First, do no harm - missing data treatment to support lake ecological state assessment

Grzegorz Chrobak [a,*], Tomasz Kowalczyk [b], Thomas B. Fischer [c], Szymon Szewrański [a], Katarzyna Chrobak [d], Jan K. Kazak [a]

[a] Institute of Spatial Management, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland,

[b] Institute of Environmental Development and Protection, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

[c] School of Environmental Sciences, Department of Geography and Planning, University of Liverpool, Liverpool, United Kingdom,

[d] Department of Urban Design & Planning and Settlement Processes, Wroclaw University of Science and Technology, Wrocław, Poland

[*] Corresponding author: Grunwaldzka St. 55, 50-357 Wroclaw, grzegorz.chrobak@upwr.edu.pl, +48 663 150 811

Abstract: Indicators of ecological potential of water bodies, that are associated with field measurements, are often subject to data gaps. This is an obstacle to constructing reliable assessments of conditions of lakes, which can lead to abandonment of assessment. Furthermore, it can lead to the use of methods, based merely on their availability. In response to these problems, a systematic approach for expert-analyst interaction for missing data treatment is proposed. In this context, a combination of algorithms with hierarchical clustering of results was used. A particular emphasis is put on the stage of preparation and interpretation of input data and the role of an expert in the workflow developed. The beneficiaries of this article are ecological data experts and analysts who work in teams to assess and interpret the state of lake ecosystems, and who present the findings in reports that are used during public consultations and discussions with key decision makers.

Keywords: ecological assessment, decision support system, missing data, lakes, water quality

1. Introduction

1.1 Data quality issues in ecological assessment

Since the publication of the Water Framework Directive in 2000, in te European Union (EU) management of water resources has become a priority, aiming to meet environmental objectives of water bodies (Di Quarto and Zinzani, 2021; Kallis and Butler, 2001). In this context, pro-ecosystem approaches require the use of methods that are based on a holistic understanding of dependencies in evaluation procedures, potentially leading to: 1) the emergence of innovative and genuine ecological approaches to water management practices (Gain et al., 2021; Giupponi, 2007; Poikane et al., 2015; Reis et al., 2017), and also to: (2) a rapid growth of methodologies, data and indicators produced by EU member states (Birk et al., 2013; Booty et al., 2001; Carey et al., 2021; Kelly et al., 2016; Zambelli et al., 2012). The number of approaches to assessing the ecological potential of water bodies is inextricably linked with issues of production, modeling and processing of observation and measurement data (Birk et al., 2012; Paruch et al., 2017; Posthuma et al., 2020). At each stage of the creation of environmental indicators, problems can arise related to the quality and availability of input values (Brito et al., 2020; Gobeyn et al., 2016; Lindholm et al., 2007; Matthies et al., 2007; Paruch et al., 2017). A key task and, at the same time, challenge are the intercalibration procedures that allow to obtain common reference levels for the classification of the ecological state of lakes. . Importantly, any unification of indicators requires a clear recognition of input data and the development of coherent methods for managing incomplete information (Gobeyn et al., 2016; Lahtinen et al., 2017). This can help to avoid undesirable consequences associated with ignoring unknowns.

1.2 Implications of missing information

The effects of a lack of data in the process of assessing ecological conditions of aquatic ecosystems can be seen at every level of data processing, including the ex-post evaluation of indicators (Yang et al., 2021; Zhang et al., 2019). Identification of the type of missing information is a critical element in the initial phase of dealing with measurement data (Little, 2021). The quantity, nature, and severity of data flaws have a direct impact on the methods that can be used to work with specific datasets. In the case of measurement sets used to assess the ecological status of lakes, deficiencies in observations often result from a type of defects, referred to as Missing At Random (MAR) (Seaman et al., 2013). In this context, there is a need to rely on substantive acceptability, as MAR is an assumption which is impossible to prove statistically (Little, 2021). Due to contingent emptiness in datasets, parameter bias can result in analyses (Schielzeth et al., 2020). How to best solve this problem depends on the assumptions made, as well as on the knowledge of the context (Koehler et al., 2017). In this context, the most common consequences of mishandling gaps in data sets include; information loss, bias in statistical inference or modeling, and results misinterpretation (Hossie et al., 2021; Noble and Nakagawa, 2021). Another problem connected with an incomplete input dataset includes an inability to use certain data analysis methods / algorithms (e.g. PCA, SVM, neural networks) (Ghannam and Techtmann, 2021). A consequence of these issues is that popular methods ted to be used, such as partial deletion, interpolation, or imputation (Curley et al., 2019; Johnson et al., 2021). Missing knowledge management requires informed decisions to be taken along the data analysis path (Likmeta et al., 2021; Newman, 2014; Wang and Xue, 2020).

1.3 Data imputation – ecological assessment perspective

The assessment of the condition and potential of aquatic ecosystems is connected with the identification of activities aimed at maintaining or improving the status of them, as required under Article 11 of the Water Framework Directive. In practice, this is associated with a planning process that takes place in a 6-year cycle. Responsible for them are water management boards together with the departments of boards of individual water sub-regions (usually within river basins). Water administration are working together on: identifying anthropogenic pressures; updating environmental objectives and protected water areas; restoring water bodies,; and setting boundary values for heavily modified and artificial water bodies. An important stage is the preparation of strategic environmental assessment (SEA; Mustow, 2021). At this key moment, assessors have the opportunity to influence the shape of the analyzes, the interpretation of the results. Furthermore, they can apply for supplementing or correcting the methodology. Comments are directed to the authors of the plan at the stage of public consultations. Among other measures, indicators of the ecological status of lake ecosystems are used to obtain results that support the definition of management practices. The evaluation of the structure and efficiency of surface water ecosystems is known as ecological status. This demonstrates how stresses (such as pollution and habitat deterioration) have an impact on specified quality components. Each surface water body has an ecological status that is assessed based on biological quality components and supported by physico-chemical and hydromorphological quality elements. According to the "one out, all out" approach, the element with the worst status out of all biological and supporting quality factors determines the overall ecological status rating for a water body. Data used to evaluate the ecological status of lakes are sets largely based on the results of field measurements. Observations are prone to errors that can occur at the stage of collecting samples (Yanai et al., 2021). There is always uncertainty over results, even if using different tools (Ejigu, 2021). Loss of data or a complete lack of it may result in abandoning the assessment, which, in some cases, significantly reduces the pool of evaluated ecosystems. This often leads to gaps in data sets that weaken results of individual measurement campaigns. Moreover, the same input data serve as components necessary to construct different environmental indicators, placing additional emphasis on the validity of an imputation attempt. In research on the ecological quality of ecosystems, various methods of supplementing missing values are used (Muharemi et al., 2019; Said et al., 2019; Zhang and Thorburn, 2022).

The so-called hot deck imputation is used for handling missing data on large scale water quality indices (Ahmed et al., 2021; Srebotnjak et al., 2012). Most extensively used are methods based on multiple imputation. These are available for most data types (Ben Aissia et al., 2017; Betrie et al., 2016; Neri et al., 2018; Ngouna et al., 2020). When faced with a high level of missingness data, machine learning techniques were adopted. These are able to troubleshoot complex data issues (Irvin et al., 2021; Kim et al., 2020; Ngouna et al., 2020; Ratolojanahary et al., 2019; Rodríguez et al., 2021). Furthermore, the spatial nature of the issue results in an introduction of time and space variables (Koki et al., 2020; Labuzzetta et al., 2021; Liu et al., 2016; Lou and Obradovic, 2011; Sojka et al., 2020; Yüksel, 2012; Zhang and Thorburn, 2021). Research with ecological water quality indicators uses methods based on a case study approach. This confirmins their effectiveness at the local scale (Bilgin and Bayraktar, 2021; Liu et al., 2011; Ren et al., 2008; Sojka et al., 2019; Weerasinghe and Handapangoda, 2019). There is a noticeable trend in the research indicating the need to develop methods that work well at the regional level, providing the option of later intercalibration of the results (Akbar et al., 2011; Botha et al., 2020; Hu et al., 2018; Jiang et al., 2017; Lepš and Šmilauer, 2006; Li et al., 2021; Luo et al., 2019). Holistic approaches facilitate macro-quality management of water resources, which is important in the context of policy design and pan-regional impact assessment. Moreover, monitoring of ecological indicators and the impact of climate change on phenomena that threaten the stability of ecosystems has lately been explored (Cheruvelil et al., 2017; Fazli et al., 2018; Hutjes, 2019; Krzeminski et al., 2019; Lizotte et al., 2014; Mankin et al., 1999; Mustajoki et al., 2004; Peters-Lidard et al., 2021).

1.4 Research goals structure of paper

The main goal of the research underpinning this paper was to present a workflow that can be used when an expert group or an ecological assessor are faced with the problem of missing values in an input dataset. In this context, a novel combined expert and analyst approach to ecological assessment is introduced. This approach gives experts the opportunity to influence (and adjust) processes by making decisions in key nodes. A further goal is the identification of possible techniques of data visualization, both with regards to raw data and analysis.

In the methodological approach, graphic representation of often complicated processes is crucial for effective cooperation in an expert team. Featured data treatment schema takes the specificity of the work of experts into account, dealing with various assessment objects with a different degree of data incompleteness in the assessment process. Thus, there are certain cross-roads highlighted where a decision is necessary, made by a specialist or requiring consultation before proceeding with the analysis (2.3. Proposed workflow). The data treatment framework guides the user through the steps of pre-selecting data (3.1. Missing data identification and triage), identifying and selecting imputation predictors (3.2. Predictor examination), the actual multiple data imputation process using the random forest algorithm (3.3. Missing data imputation), and then introduces the step of clustering similar complementary sets based on their characteristics in the context of the Ward criterion via hierarchical clustering (3.4. Clustering imputation and 3.5. Data imputation results).

2. Materials and Methods

2.1 Data, software & previous research

The input data used in this work come from the resources of the Chief Inspectorate of Environmental Protection in Poland (Appendix A) (GIOŚ, 2015). These are measurements included in the data used to develop indicators of the ecological condition of lake ecosystems. Results of the analyzes are reported to the European Commission data repositories, including information on the state of water among the Member States of the European Union (European Environment Agency, 2018) (Figure 1).

The analyzes concern a set of 499 objects for which measurements were made during the 2013-2015 measurement campaign. Chlorophyll a, nitrogen, phosphorus, phytoplankton, Ecological State Macrophyte Index (ESMI), Diatom Index for Lakes (IOJ), Phytoplankton Method for Polish Lakes (PMPL), visibility, and conductivity are some of the measures used to determine a lake’s ecological status. The basic information on data is provided in Appendix B. Data were the subject of constructing a methodology aimed at improving effectiveness and reproducibility of the procedure for determining ecological status indicators with the use of machine learning algorithms (Chrobak et al., 2021b). In the next step, the set was used to extend the methodological approach to include the use of an unsupervised tool, supporting the prioritization of lakes in the context of organizing remedial measures necessary for the ecosystem to achieve environmental goals (Chrobak et al., 2021a). An important element of working with data at each of these stages was the need to deal with the problem of missing information. In this paper, the consequences of the lack of observations in the collection are addressed, and the missing data imputation is performed and tested as a complementary solution working with workflow previously developed.

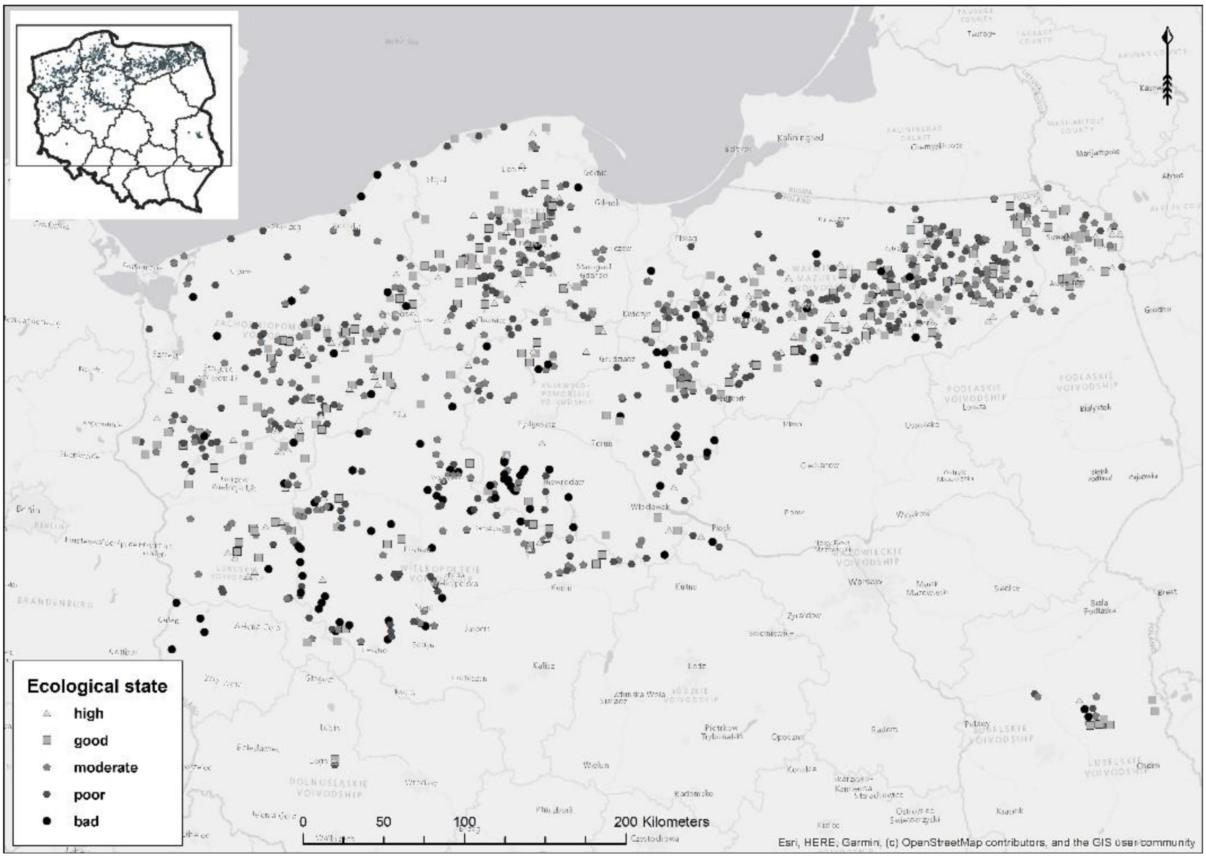

Figure 1. The map presenting lakes with shapes representing resulting ecological state according to EU classes. Originally, missing data were not included in the calculations. Instead, they influenced the appropriate value of the uncertainty of the result in the tables attached to the report.

2.2 Imputation and clustering techniques

In order to select the optimal technique for imputation of missing observations, the 'missingness' type of the dataset was identified (Zhou, 2020). The discovered systematic tendencies in the dataset show that missing observations can be predicted with use of other information present (see section 3.2 Predictor examination). It is due to existing correlations between fields and thanks to the knowledge of the data collection procedure that errors in measurements or deficiencies are not the result of a deliberate procedure. Thus, the missingness type was labelled as missing at random (MAR) (Bhaskaran and Smeeth, 2014). The following procedure of iterative imputation of missing values was preceded by stages (1); which involved applying the Pearson product-moment correlation method to analyze the degree and direction of data association (Russo, 2021) and (2); Principal Component Analysis (PCA) performed on the dataset with missing values to investigate uncertainty related absence of information (Husson et al., 2014).

Missing value imputation was done using methods of Multiple Imputation by Chained Equations (MICE) for multivariate dataset cases (Zhang, 2016). The goal to replace missing values with plausible data to estimate a more realistic layout dataset, which is affected only minimally by incomplete observations. Within the procedure, the following steps were performed on the input dataset (Raghunathan et al., 2001):

Step 1. For every missing value in the dataset, random extraction is performed from non-missing data to provide initial, basic imputation (*D*).

Step 2. The field with the least missing values ratio (*f*) is selected and transformed back to feature missing values.

Step 3. The *f* is regressed as a dependent variable onto the initially imputed dataset as $f \sim D$.

Step 4. The predicted values obtained from a regression model are used to fill missing data in *f*. Both, the non-missing and imputed values are used once *f* acts as an independent variable in regression modeling for the following dependent variables.

Step 5. Steps 2-4 are repeated for each variable with missing data identified. One iteration is understood as an operation of cycling through each of the variables. The cycle is finished once all missing values are replaced with regression predictions that match the data relationships observed in the initial dataset.

The MICE model parameters selected for this research are:

a) dataset: matrix (8 x 499) with missing values,
b) data imputation method: random forest imputation (Shah et al., 2014),
c) visit sequence: roman (left to right).

About 10 iteration cycles are performed in most research tasks (Gelman et al., 2011). However, at the conclusion of iterative cycles, the distribution of the imputation parameters (for instance, the regression model coefficients) should have converged and become stable. In order to eliminate the undesired dependency on the sequence in which variables are imputed, the authors performed 50 iterations until reaching convergence (Figure 2). The algorithm performance resulted in 30 imputed datasets, which were subject to a distribution-based clustering process.

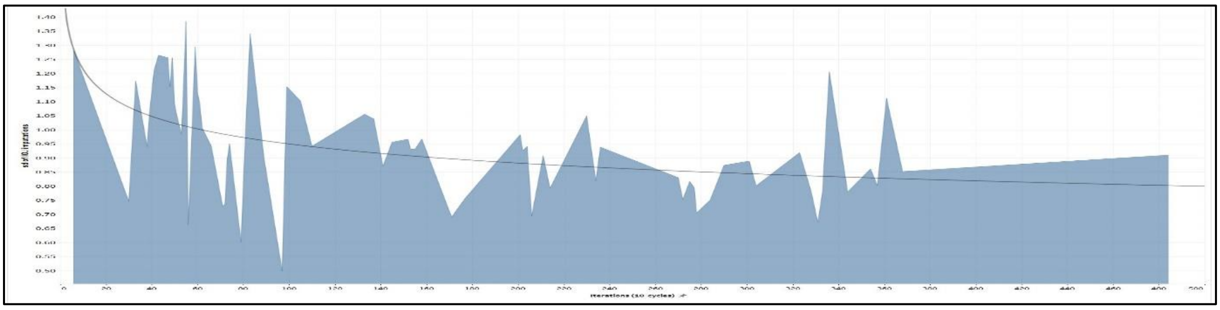

Figure 2. The formation of standard deviation for successive imputation cycles led to selection of 50 initial iterations as a default parameter for this analysis.

For each of the fields with missing values, as a result of the data imputation method, 30 versions of the possible information supplementation were obtained. The hierarchical clustering technique was used to select the imputation sets that correspond to the formation of the original variable in the context of the parameters of the similarity of the data distribution (Wu et al., 2009). Initially, each dataset was treated as a separate cluster in the agglomerative version of the algorithm. Following that, similar clusters were merged to form larger units based on predefined rules. When only one cluster emerged, the algorithm concluded that no further agglomeration is possible (Murtagh and Legendre, 2014). The clustering procedure included the following steps: (Hartigan and Wong, 1979):

Step 1. The distance matrix was computed between columns of versions of imputed columns (the original field is a feature in proximity calculation, as well, with missing values allowed, but excluded from analysis) – resulting in a cross-distance matrix.

Step 2. A cross-distance matrix was used as a dissimilarity structure for an agglomeration method to perform proximity-based merging – every column was considered as an individual cluster.

Step 3. The clusters with similar characteristics (proximity) were merged.

Step 4. The cross-distance matrix was recalculated for each cluster.

Step 5. The steps 3-4 were repeated until a single cluster remained.

In the construction of the cross-distance matrix for each of the dataset fields, the form of squared Euclidean distance matrix was used (Sarstedt and Mooi, 2014). The Ward's method, based on the optimal value of an objective function – in this case – the minimum variance was used as a criterion for choosing a pair of clusters to merge at each step (Ward, 1963). The overall within-cluster variance is reduced, using Ward's minimal variance criterion (Kruskal and Black, 2012):

$$D_{1,2} = \sqrt{\frac{2 \cdot |k| \cdot |l|}{|k| + |l|}} \cdot \left\| \underset{k}{\rightarrow} - \underset{l}{\rightarrow} \right\|$$

where:

$D_{1,2}$ – dissimilarity between cluster 1 and cluster 2,

k,l – observations from cluster 1 and cluster 2,

$\underset{k}{\rightarrow}, \underset{l}{\rightarrow}$ – centroids for clusters 1 and 2,

$\| . \|$ – Euclidean norm.

For using this approach, the pair of clusters was selected that, after merging, resulted in the least amount of total within-cluster variance. A weighted squared distance between cluster centers was used to calculate this increase.

All clusters were singletons in the first stage (clusters containing a single point). The initial distance between individual objects was proportional to the squared Euclidean distance in order to execute a recursive algorithm under the objective function (Everitt, 1980) as:

$$D_{i,j} = \sum_{v}^{d} \left(x_{v_i} - x_{vj}\right)^2$$

where:

$D_{i,j}$ – distance between cells i and j,

x_{v_i} – value of x variable at cell i,

d – number of dataset dimensions.

Every feasible cluster pair is examined at each phase, and the two clusters whose merger results in the least amount of information loss are combined. Ward defines information loss in terms of an error sum-of-squares criterion (ESS) (Ward, 1963):

$$ESS = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

where:

n – number of observations,

x_i – the value of i-tj observation.

and *0* being mean value of all the observations.

2.3 Proposed workflow

Within the block diagram of the suggested method, the proposed data analysis processes for the efficient imputation of missing values have been systematized (Figure 3). The workflow was created to supplement the methodology described in the authors' previous works on optimizing the assessment of the ecological state of lake ecosystems (Chrobak et al., 2021a, 2021b). This enabled the evaluation solutions to be tailored to the framework imposed by the Water Framework Directive, which indicates the need to conduct assessments involving expert knowledge. From the technical point of view, the approach addresses cases where the analysis cannot be performed effectively due to a significant number of missing observations. Thus, the decision whether to continue the analysis with use of data imputation is made by the expert, who is guided by experience and aided with dataset recognition led by skilled analyst. The aim is to obtain reliable premises for the implementation legitimacy of subsequent steps of ecological assessment process. In the diagram of the analytical process shown below, the dataset objects (lakes with measurements) appear as rectangles with blue border. Purple-outlined hexagonal blocks denote an analytical or computational process that could produce new data objects or serve as the basis for decision-making. In some places, these blocks are linked to orange-colored square blocks. In these cases, an expert decision is advised. Given the number or severity of missing observations, the expert may decide to end the process. If the process is not stopped during the data triage stage (section 3.1), the dataset is subjected to multivariate imputation, the results of which are clustered. The sets of imputations proposed by the algorithm are reviewed again by an expert, who is supported by the clustering results. Finally, the selected dataset with no missing values is submitted to further analyses, serving as an input for the supervised classifier of the lake ecological state class. The operation of such a classifier was described in the work that was published before this research (Chrobak et al., 2021b).

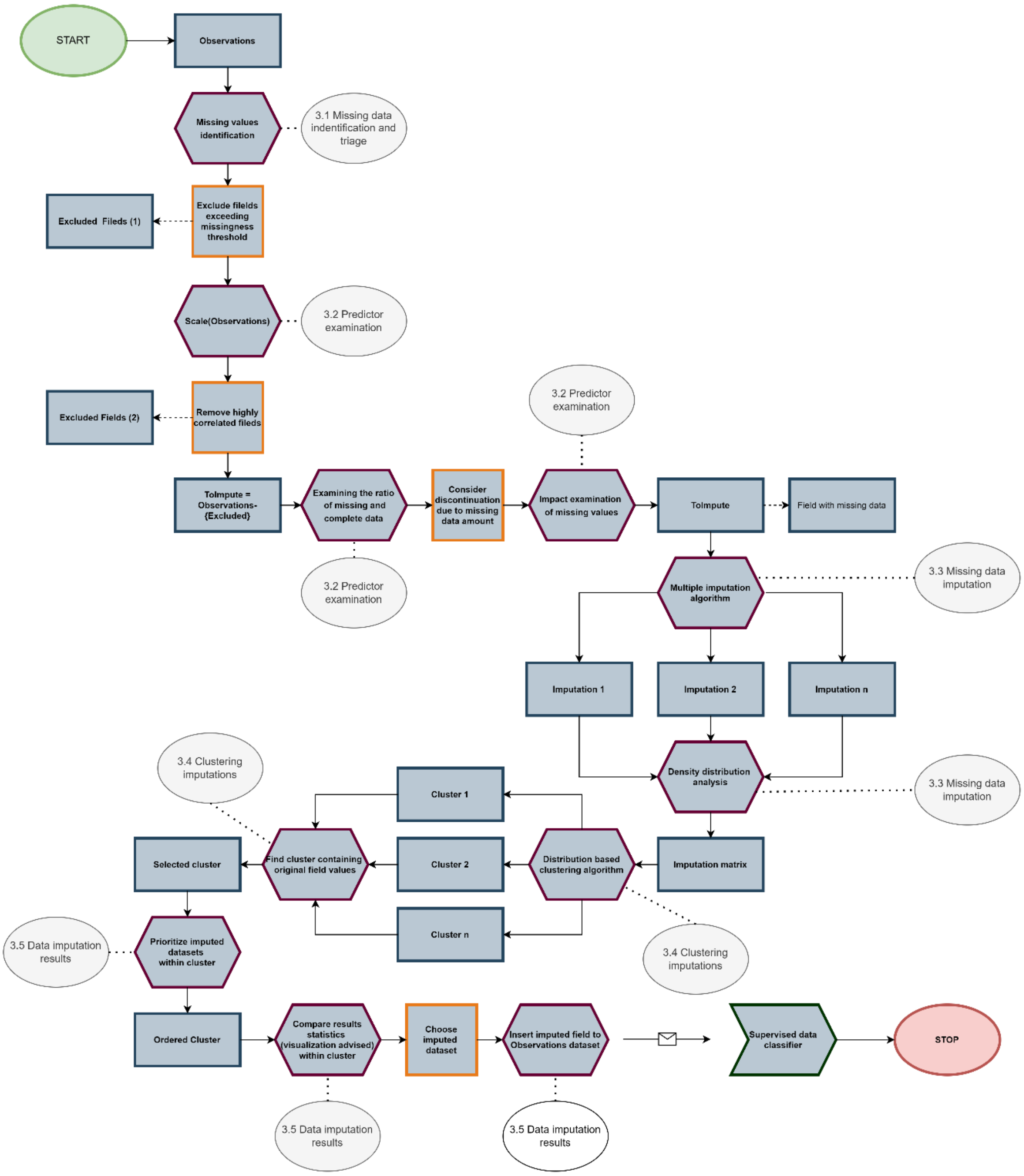

Figure 3. The workflow of missing data curation and imputation. The main purpose of arranging the steps taken into a procedural form is to systematize the methodology so that it is reproducible. Each of the process blocks enclosed by a purple frame symbolizes the action on the data. The squares with an orange frame indicate the moment of the decision made by the analyst / expert. Each of the steps of the analysis is discussed along with an example of implementation in the following subsections of this article.

3. Results

3.1 Missing data identification and triage

The input data of the analysis were characterized by a different number and structure of missing measurements. The identification of the shortcomings started with the preparation of the chart showing the scale of the problem

(Figure 4). According to the adopted classification, the so-called "missing grade", deficiencies were identified in 5 out of 8 variables used in the process of assessing the ecological condition of lakes (Khorshidi et al., 2020). The spread of NA's percentage ranged from 0.2% for the conductivity variable to 15% for the IOJ parameter. It is worth noting that the fields containing the measurement results for ESMI and IOJ together account for the existence of approx. 80% of the deficiencies. Moreover, these deficiencies are characterized in the adopted methodology of data triage as NotBad (missing <= 20% values), where the deficiencies in the field of PMPL, chlorophyll a, and conductivity are labeled as Good (less than 5% missing). Despite the lack of fields with the Bad category, it is important to remember that (1) the categories are arbitrary intervals that are largely dependent on the decision of an expert who knows the data; and (2) it is possible that there are gaps in the intersection data that, when accumulated at the intersections, will give a picture of real losses in the set of measurements' quality. IOJ and ESMI parameters are components that strongly affect the results of ecological status classification, as indicated by the PCA analysis by Chrobak et al., 2021. Leaving these fields out of the analysis may cause the final result to be skewed.

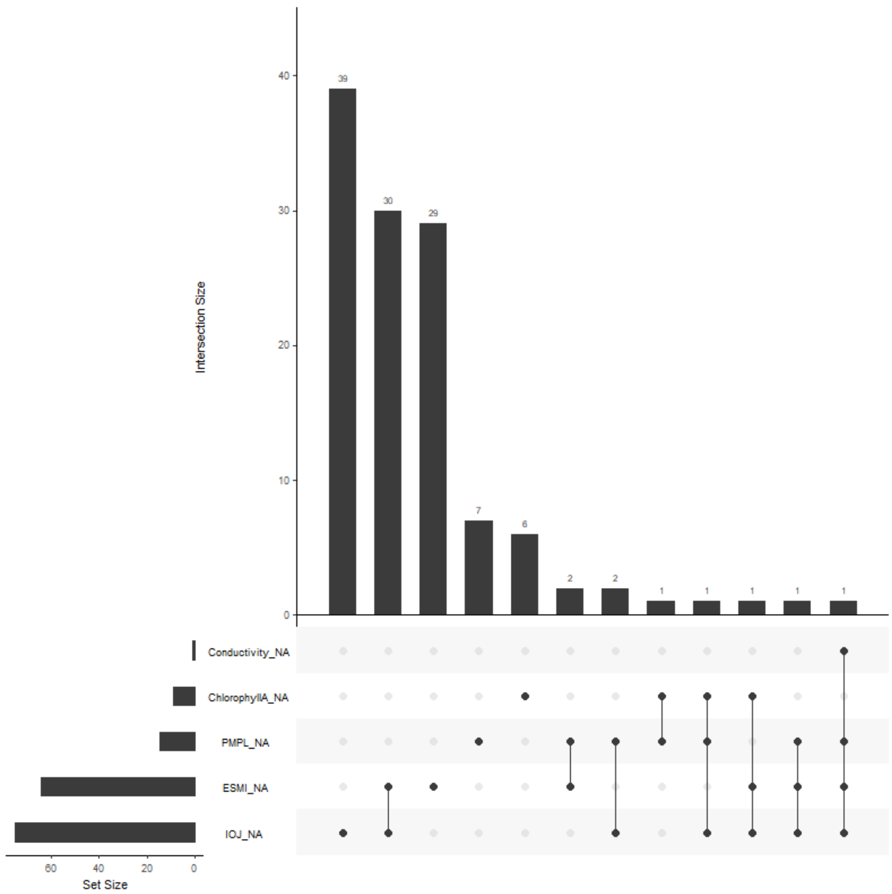

Figure 4. The visual representation of missing values across the dataset indicated deficiencies in five out of eight variables involved in the construction of the lake evaluation index. In addition, the number of objects (39) that have information gaps for more than one field is also indicated. The analysis did not reveal any cases where the object has gaps for each of the variables. The fields to note are IOJ and ESMI, together accounting for 80% of existing NA statements, which is a prerequisite for taking corrective action on the data.

3.2 Predictor examination

One of the data preprocessing steps, crucial for later decisions made during data imputation, is the exploratory analysis of predictors (Braun and Oswald, 2011). The variables were subjected to the analysis of mutual linear dependencies, which allowed for an assumption of the situation earlier referred to as MAR in the context of missing observations. Strong correlations (> = | 0.5|) were identified, e.g. for visibility-PMPL or nitrogen-chlorophyll pairs (Figure 5). Variables that are strongly associated with each other are not preferred candidates for following multiple data imputation (Ellington et al., 2015). In most situations, the selected imputation method should omit these variables during the algorithm implementation (Alice, 2015). For some instances, it is also possible for algorithms to fail or produce unreliable, overfitted results (Christie et al., 1984). Thus, highly correlated variables were excluded from the imputation process. For each of the fields with missing values a separate selection of predictors was performed, on the basis of which the calculations were continued. As a result, in the case of the IOJ variable, each of the possible predictors was qualified (the weakest correlation concerned the relationship with nitrogen, the strongest with phosphorus). The following predictors were related to the ESMI index: phosphorus, IOJ, and conductivity. For the imputation of the field containing the PMPL measurement results, the variables: IOJ, conductivity, and phosphorus were specified. IOJ and conductivity variables were used to supplement deficiencies in the chlorophyll field. It can be seen that the IOJ variable, which is one of the imputation objects, has no correlations identified in the data set, which would rule out using any of the variables due to the concern about multicollinearity-induced bias. In that case, multiple imputation was performed using all of the available predictors. The PCA plot shows the effect of IOJ on data variability in the dataset (Figure 4).Furthermore, the plot includes variable-wise uncertainty due to the presence of empty observations (Husson et al., 2018). The analysis demonstrates that variability across different possible imputation scenarios is limited, implying that PCA results may be perceived as plausible by a user (Benahmed and Houichi, 2018). It also shows the need to monitor the impact of data imputation on the shaping of leading dimensions' explanatory skills (Chrobak et al., 2021b).

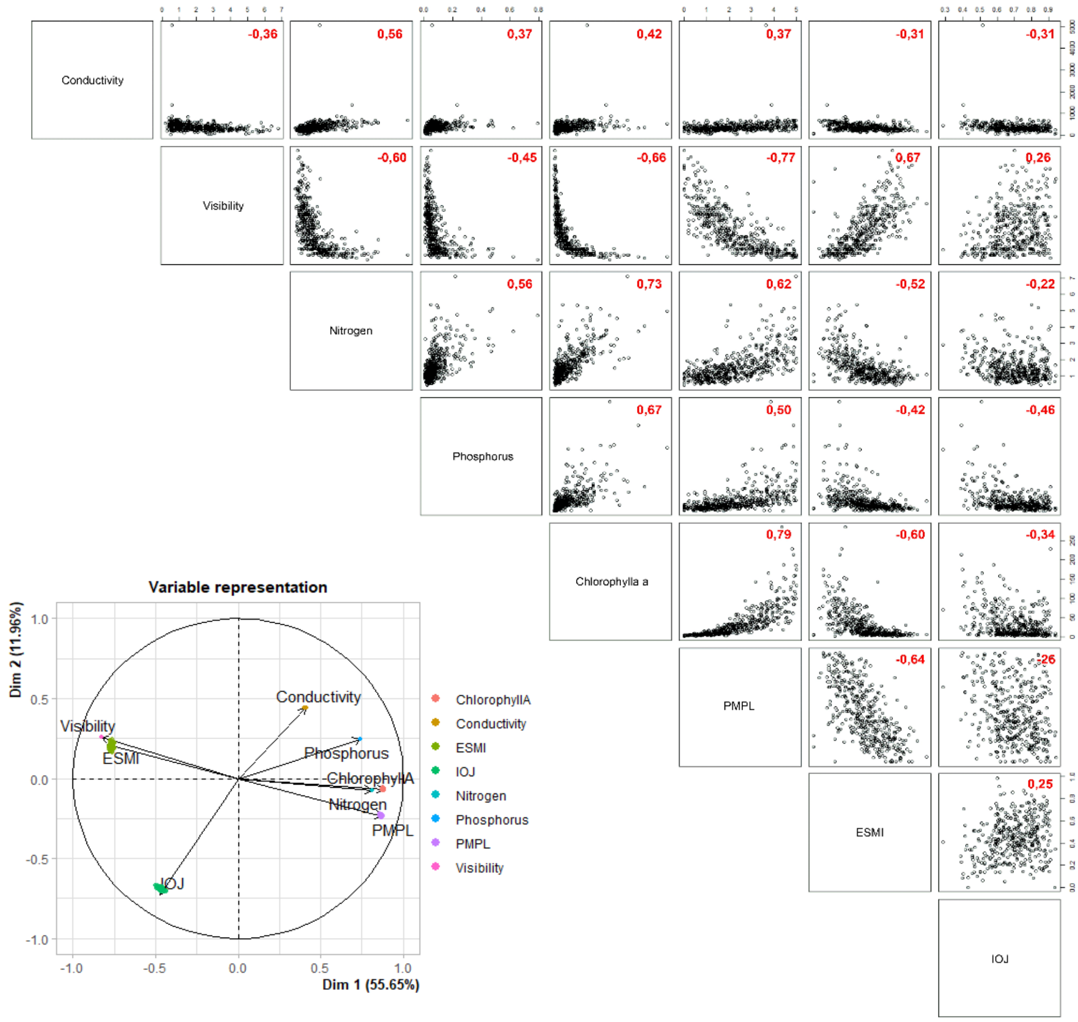

Figure 5. Evaluation of predictors preceding the data imputation process. Correlation analysis using the Pearson product-moment correlation coefficient method indicated the existence of a linear relationship between some sets of observations. This information was used to select potential predictors of imputation of missing values. The results visible on the vector PCA indicated the importance of the IOJ parameter affecting the diversity of the data set, which affects the final ability of the variable to explain the differentiation in the shaping of the first coordinate variance in the reduced observation space.

3.3 Missing data imputation

Missing data imputation concerned four variables (IOJ, ESMI, PMPL, chlorophyll a), for which individual sets of predictors were selected in the previous stage of work. The applied method of multiple imputation is the MICE approach, using the random forest algorithm (Xiao and Bulut, 2020). The method is effective when linear relationships exist between variables and does not require the use of hyperparameter calibration practices. The distributions were assumed for each variable and imputation was performed according to the distribution characteristics obtained from the original, non-imputed dataset (Figure 6). It is not possible to know the true value of intercept term due to missing data in the source field, thus introduction of a distribution assumptions was necessary.

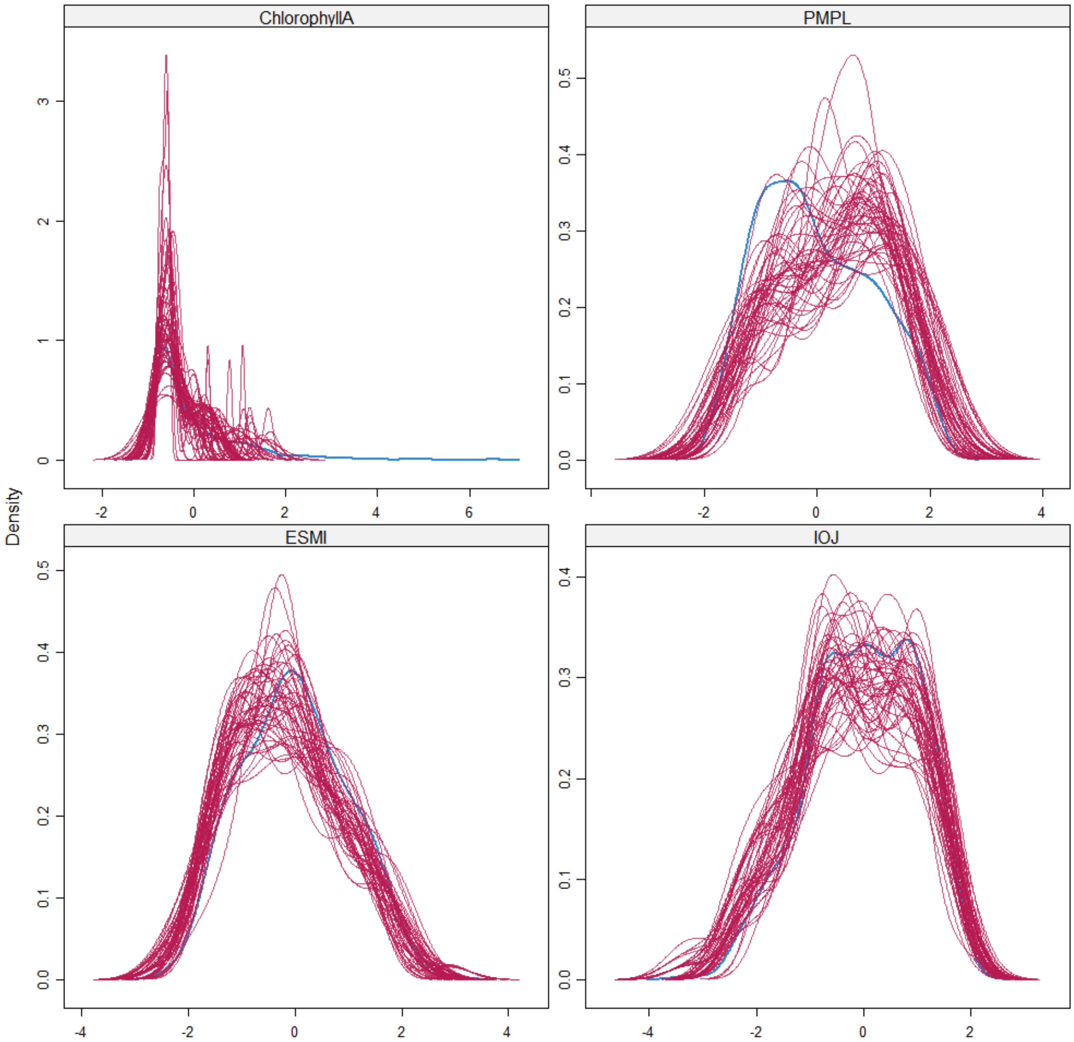

Figure 6. The density plots for each of imputation dataset are showed in red. The density of original field is displayed as blue line. The dataset desired to be the best imputation option is expected to be similar in context of data density distribution. However, different results for individual iterations of the 'rf' algorithm do not give an unambiguous fit of the optimal solution. The results also indicate the necessity of continuous monitoring of the model results in order to avoid the use of distributions, the parameters of which (e.g. kurtosis) differ significantly from the expected fit.

The selection of the set of possible imputations was carried out for the IOJ variable as a presentation of the functioning of the approach in practice. According to the results of the PCA analysis and the identification of predictors, it is a variable that significantly influences the result of the final classification of the ecological state of lakes in the adopted methodology. Gaps in observations of 15% make it an indicator that has the potential to be the most difficult imputation, compared with e.g. chlorophyll (<5%). The plot in Figure 4 indicates the presence of imputation sets that may result in an optimal but not overfitted match (Radosavljevic and Anderson, 2014).

3.4 Clustering imputations

According to the scheme of proceedings presented in the Materials & Methods section (Figure 3), the grouping of similar imputations was performed, using the hierarchical clustering method (Cohen-Addad et al., 2019). The aim of this part of the analysis was to use a tool that allows for fairly intuitive and quick interpretation of a given set of imputation sets, bearing in mind the possibility of carrying out more imputation iterations in specific cases or, if necessary, indicating many supplementary series (scenarios). In order to minimize the cluster-associated variance loss the Ward's method was applied, so that, at each algorithm performance step, the combination of every possible cluster pair was considered. It this case, the information loss was defined in terms of an error sum of squares criterion (ESS). Each of the leaves of the resulting dendrogram referred to the series obtained in the multiple imputation process. Sets of similar observations according to Ward's criterion were collected under the dendrogram branch (Figure 7). The height parameter of the combination displayed on the x axis indicated the similarity measure between two sets. Seven clusters within the data set were defined, using the so-called gap statistic method, which compared the total intra-cluster variation for different cluster quantities with their expected values under null reference distribution of the dataset generated with use of Monte Carlo simulations during the sampling procedure (Tibshirani et al., 2001). The original series of IOJ containing the missing observations (marked as 31) was introduced to the analysis, for which the distribution estimation was performed (Figure 7). The source set of observations was included within one cluster, marked as 4 with the sets: 1, 3, 20, and 24, which in the next steps will be considered as plausible and safe imputation options with regards to variance and distribution criteria. The distance obtained by the pair of objects 1 and 31 significantly differed from the other objects within cluster 4. Despite the fact that it indicates the best match according to the adopted criteria, it is advisable to perform a similarity test (e.g. z-statistic) in order to recognize the differences between the objects cluster (Ben-Zvi, 2004).

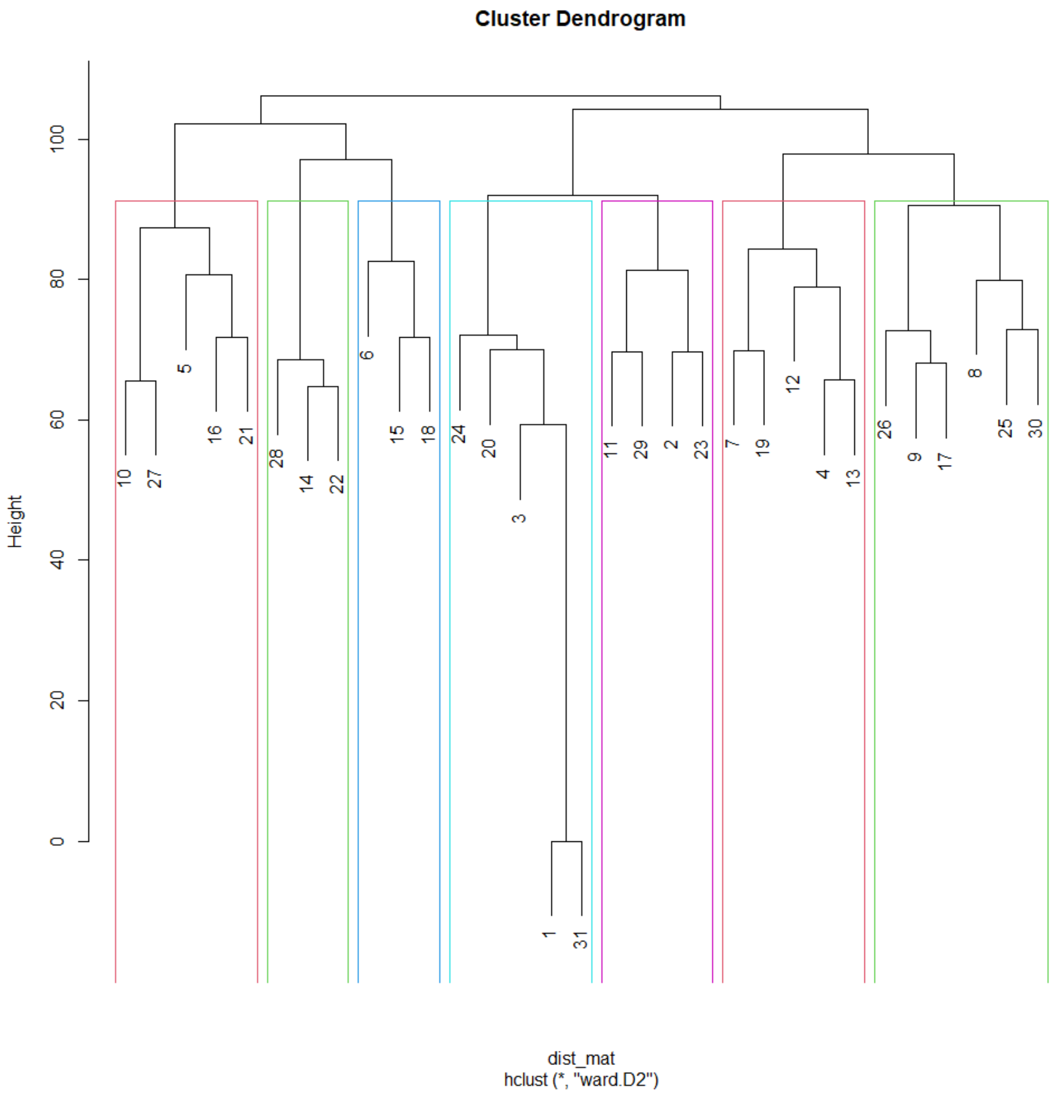

Figure 7. The dendrogram created for set of plausible imputation options for IOJ variable was based on bottom-up, distribution based hierarchical clustering algorithm. During consecutive model runs, seven separate clusters were distinguished. The 4th cluster (enclosed with blue frame) contains the original IOJ variable (marked with the number 31) entered for the analysis. High similarity in the context of distribution was recognized for imputation set no. 1. The next options of field completion with similar distribution are found in sets: 3, 20, and 24. The sets from the fourth cluster, in the given prioritization order, constituted a pool of plausible solutions to the problem of missing values.

3.5 Data imputation results

The results of data imputation for the IOJ variable were presented in the form of sequences of corresponding series, arranged according to Lake ID in the original dataset (Figure 8). It allowed for the tracing of the imputation process within Cluster No. 4, as well as the final verification of the results, using polynomial regression on each of the retrieved series. Treating the process-aspect approach to data imputation is one of the most informative ways of presenting the process-aspect approach to data imputation. It proved to be highly informative to decision-makers and water-quality experts during the presentation of results and project-group meetings. The second way for visualizing the imputation process is to arrange lakes in order of catchment area, allowing for simultaneous assessment of the degree of missing observations in spatial terms (Figure 9). The method also makes it simple to partition the sets so that specialists working on specific catchment assessments can accurately evaluate the scope of the problem in their work area and compare it to the situation in other task groups. Furthermore, the visualization enables for cross-referencing of individual implementation outcomes across the cluster (red dashed lines) and tracing of the data imputation process to identify undesired outliers generated by the method used.

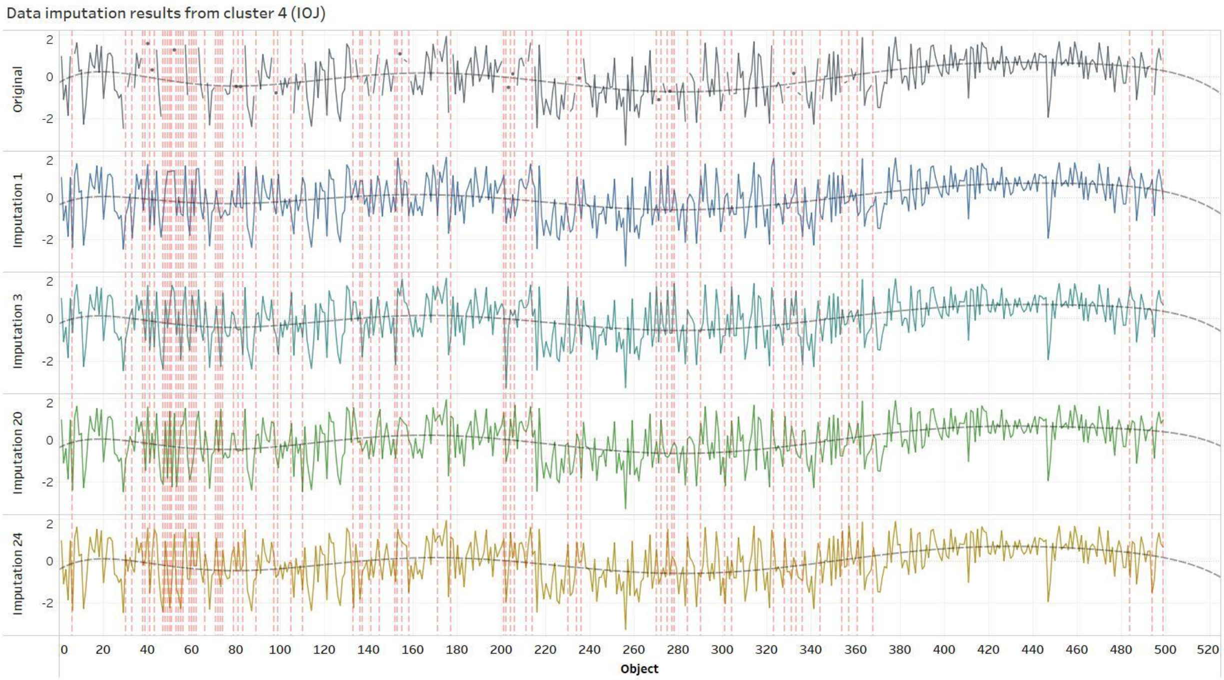

Figure 8. The chart shows a compilation of the four imputation sets (in order of priority) against the original IOJ value evolution of the evaluation set. Dashed red lines indicate where data imputation has been performed. For each of the options within the cluster no. 4, the statistics of the shaping of the variable allow for "safe" imputation of data and the use of the set in subsequent analyzes on the way to obtain a reliable indicator of the ecological condition of lakes.

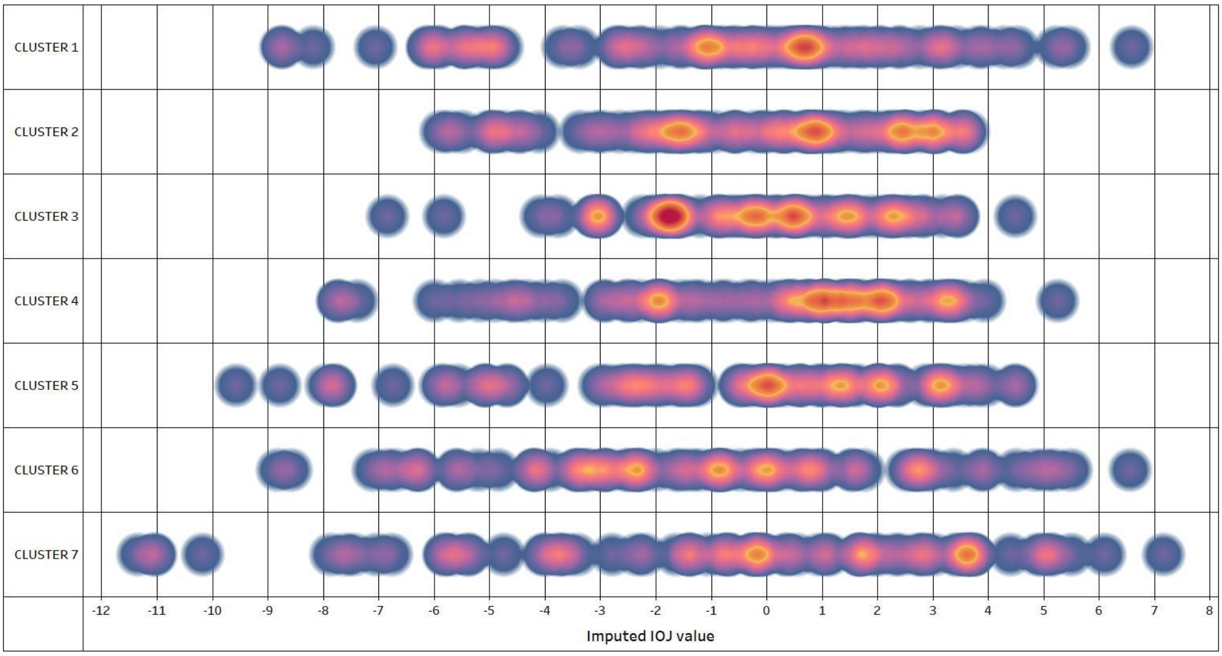

Figure 9. The distribution of clustered imputations shows point concentrations around values which triggered separation. The characteristics of each cluster can be distinguished during the reverse reasoning making it possible to determine entry requirements for next iterations of imputation algorithm when assessing dataset obtained in currently ongoing data collection campaign.

4. Discussion

This research study underlying this paper focused on how to deal with missing-at-random data curation and imputation in the process of assessing the ecological status of lake ecosystems. The study was based on a collection of 499 lakes in Poland, with missing values detected to various degrees. A methodology was designed, based on the authors' knowledge and support in the field of expert evaluations, allowing for the imputation of data gaps to be implemented. The technique is demonstrated with an example from an authentic dataset used in the ecological status assessment with the goal of submitting the results to European Union bodies in relation to WFD obligations (Reyjol et al., 2014). The presented scheme of conduct is a complementary element to the previous works, where the stage of incomplete information management is part of an extensive algorithm of ecological assessment of lakes. The tools used in the study allowed for the selection of four ranked propositions of value imputation for the IOJ index, which was characterized by a 15% share of incomplete values. Data imputation, especially in the case of the identification of relatively large gaps in data sets (e.g.> 5%), is always associated with the risk of introducing bias into the process, which may negatively (‘mis-informatively’) affect the final results and their interpretation (Krueger, 2017). As a result, it's critical to understand the facts and intentionally employ the various strategies for addressing flaws. Testing the susceptibility of values to outliers is a useful practice which is part of the input data recognition stage (Jackson and Chen, 2004). Due to the emerging need to analyze lakes in a regional (or sub-basin) perspective, the future role of ecological status indicators, which will be used to make decisions at higher (supralocal) levels of water resource quality management, should be taken into account (Mammides, 2020; Rivera-Rondón and Catalan, 2020; Wu et al., 2021). It is connected with going beyond the locally understood and evaluated indicators (Baldera et al., 2018; Kraemer et al., 2020). This is one of the challenges of the ecological evaluation of aquatic ecosystems, as the management of gaps in large-scale data requires the development of methods of analyzing the relationships between indicators and their components in the context of spatial and temporal relationships between the objects of assessment (Kolada et al., 2014; Rossaro et al., 2012; Werner et al., 2016). This may ultimately lead to the observation of a phenomenon referred to as data drift, defined as a difference in variation of the data used to construct an initial assessment framework and the observations feeding the assessment model in the next round of reporting (Brock and Carpenter, 2012; Koehnken et al., 2020). Taking the changes in ecosystems and their internal relationships into account, especially in the era of the identified impact of climate change effects, new factors may affect the variability of the ecological state of lakes over time. Thus, it is critical to create a consistent procedure for detecting data drift, defining drift percentage criteria, and configuring pro-active alerts so that the necessary action may be performed (Dong et al., 2018; Gupta et al., 2020). Shift may manifest itself in the data at the level of their covariate shift, therefore steering with data imputation should minimize the effect of completions on the distribution of the variable (Hilt et al., 2017; Martin et al., 2020).

The clustering approach used in this work to select plausible options is an alternative solution to the pooling stage within the multiple imputation process. The classification algorithm used is, comparatively speaking, easy to interpret (Cohen-Addad et al., 2019). The user also does not need to define the number of clusters a-priori. However, during the process arbitrary decisions are made (distance metric, linkage criterion), which prompts the expert to monitor the results in order to react quickly to noticeable errors, e.g. related to the use of mixed data types (Karthikeyan et al., 2020; Zhang et al., 2013). In addition, the algorithm is sensitive to the increase in the number of dimensions in the data, so an iterative analysis of successive variables requiring imputation is recommended (Contreras and Murtagh, 2015). The Ward criterion used allowed for the creation of clusters based on a minimal increase in degree in within cluster variance making the approach less susceptible to noise related to multiple imputation results (McInnes et al., 2017).

Thus, the main limitations of the proposed approach are of two types. First, in terms of the algorithms used, the method inherits some of their inherent limitations. In the case of the applied data imputation using the MICE method with the use of random forest function, the limitations result from the need to control the results of supplements. The expert should control the process so as not to allow indiscriminate acceptance of results significantly deviating from the observed data. This may affect the second element of the process, which is hierarchical clustering, which is sensitive to the presence of noise and outliers. This applies to both the original input data and the imputation results. The second type of limitation is also related to noise, however, it concerns noise generated on the side of expert judgment. The method does not allow for the complete elimination of cognitive errors resulting from the participation of expert decisions characterized by their own systematic noise or bias.

One of the indirect limitations of the whole assessment system, which this methodology also inherits, results from the dependence on measurement timing and hydrological background for subsequent analyzes. As the analysts work within a given time window, the measurement reports contain data that represent the ecological situation of the reservoir considered to characterize it in terms of "typical state". In practice, this means that the samples of the studied variables from the extremal hydrological periods (drought, flood) are included in the reports for separate analyzes in the research dealing with extraordinary situations. Thus, the relationship between extraordinary measures and “normal” periods is neglected. Undoubtedly, periods of ecological stress can affect the quality and values of measurements, being for example a delayed ecosystem response to critical phenomena. Although striving for normality of results through their early averaging and sampling in arbitrarily selected "typical" periods has a mitigating effect on the variance of results, the noise generated at the early stage of the assessment is not measured at present.

An important positive effect of the proposed imputation process is leading the data set to the smooth transition of subsequent evaluation steps, where specialists often use tools that function only with non-missing input. Due to the key nature of the input data management process, the transparency aspect of the analytical procedures used is not without significance (Romañach et al., 2014; Zasada et al., 2017). Methods that include data visualizations as inseparable elements of data processing are beneficial to supporting the ability to explain actions taken, especially at the level of expert - decision makers interactions, which are critical for the often overlooked data-sense making stage of ecological assessment (Arciniegas et al., 2013).

5. Conclusions

The missing data treatment scheme presented in the paper is aimed at systematizing the value imputation stage so that it is possible to perform an efficient, reproducible solution ready to implement within existing lake ecological state assessment methods. The analyses included eight variables. There were gaps in the measurement data for five of them. The number of missing items indicated the need to imputate data for four variables. An approach was used based on random forest multiple imputation with predictors examination. A hierarchical algorithm with a

Ward's variance minimization criterion was used to cluster plausible imputation solutions obtained in previous step. There were seven clusters of similar additions found. Cluster 4. contained the original data set as well as four completed sets that met the membership criteria. The results were presented as a dendrogram in the case of the selection of clusters, as well as with the help of ordered trajectories of the shaping of the variable for the set containing missing values in relation to the four possible supplementary series according to the adopted criteria. The stage of missing data treatment was indicated as an integral part of the process of assessing the ecological condition of lakes, influencing the selection of modeling and classification methods in subsequent stages of analyzes related to the proper ecological assessment and prioritization of ecosystems in terms of the selection of remedial solutions. The authors note the positive impact of methodological and visual communication on the experts-analyst-decision maker line, which should be carried out with the transparency of the process (Moallemi et al., 2020). This can be facilitated, for example, with the use of available data visualization techniques. This research concludes the three-step approach to lake ecological assessment, which now consists of 1) data preprocessing and missing values treatment, 2) model-based assessment, and 3) lake prioritization for remedial purposes. Taking into account the holistic view of the research results, the proposed solutions are aimed at systematization of the process of supplementing gaps in data on measurements, in contrast to the previous omission of this issue in the reports on the assessment of the ecological state of lakes. The role of the expert limnologist was also unclear in the course the analyzes. As a result, some lakes were only assessed by experts, while others using analytical approaches. Some of the assessments were carried over from previous measurement campaigns. This resulted in a conflict of results in the event that the lake apparently did not achieve environmental objectives, despite the implemented remedial measures. Thus, a certain kind of data-result asymmetry occurred. The proposed fragment of the methodology was therefore aimed at organizing the assessment process by: 1) defining the role of an expert in the course of analyzes, 2) introducing a consistent methodology of data pre-processing, which will be passed to expert judgment only in the next steps, 3) enabling the use of effective algorithms in the assessment, which are sensitive to data deficiencies (e.g. kSVM or PCA), and 4) enabling the preview of the entire assessment process so that it can be corrected or further improved in the future. With reference to the results of the next campaign to assess the ecological status of waters, future research should focus on assessing the scale of the phenomenon of ecological data drift, which, based on the observed climate change, anthropological pressure and loss of biodiversity, may have a significant impact on the broad concept indicator construction for lake water ecological assessment.

6. Software and data availability

The research was conducted with use of software providing: data visualization (*Tableau 2021.1.1, https://www.tableau.com/*), data modelling (*R 4.0.5 via RStudio 1.4.1106 „Tiger Daylily", https://www.r-project.org/, https://rstudio.com/*), and algorithm development (*draw.io 15.9.1*, https://www.diagrams.net/). Appendix B contains an R language script that converts all of the analysis procedures in this paper into an executable, reproducible workflow. The materials for this work are available from the HydroSource platform: https://www.hydroshare.org/resource/ebec024018be4c2ba04cbfa85bb14d8e/ in the repository titled "LakeEcoMissingData". Accessed as Resource: a) R-code for data preprocessing, imputation and clustering as "LakesMissingRcode.R", b) XML file of featured workflow schema as "LakeMissingWorkFlow", c) CSV file containing raw measurement results treated as input to this analysis, d) a set of results of the statistical analysis of the variables involved in the study.

7. Literature

Ahmed, M., Mumtaz, R., Zaidi, S.M.H., 2021. Analysis of water quality indices and machine learning techniques for rating water pollution: A case study of Rawal Dam, Pakistan. Water Supply. https://doi.org/10.2166/ws.2021.082

Akbar, T.A., Hassan, Q.K., Achari, G., 2011. A Methodology for Clustering Lakes in Alberta on the basis of Water Quality Parameters. Clean - Soil, Air, Water. https://doi.org/10.1002/clen.201100050

Alice, M., 2015. Imputing missing data with R; MICE package | R-bloggers. R-bloggers.

Arciniegas, G., Janssen, R., Rietveld, P., 2013. Effectiveness of collaborative map-based decision support tools: Results of an experiment. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2012.02.021

Baldera, A., Hanson, D.A., Kraft, B., 2018. Selecting indicators to monitor outcomes across projects and multiple restoration programs in the Gulf of Mexico. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2018.01.025

Ben-Zvi, D., 2004. Reasoning about variability in comparing distributions. Stat. Educ. Res. J.

Ben Aissia, M.A., Chebana, F., Ouarda, T.B.M.J., 2017. Multivariate missing data in hydrology – Review and applications. Adv. Water Resour. https://doi.org/10.1016/j.advwatres.2017.10.002

Benahmed, L., Houichi, L., 2018. The effect of simple imputations based on four variants of PCA methods on the quantiles of annual rainfall data. Environ. Monit. Assess. https://doi.org/10.1007/s10661-018-6913-y

Betrie, G.D., Sadiq, R., Tesfamariam, S., Morin, K.A., 2016. On the Issue of Incomplete and Missing Water-

Quality Data in Mine Site Databases: Comparing Three Imputation Methods. Mine Water Environ. https://doi.org/10.1007/s10230-014-0322-4

Bhaskaran, K., Smeeth, L., 2014. What is the difference between missing completely at random and missing at random? Int. J. Epidemiol. https://doi.org/10.1093/ije/dyu080

Bilgin, A., Bayraktar, H.D., 2021. Assessment of lake water quality using multivariate statistical techniques and chlorophyll-nutrient relationships: a case study of the Göksu Lake. Arab. J. Geosci. https://doi.org/10.1007/s12517-021-06871-4

Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., Solimini, A., Van De Bund, W., Zampoukas, N., Hering, D., 2012. Three hundred ways to assess Europe's surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2011.10.009

Birk, S., Willby, N.J., Kelly, M.G., Bonne, W., Borja, A., Poikane, S., van de Bund, W., 2013. Intercalibrating classifications of ecological status: Europe's quest for common management objectives for aquatic ecosystems. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2013.03.037

Booty, W.G., Lam, D.C.L., Wong, I.W.S., Siconolfi, P., 2001. Design and implementation of an environmental decision support system. Environ. Model. Softw. https://doi.org/10.1016/S1364-8152(01)00016-0

Botha, E.J., Anstee, J.M., Sagar, S., Lehmann, E., Medeiros, T.A.G., 2020. Classification of Australian waterbodies across a wide range of optical water types. Remote Sens. https://doi.org/10.3390/RS12183018

Braun, M.T., Oswald, F.L., 2011. Exploratory regression analysis: A tool for selecting models and determining predictor importance. Behav. Res. Methods. https://doi.org/10.3758/s13428-010-0046-8

Brito, A.C., Garrido-Amador, P., Gameiro, C., Nogueira, M., Moita, M.T., Cabrita, M.T., 2020. Integrating in situ and ocean color data to evaluate ecological quality under the water framework directive. Water (Switzerland). https://doi.org/10.3390/w12123443

Brock, W.A., Carpenter, S.R., 2012. Early Warnings of Regime Shift When the Ecosystem Structure Is Unknown. PLoS One. https://doi.org/10.1371/journal.pone.0045586

Carey, C.C., Woelmer, W.M., Lofton, M.E., Figueiredo, R.J., Bookout, B.J., Corrigan, R.S., Daneshmand, V., Hounshell, A.G., Howard, D.W., Lewis, A.S.L., McClure, R.P., Wander, H.L., Ward, N.K., Thomas, R.Q., 2021. Advancing lake and reservoir water quality management with near-term, iterative ecological forecasting. Inl. Waters. https://doi.org/10.1080/20442041.2020.1816421

Cheruvelil, K.S., Yuan, S., Webster, K.E., Tan, P.N., Lapierre, J.F., Collins, S.M., Fergus, C.E., Scott, C.E., Henry, E.N., Soranno, P.A., Filstrup, C.T., Wagner, T., 2017. Creating multithemed ecological regions for macroscale ecology: Testing a flexible, repeatable, and accessible clustering method. Ecol. Evol. https://doi.org/10.1002/ece3.2884

Christie, A.A., Kennelley, M.D., William King, J., Schaefer, T.F., 1984. Testing for incremental information content in the presence of collinearity. J. Account. Econ. https://doi.org/10.1016/0165-4101(84)90025-9

Chrobak, G., Kowalczyk, T., Fischer, T.B., Chrobak, K., Szewrański, S., Kazak, J.K., 2021a. Combining indicators for better decisions – Algorithms vs experts on lakes ecological status assessment. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2021.108318

Chrobak, G., Kowalczyk, T., Fischer, T.B., Szewrański, S., Chrobak, K., Kazak, J.K., 2021b. Ecological state evaluation of lake ecosystems revisited: Latent variables with kSVM algorithm approach for assessment automatization and data comprehension. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2021.107567

Cohen-Addad, V., Kanade, V., Mallmann-Trenn, F., Mathieu, C., 2019. Hierarchical clustering: Objective functions and algorithms. J. ACM. https://doi.org/10.1145/3321386

Contreras, P., Murtagh, F., 2015. Hierarchical clustering, in: Handbook of Cluster Analysis. https://doi.org/10.1201/b19706

Curley, C., Krause, R.M., Feiock, R., Hawkins, C. V., 2019. Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database. Urban Aff. Rev. https://doi.org/10.1177/1078087417726394

Di Quarto, F., Zinzani, A., 2021. European environmental governance and the post-ecology perspective: a critical analysis of the Water Framework Directive. GeoJournal. https://doi.org/10.1007/s10708-021-10402-9

Dong, F., Zhang, G., Lu, J., Li, K., 2018. Fuzzy competence model drift detection for data-driven decision support systems. Knowledge-Based Syst. https://doi.org/10.1016/j.knosys.2017.08.018

Ejigu, M.T., 2021. Overview of water quality modeling. Cogent Eng. https://doi.org/10.1080/23311916.2021.1891711

Ellington, E.H., Bastille-Rousseau, G., Austin, C., Landolt, K.N., Pond, B.A., Rees, E.E., Robar, N., Murray, D.L., 2015. Using multiple imputation to estimate missing data in meta-regression. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.12322

Europe Environment Agency, 2018. Ecological status of surface water bodies [WWW Document]. Eur. Environ. Inf. Obs. Netw.

Everitt, B., 1980. Cluster analysis. Qual. Quant. https://doi.org/10.1007/BF00154794

Fazli, B., Shafie, A., Mohamed, A., Mohamad, M.F., Yahaya, N.K.E.M., Noordin, N., 2018. Development of spatial similarity-based modelling to improve integrated lake water quality management in Malaysia. Lakes Reserv. Res. Manag. https://doi.org/10.1111/lre.12204

G.-Tóth, L., Poikane, S., Penning, W.E., Free, G., Mäemets, H., Kolada, A., Hanganu, J., 2008. First steps in the Central-Baltic intercalibration exercise on lake macrophytes: Where do we start? Aquat. Ecol. https://doi.org/10.1007/s10452-008-9184-9

Gain, A.K., Hossain, S., Benson, D., Di Baldassarre, G., Giupponi, C., Huq, N., 2021. Social-ecological system approaches for water resources management. Int. J. Sustain. Dev. World Ecol. https://doi.org/10.1080/13504509.2020.1780647

Gelman, A., Levy, M.A., Abayomi, K., 2011. Diagnostics for Multivariate Imputations. SSRN Electron. J. https://doi.org/10.2139/ssrn.1010415

Ghannam, R.B., Techtmann, S.M., 2021. Machine learning applications in microbial ecology, human microbiome studies, and environmental monitoring. Comput. Struct. Biotechnol. J. https://doi.org/10.1016/j.csbj.2021.01.028

GIOŚ, 2015. Bank danych pomiarowych [WWW Document]. URL https://powietrze.gios.gov.pl/pjp/archives

Giupponi, C., 2007. Decision Support Systems for implementing the European Water Framework Directive: The MULINO approach. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2005.07.024

Gobeyn, S., Bennetsen, E., Van Echelpoel, W., Everaert, G., Goethals, P.L.M., 2016. Impact of abundance data errors on the uncertainty of an ecological water quality assessment index. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2015.07.031

Gupta, O., Goyal, N., Anand, D., Kadry, S., Nam, Y., Singh, A., 2020. Underwater Networked Wireless Sensor Data Collection for Computational Intelligence Techniques: Issues, Challenges, and Approaches. IEEE Access. https://doi.org/10.1109/ACCESS.2020.3007502

Hartigan, J.A., Wong, M.A., 1979. Algorithm AS 136: A K-Means Clustering Algorithm. Appl. Stat. https://doi.org/10.2307/2346830

Hilt, S., Brothers, S., Jeppesen, E., Veraart, A.J., Kosten, S., 2017. Translating Regime Shifts in Shallow Lakes into Changes in Ecosystem Functions and Services. Bioscience. https://doi.org/10.1093/biosci/bix106

Hossie, T.J., Gobin, J., Murray, D.L., 2021. Confronting Missing Ecological Data in the Age of Pandemic Lockdown. Front. Ecol. Evol. https://doi.org/10.3389/fevo.2021.669477

Hu, Y., Peng, J., Liu, Y., Tian, L., 2018. Integrating ecosystem services trade-offs with paddy land-to-dry land decisions: A scenario approach in Erhai Lake Basin, southwest China. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2017.12.340

Husson, F., Josse, J., Le, S., Mazet, J., 2018. FactoMineR: multivariate exploratory data analysis and data mining. J. Stat. Softw.

Husson, F., Josse, J., Le, S., Mazet, J., 2014. Multivariate exploratory data analysis and data mining with R. R Packag. version 1.26.

Hutjes, R., 2019. Service for Water Indicators in Climate Change Adaptation (SWICCA) [WWW Document]. Web page.

Irvin, J., Zhou, S., McNicol, G., Lu, F., Liu, V., Fluet-Chouinard, E., Ouyang, Z., Knox, S.H., Lucas-Moffat, A., Trotta, C., Papale, D., Vitale, D., Mammarella, I., Alekseychik, P., Aurela, M., Avati, A., Baldocchi, D., Bansal, S., Bohrer, G., Campbell, D.I., Chen, J., Chu, H., Dalmagro, H.J., Delwiche, K.B., Desai, A.R., Euskirchen, E., Feron, S., Goeckede, M., Heimann, M., Helbig, M., Helfter, C., Hemes, K.S., Hirano, T., Iwata, H., Jurasinski, G., Kalhori, A., Kondrich, A., Lai, D.Y., Lohila, A., Malhotra, A., Merbold, L., Mitra, B., Ng, A., Nilsson, M.B., Noormets, A., Peichl, M., Rey-Sanchez, A.C., Richardson, A.D., Runkle, B.R., Schäfer, K.V., Sonnentag, O., Stuart-Haëntjens, E., Sturtevant, C., Ueyama, M., Valach, A.C., Vargas, R., Vourlitis, G.L., Ward, E.J., Wong, G.X., Zona, D., Alberto, M.C.R., Billesbach, D.P., Celis, G., Dolman, H., Friborg, T., Fuchs, K., Gogo, S., Gondwe, M.J., Goodrich, J.P., Gottschalk, P., Hörtnagl, L., Jacotot, A., Koebsch, F., Kasak, K., Maier, R., Morin, T.H., Nemitz, E., Oechel, W.C., Oikawa, P.Y., Ono, K., Sachs, T., Sakabe, A., Schuur, E.A., Shortt, R., Sullivan, R.C., Szutu, D.J., Tuittila, E.S., Varlagin, A., Verfaillie, J.G., Wille, C., Windham-Myers, L., Poulter, B., Jackson, R.B., 2021. Gap-filling eddy covariance methane fluxes: Comparison of machine learning model predictions and uncertainties at FLUXNET-CH4 wetlands. Agric. For. Meteorol. https://doi.org/10.1016/j.agrformet.2021.108528

Jackson, D.A., Chen, Y., 2004. Robust principal component analysis and outlier detection with ecological data. Environmetrics. https://doi.org/10.1002/env.628

Jannicke Moe, S., Schartau, A.K., Bækken, T., Mcfarland, B., 2010. Assessing macroinvertebrate metrics for classifying acidified rivers across northern Europe. Freshw. Biol. https://doi.org/10.1111/j.1365-2427.2010.02413.x

Jiang, Q., Liang, Z., Zhao, L., Li, Y., Wu, S., Liu, Y., 2017. Integrated PCA-BN Approach for Identifying the Water Quality Response Patterns for Lakes in Yunnan Plateau. Beijing Daxue Xuebao (Ziran Kexue

Ban)/Acta Sci. Nat. Univ. Pekin. https://doi.org/10.13209/j.0479-8023.2017.113

Johnson, T.F., Isaac, N.J.B., Paviolo, A., González-Suárez, M., 2021. Handling missing values in trait data. Glob. Ecol. Biogeogr. https://doi.org/10.1111/geb.13185

Kallis, G., Butler, D., 2001. The EU water framework directive: Measures and implications. Water Policy. https://doi.org/10.1016/S1366-7017(01)00007-1

Karthikeyan, B., George, D.J., Manikandan, G., Thomas, T., 2020. A comparative study on k-means clustering and agglomerative hierarchical clustering. Int. J. Emerg. Trends Eng. Res. https://doi.org/10.30534/ijeter/2020/20852020

Kelly, M.G., Birk, S., Willby, N.J., Denys, L., Drakare, S., Kahlert, M., Karjalainen, S.M., Marchetto, A., Pitt, J.A., Urbanič, G., Poikane, S., 2016. Redundancy in the ecological assessment of lakes: Are phytoplankton, macrophytes and phytobenthos all necessary? Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2016.02.024

Khorshidi, H.A., Kirley, M., Aickelin, U., 2020. Machine learning with incomplete datasets using multi-objective optimization models, in: Proceedings of the International Joint Conference on Neural Networks. https://doi.org/10.1109/IJCNN48605.2020.9206742

Kim, Z., Jeong, H., Shin, S., Jung, J., Kim, J.H., Ki, S.J., 2020. Characterizing water quality and quantity profiles with poor quality datin a machine learning algorithm. Desalin. Water Treat. https://doi.org/10.5004/dwt.2020.25481

Kindsvater, H.K., Dulvy, N.K., Horswill, C., Juan-Jordá, M.J., Mangel, M., Matthiopoulos, J., 2018. Overcoming the Data Crisis in Biodiversity Conservation. Trends Ecol. Evol. https://doi.org/10.1016/j.tree.2018.06.004

Koehler, M., Bogatu, A., Civili, C., Konstantinou, N., Abel, E., Fernandes, A.A.A., Keane, J., Libkin, L., Paton, N.W., 2017. Data context informed data wrangling, in: Proceedings - 2017 IEEE International Conference on Big Data, Big Data 2017. https://doi.org/10.1109/BigData.2017.8258015

Koehnken, L., Rintoul, M.S., Goichot, M., Tickner, D., Loftus, A.C., Acreman, M.C., 2020. Impacts of riverine sand mining on freshwater ecosystems: A review of the scientific evidence and guidance for future research. River Res. Appl. https://doi.org/10.1002/rra.3586

Koki, I.B., Low, K.H., Zain, S.M., Juahir, H., Bayero, A.S., Azid, A., Zali, M.A., 2020. Spatial variability in surface water quality of lakes and ex-mining ponds in malacca, malaysia: The geochemical influence. Desalin. Water Treat. https://doi.org/10.5004/dwt.2020.25982

Kolada, A., Willby, N., Dudley, B., Nõges, P., Søndergaard, M., Hellsten, S., Mjelde, M., Penning, E., Van Geest, G., Bertrin, V., Ecke, F., Mäemets, H., Karus, K., 2014. The applicability of macrophyte compositional metrics for assessing eutrophication in European lakes. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2014.04.049

Kraemer, S.A., Barbosa da Costa, N., Shapiro, B.J., Fradette, M., Huot, Y., Walsh, D.A., 2020. A large-scale assessment of lakes reveals a pervasive signal of land use on bacterial communities. ISME J. https://doi.org/10.1038/s41396-020-0733-0

Krueger, T., 2017. Bayesian inference of uncertainty in freshwater quality caused by low-resolution monitoring. Water Res. https://doi.org/10.1016/j.watres.2017.02.061

Kruskal, J.B., Black, P., 2012. Ward’s hierarchical agglomerative clustering method: Which algorithms implement Ward’s criterion? J. Classif.

Krzeminski, P., Tomei, M.C., Karaolia, P., Langenhoff, A., Almeida, C.M.R., Felis, E., Gritten, F., Andersen, H.R., Fernandes, T., Manaia, C.M., Rizzo, L., Fatta-Kassinos, D., 2019. Performance of secondary wastewater treatment methods for the removal of contaminants of emerging concern implicated in crop uptake and antibiotic resistance spread: A review. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2018.08.130

Labuzzetta, C., Zhu, Z., Chang, X., Zhou, Y., 2021. A submonthly surface water classification framework via gap-fill imputation and random forest classifiers of landsat imagery. Remote Sens. https://doi.org/10.3390/rs13091742

Lahtinen, T.J., Hämäläinen, R.P., Liesiö, J., 2017. Portfolio decision analysis methods in environmental decision making. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2017.04.001

Lazaridou, M., Ntislidou, C., Karaouzas, I., Skoulikidis, N., Birk, S., 2018. Harmonization of the assessment method for classifying the ecological quality status of very large Greek rivers. Knowl. Manag. Aquat. Ecosyst. https://doi.org/10.1051/kmae/2018038

Lepš, J., Šmilauer, P., 2006. Multivariate Analysis of Ecological Data. Bull. Ecol. Soc. Am. https://doi.org/10.1890/0012-9623(2006)87[193:maoed]2.0.co;2

Li, J., Tian, L., Wang, Y., Jin, S., Li, T., Hou, X., 2021. Optimal sampling strategy of water quality monitoring at high dynamic lakes: A remote sensing and spatial simulated annealing integrated approach. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2021.146113

Likmeta, A., Metelli, A.M., Ramponi, G., Tirinzoni, A., Giuliani, M., Restelli, M., 2021. Dealing with multiple experts and non-stationarity in inverse reinforcement learning: an application to real-life problems. Mach. Learn. https://doi.org/10.1007/s10994-020-05939-8

Lindholm, O., Greatorex, J.M., Paruch, A.M., 2007. Comparison of methods for calculation of sustainability indices for alternative sewerage systems-Theoretical and practical considerations. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2005.10.002

Little, R.J., 2021. Missing data assumptions. Annu. Rev. Stat. Its Appl. https://doi.org/10.1146/annurev-statistics-040720-031104

Liu, J., Liu, Q., Yang, H., 2016. Assessing water scarcity by simultaneously considering environmental flow requirements, water quantity, and water quality. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2015.07.019

Liu, W.C., Yu, H.L., Chung, C.E., 2011. Assessment of water quality in a subtropical alpine lake using multivariate statistical techniques and geostatistical mapping: a case study. Int. J. Environ. Res. Public Health. https://doi.org/10.3390/ijerph8041126

Lizotte, R.E., Knight, S.S., Locke, M.A., Bingner, R.L., 2014. Influence of integrated watershed-scale agricultural conservation practices on lake water quality. J. Soil Water Conserv. https://doi.org/10.2489/jswc.69.2.160

Lou, Q., Obradovic, Z., 2011. Modeling multivariate spatio-temporal remote sensing data with large gaps, in: IJCAI International Joint Conference on Artificial Intelligence. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-287

Luo, W., Zhu, S., Wu, S., Dai, J., 2019. Comparing artificial intelligence techniques for chlorophyll-a prediction in US lakes. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-019-06360-y

Lyche Solheim, A., Rekolainen, S., Moe, S.J., Carvalho, L., Phillips, G., Ptacnik, R., Penning, W.E., Toth, L.G., O'Toole, C., Schartau, A.K.L., Hesthagen, T., 2008. Ecological threshold responses in European lakes and their applicability for the Water Framework Directive (WFD) implementation: Synthesis of lakes results from the REBECCA project. Aquat. Ecol. https://doi.org/10.1007/s10452-008-9188-5

Mammides, C., 2020. A global assessment of the human pressure on the world's lakes. Glob. Environ. Chang. https://doi.org/10.1016/j.gloenvcha.2020.102084

Mankin, K.R., Koelliker, J.K., Kalita, P.K., 1999. Watershed and lake water quality assessment: An integrated modeling approach. J. Am. Water Resour. Assoc. https://doi.org/10.1111/j.1752-1688.1999.tb04194.x

Martin, R., Radosavljevic, S., Schlüter, M., 2020. Short-term decisions in lake restoration have long-term consequences for water quality. Reg. Environ. Chang. https://doi.org/10.1007/s10113-020-01643-4

Matthies, M., Giupponi, C., Ostendorf, B., 2007. Environmental decision support systems: Current issues, methods and tools. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2005.09.005

McInnes, L., Healy, J., Astels, S., 2017. hdbscan: Hierarchical density based clustering. J. Open Source Softw. https://doi.org/10.21105/joss.00205

Moallemi, E.A., Zare, F., Reed, P.M., Elsawah, S., Ryan, M.J., Bryan, B.A., 2020. Structuring and evaluating decision support processes to enhance the robustness of complex human–natural systems. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2019.104551

Muharemi, F., Logofătu, D., Leon, F., 2019. Machine learning approaches for anomaly detection of water quality on a real-world data set. J. Inf. Telecommun. https://doi.org/10.1080/24751839.2019.1565653

Murtagh, F., Legendre, P., 2014. Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion? J. Classif. https://doi.org/10.1007/s00357-014-9161-z

Mustajoki, J., Hämäläinen, R.P., Marttunen, M., 2004. Participatory multicriteria decision analysis with Web-HIPRE: A case of lake regulation policy. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2003.07.002Mustow, S. E. 2021. Strategic environmental assessment in the water sector, in: Fischer, T. B. and González, A. (eds.). *Handbook on Strategic Environmental Assessment*, Cheltenham: Edward Elgar (chapter 13).

Neri, L., Coscieme, L., Giannetti, B.F., Pulselli, F.M., 2018. Imputing missing data in non-renewable empower time series from night-time lights observations. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2017.08.040

Newman, D.A., 2014. Missing Data: Five Practical Guidelines. Organ. Res. Methods. https://doi.org/10.1177/1094428114548590

Ngouna, R.H., Ratolojanahary, R., Medjaher, K., Dauriac, F., Sebilo, M., Junca-Bourié, J., 2020. A data-driven method for detecting and diagnosing causes of water quality contamination in a dataset with a high rate of missing values. Eng. Appl. Artif. Intell. https://doi.org/10.1016/j.engappai.2020.103822

Noble, D.W.A., Nakagawa, S., 2021. Planned missing data designs and methods: Options for strengthening inference, increasing research efficiency and improving animal welfare in ecological and evolutionary research. Evol. Appl. https://doi.org/10.1111/eva.13273

Paruch, A.M., 2014. The impact of wastewater irrigation on the chemical quality of groundwater. Water Environ. J. https://doi.org/10.1111/wej.12064

Paruch, L., Paruch, A.M., Blankenberg, A.G.B., Haarstad, K., Mæhlum, T., 2017. Norwegian study on microbial source tracking for water quality control and pollution removal in constructed wetland treating catchment run-off. Water Sci. Technol. https://doi.org/10.2166/wst.2017.303

Peters-Lidard, C.D., Rose, K.C., Kiang, J.E., Strobel, M.L., Anderson, M.L., Byrd, A.R., Kolian, M.J., Brekke, L.D., Arndt, D.S., 2021. Indicators of climate change impacts on the water cycle and water management. Clim. Change. https://doi.org/10.1007/s10584-021-03057-5

Poikane, S., Birk, S., Böhmer, J., Carvalho, L., De Hoyos, C., Gassner, H., Hellsten, S., Kelly, M., Lyche Solheim, A., Olin, M., Pall, K., Phillips, G., Portielje, R., Ritterbusch, D., Sandin, L., Schartau, A.K., Solimini, A.G., Van Den Berg, M., Wolfram, G., Van De Bund, W., 2015. A hitchhiker's guide to European lake ecological assessment and intercalibration. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2015.01.005

Posthuma, L., Zijp, M.C., De Zwart, D., Van de Meent, D., Globevnik, L., Koprivsek, M., Focks, A., Van Gils, J., Birk, S., 2020. Chemical pollution imposes limitations to the ecological status of European surface waters. Sci. Rep. https://doi.org/10.1038/s41598-020-71537-2

Radosavljevic, A., Anderson, R.P., 2014. Making better Maxent models of species distributions: Complexity, overfitting and evaluation. J. Biogeogr. https://doi.org/10.1111/jbi.12227

Raghunathan, T., Lepkowski, J., Van Hoewyk, J., Solenberger, P., 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Surv. Methodol.

Ratolojanahary, R., Ngouna, R.H., Medjaher, K., Dauriac, F., Sebilo, M., 2019. Groundwater quality assessment combining supervised and unsupervised methods, in: IFAC-PapersOnLine. https://doi.org/10.1016/j.ifacol.2019.10.054

Reis, S., Voigt, K., Oxley, T., 2017. Thematic issue on modelling human and ecological health risks. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2017.02.029

Ren, C., Li, C., Jia, K., Zhang, S., Li, W., Cao, Y., 2008. Water quality assessment for Ulansuhai Lake using fuzzy clustering and pattern recognition. Chinese J. Oceanol. Limnol. https://doi.org/10.1007/s00343-008-0339-2

Reyjol, Y., Argillier, C., Bonne, W., Borja, A., Buijse, A.D., Cardoso, A.C., Daufresne, M., Kernan, M., Ferreira, M.T., Poikane, S., Prat, N., Solheim, A.L., Stroffek, S., Usseglio-Polatera, P., Villeneuve, B., van de Bund, W., 2014. Assessing the ecological status in the context of the European Water Framework Directive: Where do we go now? Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2014.07.119

Rivera-Rondón, C.A., Catalan, J., 2020. Diatoms as indicators of the multivariate environment of mountain lakes. Sci. Total Environ. https://doi.org/10.1016/j.scitotenv.2019.135517

Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., Gorgoglione, A., 2021. Water-quality data imputation with a high percentage of missing values: A machine learning approach. Sustain. https://doi.org/10.3390/su13116318

Romañach, S.S., McKelvy, M., Conzelmann, C., Suir, K., 2014. A visualization tool to support decision making in environmental and biological planning. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2014.09.008

Rossaro, B., Boggero, A., Lods-Crozet, B., Free, G., Lencioni, V., Marziali, L., Wolfram, G., 2012. A benthic quality index for European alpine lakes. Fauna Nor. https://doi.org/10.5324/fn.v31i0.1364

Rubin, D.B., 1976. Inference and missing data. Biometrika. https://doi.org/10.1093/biomet/63.3.581

Russo, R., 2021. The Pearson product-moment correlation coefficient r, in: Statistics for the Behavioural Sciences. https://doi.org/10.4324/9780203641576-23

Said, N.M., Zin, Z.M., Ismail, M.N., Bakar, T.A., 2019. Comparative analysis of missing data imputation methods for continuous variables in water consumption data. Int. J. Adv. Trends Comput. Sci. Eng. https://doi.org/10.30534/ijatcse/2019/6981.62019

Sarstedt, M., Mooi, E., 2014. A Concise Guide to Market Research: Cluster analysis. Springer.

Schielzeth, H., Dingemanse, N.J., Nakagawa, S., Westneat, D.F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N.A., Garamszegi, L.Z., Araya-Ajoy, Y.G., 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. Methods Ecol. Evol. https://doi.org/10.1111/2041-210X.13434

Seaman, S., Galati, J., Jackson, D., Carlin, J., 2013. What is meant by “missing at random”? Stat. Sci. https://doi.org/10.1214/13-STS415

Shah, A.D., Bartlett, J.W., Carpenter, J., Nicholas, O., Hemingway, H., 2014. Comparison of random forest and parametric imputation models for imputing missing data using MICE: A CALIBER study. Am. J. Epidemiol. https://doi.org/10.1093/aje/kwt312

Sojka, M., Choiński, A., Ptak, M., Siepak, M., 2020. The variability of lake water chemistry in the bory tucholskie national park (Northern Poland). Water (Switzerland). https://doi.org/10.3390/w12020394

Sojka, M., Jaskuła, J., Wróżyński, R., 2019. ANALYSIS OF HEAVY METALS CONTAMINATION IN BOTTOM SEDIMENTS OF LAKES LOCATED IN THE GNIEZNO LAKELAND. Acta Sci. Pol. Form. Circumiectus. https://doi.org/10.15576/asp.fc/2019.18.4.137

Srebotnjak, T., Carr, G., De Sherbinin, A., Rickwood, C., 2012. A global Water Quality Index and hot-deck imputation of missing data. Ecol. Indic. https://doi.org/10.1016/j.ecolind.2011.04.023

Tibshirani, R., Walther, G., Hastie, T., 2001. Estimating the number of clusters in a data set via the gap statistic. J. R. Stat. Soc. Ser. B Stat. Methodol. https://doi.org/10.1111/1467-9868.00293

Wang, L., Xue, H., 2020. Group decision-making method based on expert classification consensus information integration. Symmetry (Basel). https://doi.org/10.3390/sym12071180

Ward, J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. J. Am. Stat. Assoc. https://doi.org/10.1080/01621459.1963.10500845

Weerasinghe, V.P.A., Handapangoda, K., 2019. Surface water quality analysis of an urban lake; East Beira, Colombo, Sri Lanka. Environ. Nanotechnology, Monit. Manag. https://doi.org/10.1016/j.enmm.2019.100249

Werner, P., Adler, S., Dreßler, M., 2016. Effects of counting variances on water quality assessments: Implications from four benthic diatom samples, each counted by 40 diatomists. J. Appl. Phycol. https://doi.org/10.1007/s10811-015-0760-9

Wu, J., Xiong, H., Chen, J., 2009. Towards understanding hierarchical clustering: A data distribution perspective. Neurocomputing. https://doi.org/10.1016/j.neucom.2008.12.011

Wu, Y., Duguay, C.R., Xu, L., 2021. Assessment of machine learning classifiers for global lake ice cover mapping from MODIS TOA reflectance data. Remote Sens. Environ. https://doi.org/10.1016/j.rse.2020.112206

Xiao, J., Bulut, O., 2020. Evaluating the Performances of Missing Data Handling Methods in Ability Estimation From Sparse Data. Educ. Psychol. Meas. https://doi.org/10.1177/0013164420911136

Yanai, R.D., Mann, T.A., Hong, S.D., Pu, G., Zukswert, J.M., 2021. The current state of uncertainty reporting in ecosystem studies: a systematic evaluation of peer-reviewed literature. Ecosphere. https://doi.org/10.1002/ecs2.3535

Yang, Y., Xiong, Q., Wu, C., Zou, Q., Yu, Y., Yi, H., Gao, M., 2021. A study on water quality prediction by a hybrid CNN-LSTM model with attention mechanism. Environ. Sci. Pollut. Res. https://doi.org/10.1007/s11356-021-14687-8

Yüksel, I., 2012. Developing a Multi-Criteria Decision Making Model for PESTEL Analysis. Int. J. Bus. Manag. https://doi.org/10.5539/ijbm.v7n24p52

Zambelli, P., Lora, C., Spinelli, R., Tattoni, C., Vitti, A., Zatelli, P., Ciolli, M., 2012. A GIS decision support system for regional forest management to assess biomass availability for renewable energy production. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2012.05.016

Zasada, I., Piorr, A., Novo, P., Villanueva, A.J., Valánszki, I., 2017. What do we know about decision support systems for landscape and environmental management? A review and expert survey within EU research projects. Environ. Model. Softw. https://doi.org/10.1016/j.envsoft.2017.09.012

Zhang, S., Xia, Z., Wang, T., Williams, B.K., Szaro, R.C., Shapiro, C.D., Brown, E.D., Wieland, R., Gutzler, C., White, L.G., Welsh, W.D., Wassen, M.J., Runhaar, H., Barendregt, A., Okruszko, T., Walker, J.D., Chapra, S.C., Voinov, A.A., Seppelt, R., Reis, S., Nabel, J.E.M.S., Shokravi, S., Gaddis, E.J.B., Bousquet, F., Costanza, R., Videira, N., Antunes, P., Santos, R., Lopes, R., Vayssières, J., Vigne, M., Alary, V., Lecomte, P., van der Zee, D.-J., Holkenborg, B., Robinson, S., Uusitalo, L., Lehikoinen, A., Helle, I., Myrberg, K., Tversky, A., Kahneman, D., Tsouvalis, J., Waterton, C., Tay, L., Diener, E., Swetnam, R.D., Fisher, B., Mbilinyi, B.P., Munishi, P.K.T., Willcock, S., Ricketts, T., Mwakalila, S., Balmford, A., Burgess, N.D., Marshall, A.R., Lewis, S.L., Surowiecki, J., Sun, A., Stanovich, K.E., Stanovich, P.J., Squires, H., Renn, O., Simon, H.A., Silvertown, J., Shirk, J.L., Ballard, H.L., Wilderman, C.C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B. V., Krasny, M.E., Bonney, R., Sheppard, S.R.J., Cizek, P., Seidl, R., Scholten, L., Scheidegger, A., Reichert, P., Maurer, M., Sawyer, B., Sauvé, L., Renaud, L., Kaufman, D., Sanò, M., Richards, R., Medina, R., Sahin, O., Siems, R.S., Stewart, R.A., Porter, M.G., Rockström, J., Steffen, W., Noone, K., Persson, A., Chapin, F.S., Lambin, E.F., Lenton, T.M., Scheffer, M., Folke, C., Schellnhuber, H.J., Nykvist, B., de Wit, C.A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P.K., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R.W., Fabry, V.J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., Foley, J.A., Röckmann, C., Ulrich, C., Dreyer, M., Bell, E., Borodzicz, E., Haapasaari, P., Hauge, K.H., Howell, D., Mäntyniemi, S., Miller, D.G.D., Tserpes, G., Pastoors, M., Robson, B.J., HAMILTON, D., WEBSTER, I., CHAN, T., Ritzer, G., Ritzema, H., Froebrich, J., Raju, R., Sreenivas, C., Kselik, R., Rinderknecht, S.L., Borsuk, M.E., Schuwirth, N., Langhans, S., Reed, M.S., Raymond, C.M., Bryan, B.A., MacDonald, D.H., Cast, A., Strathearn, S., Grandgirard, A., Kalivas, T., Prestopnik, N.R., Crowston, K., Pettit, C.J., Lewis, H., Petsko, G.A., Papathanasiou, J., Kenward, R., Page, T., Heathwaite, A.L.L., Thompson, L.J., Pope, L., Willows, R., Oxley, T., Jeffrey, P., Lemon, M., Oliver, D.M., Fish, R.D., Winter, M., Hodgson, C.J., Chadwick, D.R., O'Hagan, A., Nyaki, A., Gray, S.A.S., Lepczyk, C.A., Skibins, J.C., Rentsch, D., Nino-Ruiz, M., Bishop, I., Nicolson, C.R., Starfield, A.M., Kofinas, G.P., Kruse, J.A., Nettley, A., Desilvey, C., Anderson, K., Wetherelt, A., Caseldine, C., Nativi, S., Mazzetti, P., Geller, G.N., Nash, U.W., Mustajoki, J., Hämäläinen, R.P., Marttunen, M., Murray-Rust, D., Rieser, V., Robinson, D.T., Miličič, V., Rounsevell, M., Morris, D.E., Oakley, J.E., Crowe, J.A., McKinnon, J., McCall, M.K., Martinez, J., Verplanke, J., Dunn, C.E., Peters-Guarin., G., Matthews, K.B., Rivington, M., Blackstock, K., McCrum, G., Buchan, K., Maslow, A.H., Martin, G., Felten, B., Duru, M., Mackay, C., Lynam, T., Jong, W. de, Sheil, D., Kusumanto, T., Evans, K., Liu, S.B., Poore, B.S., Snell, R.J., Goodman, A., Plant, N.G., Stockdon, H.F., Morgan, K.L.M., Krohn, M.D., Lippe, M., Thai Minh, T., Neef, A., Hilger, T., Hoffmann, V., Lam, N.T., Cadisch, G., Li, Y., Zhu, Z., Leys, A.J., Vanclay, J.K., Latre, M.Á., Lopez-Pellicer, F.J., Nogueras-Iso, J., Béjar, R., Zarazaga-Soria, F.J., Muro-Medrano, P.R., Lange, E., Morgan, E., Romano, D., Lai, J.-S., Chang, W.-Y., Chan, Y.-C., Kang, S.-C., Tan, Y.-C., Labiosa, W.B., Forney, W.M., Esnard, A.-M., Mitsova-Boneva, D., Bernknopf, R., Hearn, P., Hogan, D., Pearlstine, L., Strong, D., Gladwin, H., Swain, E., Kuhn, A., Britz, W., Willy, D.K., van Oel, P., Krueger, T., Hubacek, K., Smith, L., Hiscock, K., Kraker, J. de, Kroeze, C., Kirschner, P., Kragt, M.E., Macleod, C.J.A., Korff, Y. von, Daniell, K.A., Moellenkamp, S., Bots, P.W.G., Bijlsma, R.M., Koltko-Rivera, M.E., Knapp, C.N., Fernandez-Gimenez, M., Kachergis, E., Rudeen, A., Kelly (Letcher), R.A., Jakeman, A.J.A.J., Barreteau, O., Elsawah, S., Hamilton, S.H., Henriksen, H.J., Kuikka, S., Maier, H.R., Rizzoli, A.E.A.E., van Delden, H., Kalaugher, E., Bornman, J.F., Clark, A., Beukes, P., Kahan, D., Jones, N., Ross, H., Perez, P., Leitch, A., Jankowski, P., Nyerges, T., Letcher, R.A., Norton, J.P., Irwin, A., IAP2, Howe, J., Hovmand, P.S., Andersen, D.F., Rouwette, E., Richardson, G.P., Rux, K., Calhoun, A., Hossard, L., Jeuffroy, M.H., Pelzer, E., Pinochet, X., Souchere, V., Højberg, A.L., Troldborg, L., Stisen, S., Christensen, B.B.S., Hoffmann, M., Borenstein, J., Heylighen, F., Chielens, K., Hewitt, R., Escobar, F., Hardcastle, A., Rambaldi, G., Long, B., Lanh, L. Van, Son, D.Q., Hall, D.M., Lazarus, E.D., Swannack, T.M., Gilbertz, S.J., Horton, C.C., Peterson, T.R., Halbrendt, J., Crow, S., Radovich, T., Kimura, A.H., Tamang, B.B., Haklay, M., Groen, E.A., Heijungs, R., Bokkers, E.A.M., de Boer, I.J.M., Greene, J.C., Greenblat, C.S., Chan, A., Clark, D., Cox, L.J., Henly-Shepard, S., Graveline, N., Aunay, B., Fusillier, J.L., Rinaudo, J.D., Glynn, P.D., Giupponi, C., de Vries, B.J.M., Hasselmann, K., Giordano, R., Liersch, S., Gerd Gigerenzer, Brighton, H., Gaillard, J.J.C., Monteil, C., Perrillat-Collomb, A., Chaudhary, S., Chaudhary, M., Chaudhary, O., Giazzi, F., Cadag, J.R.D., Fung, A., Russon Gilman, H., Shkabatur, J., Fulton, E.A., Boschetti, F., Sporcic, M., Jones, T., Little, L.R., Dambacher, J.M., Gray, R., Scott, R., Gorton, R., Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., van der Velde, M., Kraxner, F., Obersteiner, M., Fraternali, P., Castelletti, A., Soncini-Sessa, R., Vaca Ruiz, C., Foster, A., Dunham, I.M., Fisher, R., O'Leary, R.A., Low-Choy, S., Mengersen, K., Caley, M.J., Fischer, F., Fagin, R., Halpern, J.Y., Estelles-Arolas, E., Gonzalez-Ladron-de-

Guevara, F., Enserink, B., Patel, M., Kranz, N., Maestu, J., Guillaume, J.H.A., Filatova, T., Rook, J., Economist, Djaouti, D., Alvarez, J., Jessel, J.-P., Rampnoux, O., Delgado-Galván, X., Izquierdo, J., Benítez, J., Pérez-García, R., Debolini, M., Marraccini, E., Rizzo, D., Galli, M., Bonari, E., Dean, J., Ghemawat, S., d'Aquino, P., Bah, A., Creighton, J.L., Craig, R.K., Ruhl, J.B., Cohn, J.P., Cobb, A.N., Thompson, J.L., Chow, T.E., Sadler, R., Chingombe, W., Pedzisai, E., Manatsa, D., Mukwada, G., Taru, P., Chen, Y., Yu, J., Khan, S., Chen, S.H., Pollino, C.A., Chabris, C.F., Simons, D.J., Catenacci, M., Galelli, S., Ratto, M., Young, P.C., Carmona, G., Varela-Ortega, C., Bromley, J., Campo, P.C., Villanueva, T.R., Butler, M.P., Reed, P.M., Fisher-Vanden, K., Keller, K., Wagener, T., Buss, D., Brooking, C., Hunter, J., Bizikova, L., Burch, S., Robinson, J., Shaw, A., Wolters, H.A., Hoekstra, A.Y., BBC, Bastin, L., Cornford, D., Jones, R., Heuvelink, G.B.M., Pebesma, E., Stasch, C., Williams, M., Le Page, C., Barnaud, C., Page, C. Le, Dumrongrojwatthana, P., Trebuil, G., Aumann, C.A., Audubon, Arnstein, S.R., Arnold, T.R., Ariely, D., Argent, R.M., Arciniegas, G., Janssen, R., Rietveld, P., Anderson, C.A., Lepper, M.R., Ross, L., 2013. Serious games: Improving public policy through game-based learning and simulation. Environ. Model. Softw.

Zhang, Y., Thorburn, P.J., 2022. Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. Futur. Gener. Comput. Syst. https://doi.org/10.1016/j.future.2021.09.033

Zhang, Y., Thorburn, P.J., 2021. A dual-head attention model for time series data imputation. Comput. Electron. Agric. https://doi.org/10.1016/j.compag.2021.106377

Zhang, Y.F., Thorburn, P.J., Vilas, M.P., Fitch, P., 2019. Machine learning approaches to improve and predict water quality data, in: 23rd International Congress on Modelling and Simulation - Supporting Evidence-Based Decision Making: The Role of Modelling and Simulation, MODSIM 2019. https://doi.org/10.36334/modsim.2019.d5.zhangyif

Zhang, Z., 2016. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann. Transl. Med. https://doi.org/10.3978/j.issn.2305-5839.2015.12.63

Zhou, X.H., 2020. Challenges and strategies in analysis of missing data. Biostat. Epidemiol. https://doi.org/10.1080/24709360.2018.1469810