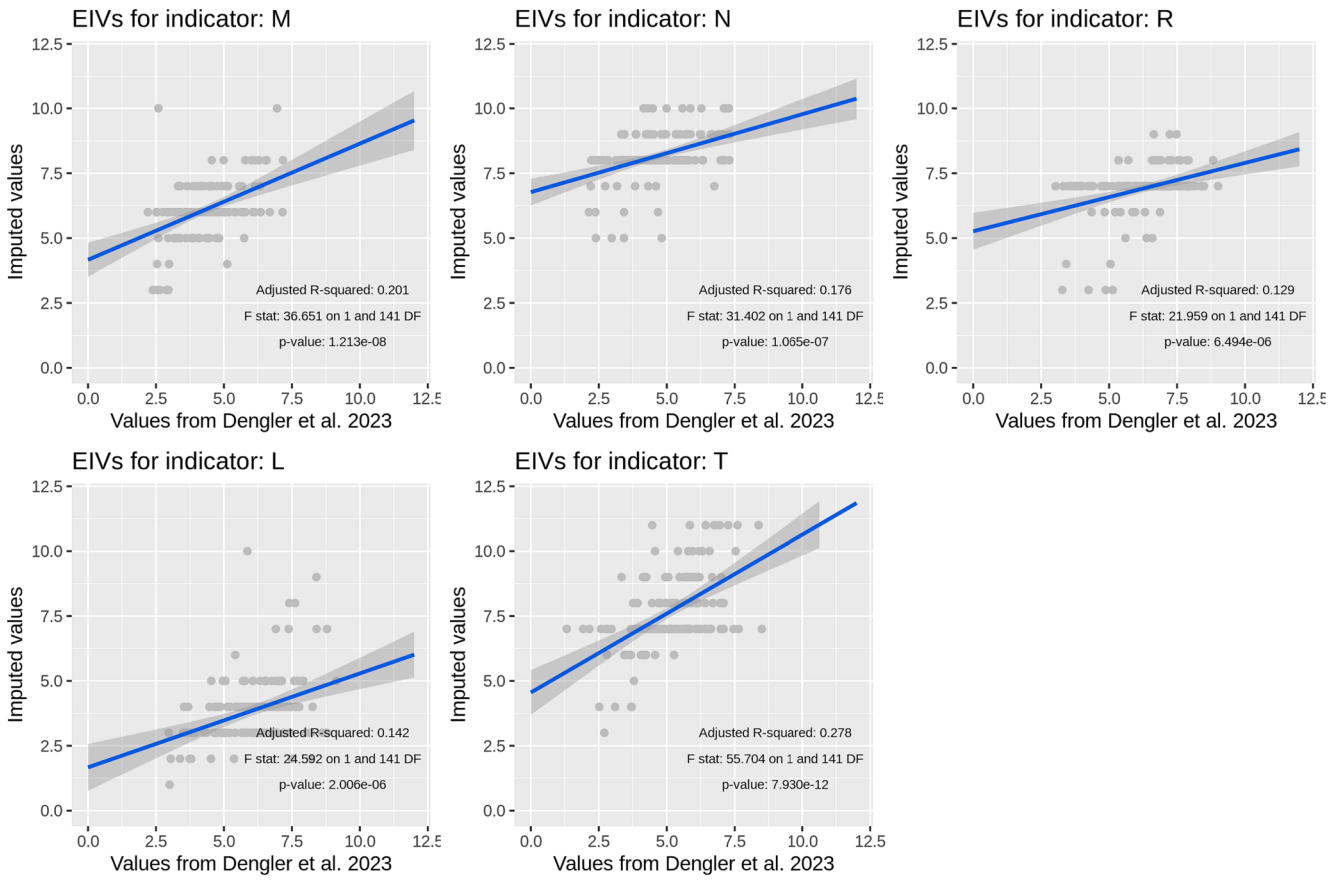


Fig. 2. Scatterplots and results of the linear regressions between EIVs from Dengler et al. (2023) and imputed values obtained in this study (available in SM).

(Dengler et al., 2023), as it can be evinced by looking at the results of the correlation and regression analysis (Fig. 2, Table 3). The correlation coefficients computed between some of the EIVs in our data and data from literature are not very high, but this is not unexpected since the data from Dengler et al. (2023) encompass the entire Europe, whereas our analyses are restricted to a specific geographic area with a distinct climatic and biogeographical context.

The EIVs imputed by KNN for some species (reported in the Tables available as Supplementary Material) may appear unreliable, especially to a scholar with a central European perspective. For instance, *Fagus sylvatica* and *F. orientalis* have quite different EIVs, which may look surprising given their morphological similarities. As a matter of fact, Fagus *sylvatica* and *F. orientalis* have hugely different ecology: *F. sylvatica* is a typical central European species, whereas *F. orientalis* is a Colchic species. Moreover, *F. sylvatica* seems of relatively recent origin, whereas *F. orientalis* is fundamentally a tertiary relic. This is also confirmed by phytosociology, where communities with *F. orientalis* have quite different ecology from the ones with *F. sylvatica* (Willner et al. 2017).

Another unexpected result is related to some *Quercus* species, for instance *Q. ilex*, whose T value is 9 in Ellenberg et al. 1991 and 11 in our results (Table S5), or *Q. robur* (6 in Ellenberg et al. 1991, 8 in our results). These apparent inconsistencies are actually due to the fact that our study area is climatically and biogeographically different from the one considered in the literature. For instance, the T value reported in Ellenberg et al. (1991) for *Q. ilex* reflects the presence of this species in central Europe in extra-zonal areas such as the Insubrian lakes, but the same value is too low for this species in southeastern Europe. It is true that lists for southern Europe have already included data for this species (e.g. Pignatti et al. YYYY), but they did not adjust their results to the Mediterranean contexts.

In a few cases, our results are actually unreliable (e.g. *Phoenix theophrasti*) due to the very restricted distribution of the species that doesn't allow for a correct imputation by statistical methods, but the number of such cases could be considered negligible if compared to the total number of species included in our dataset.

The results we have obtained should be regarded as a starting point for further development. The applied procedure for choosing the imputation method can be applied first to a specific subset of a dataset and, once the best method has been identified, extended to the remaining subsets. Potentially, this approach of imputing missing EIV would tremendously facilitate monitoring species with indicators not included in the original lists.

CRediT authorship contribution statement

Letizia Leccese: Writing – review & editing, Writing – original draft, Formal analysis, Data curation. **Giuliano Fanelli:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization. **Vito Emanuele Cambria:** Writing – review & editing, Supervision, Methodology. **Marco Massimi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Fabio Attorre:** Supervision, Methodology, Conceptualization. **Marco Alfò:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization. **Svetlana Aćić:** Writing – review & editing, Validation. **Erwin Bergmeier:** Writing – review & editing, Validation. **Andraž Čarni:** . **Mirjana Cuk:** Writing – review & editing, Validation. **Renata Custerevska:** Writing – review & editing, Validation. **Panayotis Dimopoulos:** Writing – review & editing, Validation. **Petrit Hoda:** Writing – review & editing, Validation. **Alfred Mullaj:** Writing – review & editing, Validation. **Urban Šilc:** Writing – review & editing, Validation. **Zeljko Skvorc:** Writing – review & editing,

Validation. **Zvjezdana Stancic:** Writing – review & editing, Validation. **Zora Dajic Stevanovic:** Writing – review & editing, Validation. **Rossen Tzonev:** Writing – review & editing, Validation. **Kiril Vassilev:** Writing – review & editing, Validation. **Luca Malatesta:** Writing – review & editing, Methodology, Formal analysis. **Michele De Sanctis:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Data curation, Conceptualization.

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.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ecolind.2024.111851.

References

Austin, P.C., White, I.R., Lee, D.S., van Buuren, S., 2021. Missing data in clinical research: a tutorial on multiple imputation. Can. J. Cardiol. 37 (9), 1322–1331. https://doi.org/10.1016/j.cjca.2020.11.010.

Berg, C., Welk, E., Jäger, E.J., 2017. Revising Ellenberg's indicator values for continentality based on global vascular plant species distribution. Appl. Veg. Sci. 20 (3), 482–493. https://doi.org/10.1111/avsc.12306.

Berrar, D., 2019. Cross-validation. In: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology, Volume I. Elsevier, pp. 542–545. https://doi.org/10.1016/B978-0-12-809633-8.20349-X.

Bertelheimer, M., Poschlod, P., 2015. Functional characterisations of Ellenberg indicator values – a review on ecophysiological determinants. Funct. Ecol. 30 (4), 506–516. https://doi.org/10.1111/1365-2435.12531.

Chemolli, E., Pasini, M., 2008. I dati mancanti. DiPAV-QUADERNI (2007/20).

Chytrý, M., Hennekens, S.M., Borja, J.-A., Knollová, I., Dengler, J., Jansen, F., et al., 2016. European vegetation archive (EVA): an integrated database of European vegetation plots. Appl. Veg. Sci. 19 (1), 173–180. https://doi.org/10.1111/avsc.12191.

Dengler, J., Jansen, F., Chusova, O., Hüllbusch, E., Nobis, M.P., van Meerbeek, K., 2023. Ecological indicator values for Europe (EIVE) 1.0. Veget. Classif. Surv. 4, 7–29. https://doi.org/10.3897/VCS.98324.

Diekmann, M., 2003. Species indicator values as an important tool in applied plant ecology–a review. Basic Appl. Ecol. 4 (6), 493–506. https://doi.org/10.1078/1439-1791-00185.

Domina, G., Galasso, G., Bartolucci, F., Guarino, R., 2018. Ellenberg indicator values for the vascular flora alien to Italy. Fl. Medit 28, 53–61 https://doi.org/0.7320/FlMedit28.053.1.

Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the. 632+ bootstrap method. J. Am. Stat. Assoc. 92 (438), 548–560. https://doi.org/10.2307/2965703.

Ellenberg, H., Weber, H.E., Dull, R., Wirth, V., Werner, W., Paulissen, D., 2001. Zeigerwerte von Pflanzen in Mitteleuropa (indicator values of plants in Central Europe). Scripta Geobotanica 18, 1–262.

Ellenberg, H., 1979. Zeigerwerte der Gefäßpflanzen Mitteleuropas. 2. Aufl. (Indicator values of vascular plants in Central Europe. 2° edition). Scripta Geobotanica 9, 1-122.

Fanelli, G., De Lillis, M., 2004. Relative growth rate and hemerobiotic state in the assessment of disturbance gradients. Appl. Veg. Sci. 7 (1), 133–140. https://doi.org/10.1111/j.1654-109X.2004.tb00603.x.

Fanelli, G., Testi, A., Pignatti, S., 2006. Prototipo di flora ecologica per specie dell'Italia Centro-Meridionale. Il sistema ambientale della Tenuta Presidenziale di Castelporziano. Ricerche sulla complessità di un ecosistema mediterraneo (Prototipo di flora ecologica per specie dell'Italia Centro-Meridionale. Il Sistema Ambientale Della Tenuta Presidenziale Di Castelporziano. Ricerche Sulla Complessità Di Un Ecosistema Mediterraneo). Accademia Nazionale Delle Scienze 505–564.

Fanelli, G., Pignatti, S., Testi, A., 2007. An application case of ecological indicator values (Zeigerwerte) calculated with a simple algorithmic approach. Plant Biosystems 141 (1), 15–21. https://doi.org/10.1080/11263500601153685.

Friedman, J., Tibshirani, R., Hastie, T., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22. https://doi.org/10.18637/jss.v033.i01.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction. Springer. ISBN: 978-0387848570.

Hedwall, P.O., Brunet, J., Diekmann, M., 2019. With Ellenberg indicator values towards the north: does the indicative power decrease with distance from Central Europe? J. Biogeogr. 46 (5), 1041–1053. https://doi.org/10.1111/jbi.13565.

Hill, M.O., Roy, D.B., Mountford, J.O., Bunce, R.G.H., 2000. Extending Ellenberg's indicator values to a new area: an algorithmic approach. J. Appl. Ecol. 37 (1), 3–15. https://doi.org/10.1046/j.1365-2664.2000.00466.x.

Hughes, R.A., Heron, J., Sterne, J.A., Tilling, K., 2019. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. Int. J. Epidemiol. 48 (4), 1294–1304. https://doi.org/10.1093/ije/dyz032.

Jaccard, P., 1901. Etude comparative de la distribution florale dans Une portion des Alpes et du Jura (Comparative study of floral distribution in a portion of the Alps and Jura). Bull Soc. Vaudoise Sc. Nat. 37, 547–579.

Jamshidian, M., Jalal, S., 2010. Tests of homoscedasticity, normality, and missing completely at random for incomplete multivariate data. Psychometrika 75 (4), 649–674. https://doi.org/10.1007/s11336-010-9175-3.

Jansen, I., Molenberghs, G., Aerts, M., Thijs, H., Van Steen, K., 2003. A local influence approach applied to binary data from a psychiatric study. Biometrics 59, 409–418.

Jonsson, B.G., Dahlgren, J., Ekström, M., Esseen, P.A., Grafström, A., Ståhl, G., Westerlund, B., 2021. Rapid changes in ground vegetation of mature boreal forests—An analysis of Swedish National Forest Inventory data. Forests 12 (4), 475. https://doi.org/10.3390/f12040475.

Juggins, S. 2023. rioja: Analysis of Quaternary Science Data. R package version 1.0-6. https://cran.r-project.org/package=rioja.

Kamshidian, M., Jalal, S., Jansen, C., 2014. MissMech: an R package for testing homoscedasticity, multivariate normality, and missing completely at random (MCAR). J. Stat. Softw. 56, 1–31. https://doi.org/10.18637/jss.v056.i06.

Kowarik, I., 1990. Some responses of flora and vegetation to urbanization in Central Europe. In: Sukopp, H., Heinÿ, S., Kowarik, I. (Eds.), Urban Ecology. Plant and Plant Communities in in Urban Environments. SPB Academic Publishing, The Hague, pp. 45–74.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 1–26. https://doi.org/10.18637/jss.v028.i05.

Little, R.J.A., 1988. A test of missing completely at random for multivariate data with missing values. J. Am. Stat. Assoc. 83 (404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722.

Little, R.J.A., Rubin, D.B., 2022. Statistical analysis with missing data. Wiley 389. https://doi.org/10.1002/9781119013563.

Molenberghs, G., Beunckens, C., Sotto, C., Kenward, M.G., 2008. Every missing not at random model has got a missing at random counterpart with equal fit. J. R. Stat. Soc. Ser. B 70, 371–388.

Monsteller, F., Tukey, J.W., 1968. Data analysis, including statistics. Handbook Soc. Psychol. 2, 80–203.

Müller, F., Burkhard, B., 2012. The indicator side of ecosystem services. Ecosyst. Serv. 1, 26–30.

Nakagawa, S., Freckleton, R.P., 2008. Missing inaction: the dangers of ignoring missing data. Trends Ecol. Evol. 23 (11), 592–596. https://doi.org/10.1016/j.tree.2008.06.014.

Nugroho, H., Utama, N.P., Surendro, K., 2018. A class center based approach for missing value imputation. Knowl.-Based Syst. 151, 124–135. https://doi.org/10.1016/j.knosys.2018.03.026.

Nugroho, H., Utama, N.P., Surendro, K., 2021. Class center-based firefly algorithm for handling missing data. J. Big Data 8 (1), 37. https://doi.org/10.1186/s40537-021-00424-y.

Onkelinx, T., Devos, K., Quataert, P., 2017. Working with population totals in the presence of missing data comparing imputation methods in terms of bias and precision. J. Ornithol. 158, 603–615. https://doi.org/10.1007/s10336-016-1404-9.

Pignatti, S., Bianco, P.M., Fanelli, G., Paglia, S., Pietrosanti, S., Tescarollo, P., 2001. Le piante come indicatori ambientali, manuale tecnico-scientifico (plants as environmental indicators, technical-scientific manual). Agenzia Nazionale Protezione Ambiente, Roma, Italy.

Pignatti, S., Menegoni, P., Pietrosanti, S., 2005. Biondicazione attraverso le piante vascolari. Valori di indicazione secondo Ellenberg (Zeigerwerte) per le specie della Flora D'italia (Biondication through vascular plants. indication values according to Ellenberg (Zeigerwerte) for the species of the Flora of Italy). Braun-Blanquetia 39, 1–97.

R Core Team, 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Schafer, J.L., Graham, J.W., 2002. Missing data: our view of the state of the art. Psychol. Methods 7 (2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147.

Schaffers, A.P., Sýkora, K.V., 2000. Reliability of ellenberg indicator values for moisture, nitrogen and soil reaction: a comparison with field measurements. J. Veg. Sci. 11 (22), 225–244. https://doi.org/10.2307/3236802.

Schaffers, A.P., Sýkora, K.V., 2002. Synecology of species-rich plant communities on roadside verges in the Netherlands. Phytocoenologia 32 (1), 29–84.

Ter Braak, C.J.F., Gremmen, N.J.M., 1987. Ecological amplitudes of plant species and the internal consistency of Ellenberg's indicator values for moisture. Vegetatio 68, 79–82.

Ter Braak, C.J.S., Juggins, S., 1993. Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from. In: species assemblages. Springer Netherlands.

Ter Braak, C.J.F., Juggins, S., Birks, H.J.B., Van der Voet, H., 1993. Weighted averaging partial least squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration. In: Patil, G.P., Rao, C.R. (Eds.), Multivariate Environmental Statistics. Elsevier Science Publishers B.V, NorthHolland), Amsterdam, pp. 525–560.

Tichý, L., Axmanová, I., Dengler, J., Guarino, R., Jansen, F., Midolo, G., Chytrý, M., 2023. Ellenberg-type indicator values for European vascular plant species. J. Veg. Sci. 34 (1), e13168.

Tsai, C.F., Li, M.L., Lin, W.C., 2018. A class center based approach for missing value imputation. Knowl.-Based Syst. 151, 124–135. https://doi.org/10.1016/j.knosys.2018.03.026.

van Buuren, S., 2018. Flexible imputation of missing data, second ed. CRC Press, Boca Raton.

van der Maarel, E., 1993. Some remarks on disturbance and its relations to diversity and stability. J. Veg. Sci. 4 (6), 733–736. https://doi.org/10.2307/3235608.

Vehtari, A., Gelman, A., Gabry, J., 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat. Comput. 27, 1413–1432. https://doi.org/10.1007/s11222-016-9696-4.

Wickham, H., François, R., Henry, L., Müller, K., Vaughan, D., 2023. dplyr: a grammar for data manipulation. R Package Version 1 (1), 4. https://dplyr.tidyverse.org.

Willner, W., Jiménez-Alfaro, B., Agrillo, E., Biurrun, I., Campos, J.A., Čarni, A., Chytrý, M., 2017. Classification of European beech forests: a Gordian Knot? Appl. Veg. Sci. 20 (3), 494–512.

Zelený, D., Schaffers, A.P., 2012. Too good to be true: pitfalls of using Ellenberg indicator values in vegetation analyses. J. Veg. Sci. 23 (3), 419–431. https://doi.org/10.1111/j.1654-1103.2011.01366.x.

Zhang, Z., 2016. Introduction to machine learning: k-nearest neighbour. Ann. Transl. Med. 4 (11). https://doi.org/10.21037/atm.2016.03.37.