## **Combining indicators for better decisions – Algorithms vs experts on lakes ecological status assessment**

Chrobak G, Kowalczyk T, Fischer TB, Chrobak K, Szewrański S, et al., Ecological Indicators, 2021

**1. Introduction**

**2.1. Measuring ecological state**

There is increasing attention given to the assessment of the ecological conditions of water-based ecosystems. This is closely connected with efforts to develop a better understanding of human impacts on surrounding environments. A principle in current water management practices is the optimization of economic and social well-being of water users without undermining the health of vital ecosystems (Forslund & House, 2009; UN-Water, 2017).There are numerous methodologies that allow for an assessment of the ecological status of aquatic ecosystems (Søndergaard, Jeppesen, Jensen, & Amsinck, 2005; Penning et al., 2008; Bennion et al., 2014). These are normally based on the assessment of a set of component parameters to determine a consistent indicator of ecological fitness (Alahuhta et al., 2009; Carvalho, Cortes, & Bordalo, 2011). In this context, one of the challenges is the need to systematize the methodologies for analysis and assessment of conditions within reproducible models, functioning as frameworks for conducting informative prognostic research (Hampton, Anderson, Bagby, & Gries, 2014; Society, 2017; Powers & Hampton, 2019; Schneider et al., 2019). This structuring creates a possibility for conducting analyses linking the ecological status of ecosystems with climate change models, for example, through statistical downscaling (Flint & Flint, 2012; Keil & Jetz, 2014; Wang et al., 2017). Moreover, in order to achieve robust results, model ensembling allows for optimal use of the latest models in the field of applied ecosystems ecology (Jones-Farrand et al., 2011; Bode et al., 2017).

Currently, within the European Union, systematics originating from the Water Framework Directive 2000/60/EC (WFD) operate in the field of aquatic ecosystems (Grizzetti, Lanzanova, Liquete, Reynaud, & Cardoso, 2016). Within the Directive, the concept of Ecological Status revolves around the quality of aquatic flora and fish fauna, nutrients availability, and factors such as salinity, temperature, and chemical pollutant emissions. Morphological features such as size, streamflow, water levels, and river bed configurations are also taken into account (Acreman & Ferguson, 2010). The WFD's ecological status classification scheme for surface water comprises five quality categories: high, good, moderate, poor, and bad. These categories are assigned to water bodies in a complex evaluation process (Xu, Tao, Dawson, Li, & Cao, 2001; Hambright & Berman, 2014).

Among freshwater ecosystems, lakes are considered composite and dynamic structures, each situated in a unique landscape context (Aoki, 2012). One of the main reasons of the complexity of their assessment is that, even though each aquatic environment is unique, there are only a few that are studied in great detail ecologically (Van Der Lee, Van Der Molen, Van Den Boogaard, & Van Der Klis, 2006). In case of most lakes, for which no comprehensive studies are available, there in a need for theoretical knowledge or forecasts about such factors such as as fisheries, tourism, and water quality (Kaklauskas, 2014; Loucks, van Beek, Loucks, & van Beek, 2017). In these situations, ecosystem modeling is the key method from which predictions and analyses can be obtained (Pasetto et al., 2018; Castellanos, Huntley, Voelker, & Lawing, 2019).

**1.2. Modelling aspects**

The ecological status modelling of lakes and other aquatic environments is relevant not only in practical matters relating to the water management or environmental solutions, but also in demonstrating how multiple factors simultaneously affect ecosystem features, such as population composition, biological productivity, and overall ecological fitness (Piroddi et al., 2015; Moe, Haande, & Couture, 2016). Modelling is, therefore, useful not only to provide practical knowledge for problem solving, but also to enhance the basic understanding of lake ecosystems considered as a group of individual objects, yet with underlying common features. Generally speaking, ecological state models of lake ecosystems are composed of predictive equations that integrate multiple variables of the environment (Schuwirth et al., 2019). Environmental variables may be physical, chemical or biological, and are expressed as quantities per unit or volume per unit area, or as flows (Mantzouki et al., 2018). The relevant indicators for lake ecosystem assessment do not only include those specific to the lake itself but also to the basin from which the lake receives water and substances, dissolved or suspended, that influence ecosystem processes (Xu et al., 2001; Martín-López, Gómez-Baggethun, García-Llorente, & Montes, 2014). Technically, ecological ecosystem models of high complexity can be developed (Hooper et al., 2005; Landi, Minoarivelo, Brännström, Hui, & Dieckmann, 2018). Nonetheless, previous experiments found that complexity is weakening the predictive reliability of ecosystem models, making them not resistant to, for example, an increase of the size of the analyzed set of objects (Parrott, 2002), or existence of outliers (Benhadi-Marín, 2018).

Various models for lake ecosystem condition assessments coexist, featuring diverse research techniques worldwide. Some of them are based directly on the implementation of the requirements from the WFD. Chemical and biological data - alkalinity and water depth (Xu et al., 2005; Zhang, Liu, Yang, Li, & Yang, 2013), chlorophyll *a* and cyanobacteria (Søndergaard, Larsen, Jørgensen, & Jeppesen, 2011), phytoplankton assemblages (Padisák, Borics, Grigorszky, & Soróczki-Pintér, 2006), macrophytes and phytobenthos (Schaumburg et al., 2004) have been used to assign ecosystems to one of five classes. Fish-based assessment methods are also still popular; however, their use is often limited by the availability of data on fish communities (Rask, Olin, & Ruuhijärvi, 2010; F. L. Kelly, Harrison, Allen, Connor, & Rosell, 2012). As a result of numerous datasets being used and different levels of data access throughout the world – many diverse approaches have emerged, some with converging views of ecological status – striving for proposing a ‘guiding image’ for classification problems (Birk, Willby, Kelly, Bonne, Borja, Poikane, & van de Bund, 2013; Mayer, Winkler, & Fry, 2014).

One of the main goals for ecological status assessment is the intercalibration of classification models for the sake of common lake management objectives (Birk, Willby, Kelly, Bonne, Borja, Poikane, & Bund, 2013; Poikane et al., 2014). However, according to some researchers, obstacles towards achieving consistent classification methods can lie in data quality, the uncertainty indicators, temporal variations (year-to-year variability), and data redundancy in the ecological assessment of water ecosystems (Baho, Futter, Johnson, & Angeler, 2015; M. G. Kelly et al., 2016; Morelli & Tryjanowski, 2016). These shortcomings negatively affect the achievement of the necessary compromise between science and management (Navarro et al., 2009). Some of the problems described above are addressed by the rapidly growing field of machine learning and artificial intelligence (Willcock et al., 2018; Desjardins-Proulx, Poisot, & Gravel, 2019), as is subsequently explained.

**1.3. Intelligent information retrieval and reproducibility**

Machine Learning is used in modeling the ecological status of lakes, especially in cases of individual examples with self-context characteristics (Peterson et al., 2003; Weyl, Ribbink, & Tweddlel, 2010). Regional scale assessments are mainly produced with the use of remote sensing techniques (Dörnhöfer & Oppelt, 2016; Harwood et al., 2016). However, in the case of classification problems regarding ecological parameters, methods derived directly from machine learning are used, such as decision trees (Everaert, Boets, Lock, Džeroski, & Goethals, 2011; Yajima & Derot, 2018), artificial neural networks (Chou, Ho, & Hoang, 2018), or genetic programming (Hamblin, 2013). Thanks to the use of dimensionality reduction methods, such as principal component analysis or linear discriminant analysis (Muttil & Chau, 2007), the problem of data redundancy can be addressed (Morrissey, 2014). Besides, many previously effective expert techniques for assessing ecological status can be processed using machine learning models to function in a systematic environment as reproducible decision support tools (Christin, Hervet, & Lecomte, 2019).

Creating reproducible classifiers supports the universality of methodologies for assessing the ecological status of ecosystems, and also allows for an assessment of a given object in the context of its appropriate group. The solutions proposed based on classification algorithms are bringing researchers closer to solving the problem of assessing the internal dynamics of objects in their respective subgroups without the need to extend (or change the granulation of) the adopted structure of separate classes. Thus, it constitutes a step forward towards a new approach, focusing on object membership level within classes instead of binary classification outcome. The consequences of such approach are currently widely discussed by researchers dealing with clustering algorithms in applied sciences (Chen et al., 2015; Lauer & Bloch, 2019; Qu, Bengio, & Tang, 2019; Chrobak, Chrobak, & Kazak, 2020).

**1.4. Research goals and organization**

The main goal of rrsearch study underlying this paper was to obtain latent information from the multidimensional assessment of the ecological state of X lakes in Poland in order to facilitate the insightful visualization and geometric interpretation of the evaluation results for the user (policy maker, planner, ecologist). At the same time, an aim was to minimize the loss of explanatory ability of the original set of measurements. The reconstruction of the assumptions of the source methodology in the process of supervised learning allowed for the creation of a model of clusters for individual ecological status classes in a two-dimensional space. An additional goal was to draw attention to the visually extended information about the position of a given object (e.g. xxx) within its proper class, instead of knowing only about the fact of belonging. This created a premise for the object-oriented evaluation of changes in the ecological status not only between but also within classes, and (using the obtained model) between measurement campaigns.

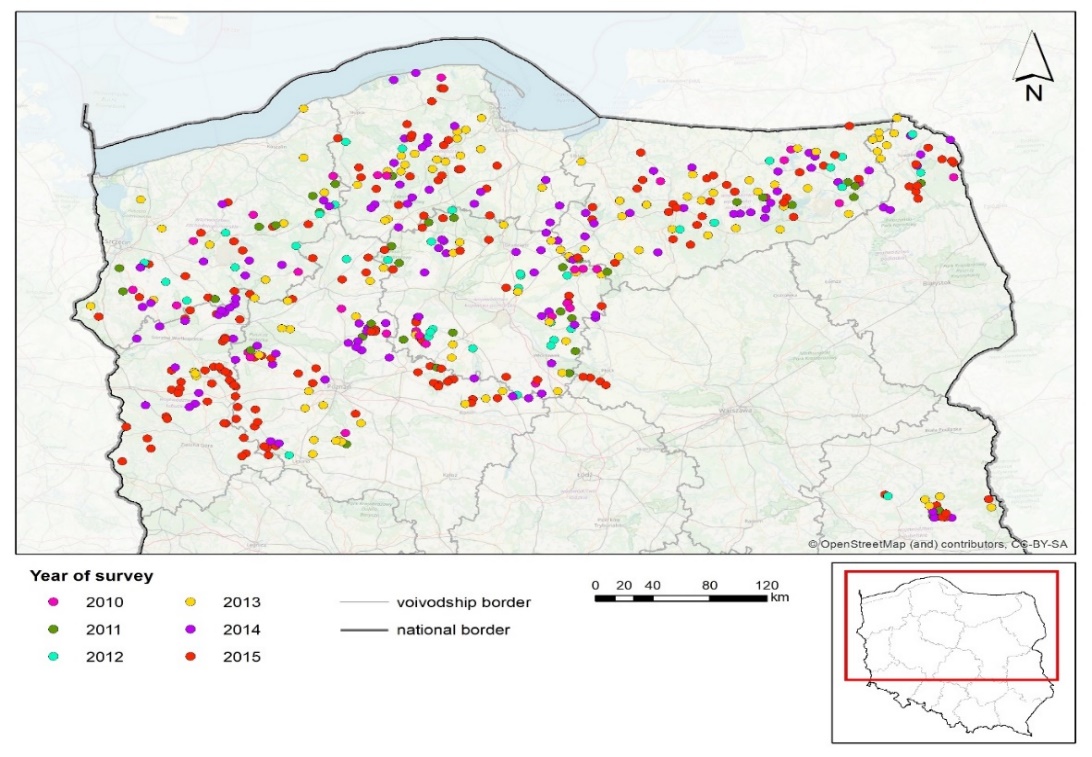
In the research underlying this paper, an ecological status classifier was developed for lake ecosystems, based on the results of the work of experts following the WFD guidelines. The available measurement data from four campaigns were used for the calculations. An analysis of missing observations was made, and a method of supplementing them was proposed. Then the dimensionality reduction was performed on the set of variables. The resulting classifier was based on the kernel Support Vector Machine method.

This paper is organized as follows. In the second section – Materials and Methods – the basic description of acquired measurements is presented together with the outline of the general analysis scope, followed by a description of particular pre-processing and machine learning-based solutions used in the research. The developed classification is then presented (third section – Results), accompanied by results of previous analyses results and detailed model schema. Next, in the Discussion, the results are summarized, featuring a description of encountered issues and applicability of research outcomes is addressed. Moreover, in this section the problem which stems from the analysis of conducted research outcomes is formulated. Finally, we point to the practical nature of the obtained results and provide information on further research.

**2. Materials and Methods**

**2.1. Data**

Input data for analysis were obtained from the repository of the Chief Inspectorate of Environmental Protection in Poland (Appendix A) (GIOŚ, 2015), the same which is used for reporting to the European Comission databases concerning environmental monitoring. Measurements are carried out following the requirements set out in national law and the European Union, due to the implementation of the aquis communautaire obligations in the field of monitoring and assessment of surface water status (European Environmental Agency, 2018). The obtained dataset covers the ecological status of 499 lake ecosystems in Poland, measured in the period 2010-2015 (Figure 1). Information on ecological status includes the following parameters: Chlorophyll *a*, Nitrogen, Phosphorus, Phytoplankton, Ecological State Macrophyte Index (ESMI), Diatom Index for Lakes (IOJ), Phytoplankton Method for Polish Lakes (PMPL), Visibility, and Conductivity.



*Figure 1. Lake objects assessed ecologically in Poland during 2011-2015 campaign.*

**2.2. Analysis framework**

The general flow chart of the analyzes is presented in Figure 2. As part of the preparation of the data set for modeling, data implementation was carried out to supplement the missing observations. Next, scaled data were subjected to dimensional reduction, and then divided into two subsets: training and testing. The training subset (80%) was used in the k-fold Cross Validation process, within which a classifier was created, assisted with the grid search method to determine the optimal hyperparameter for Support Vector Machine model. Classifiers were used to create results predictions within k test folds. Then, after determining the confusion matrix and the average accuracy model, the classification model was used on the test set (20%) and classification results were obtained. A detailed diagram of the actual modeling process is presented in the sub-section Classification Modeling with Hyperparameter Optimization of the Results section.

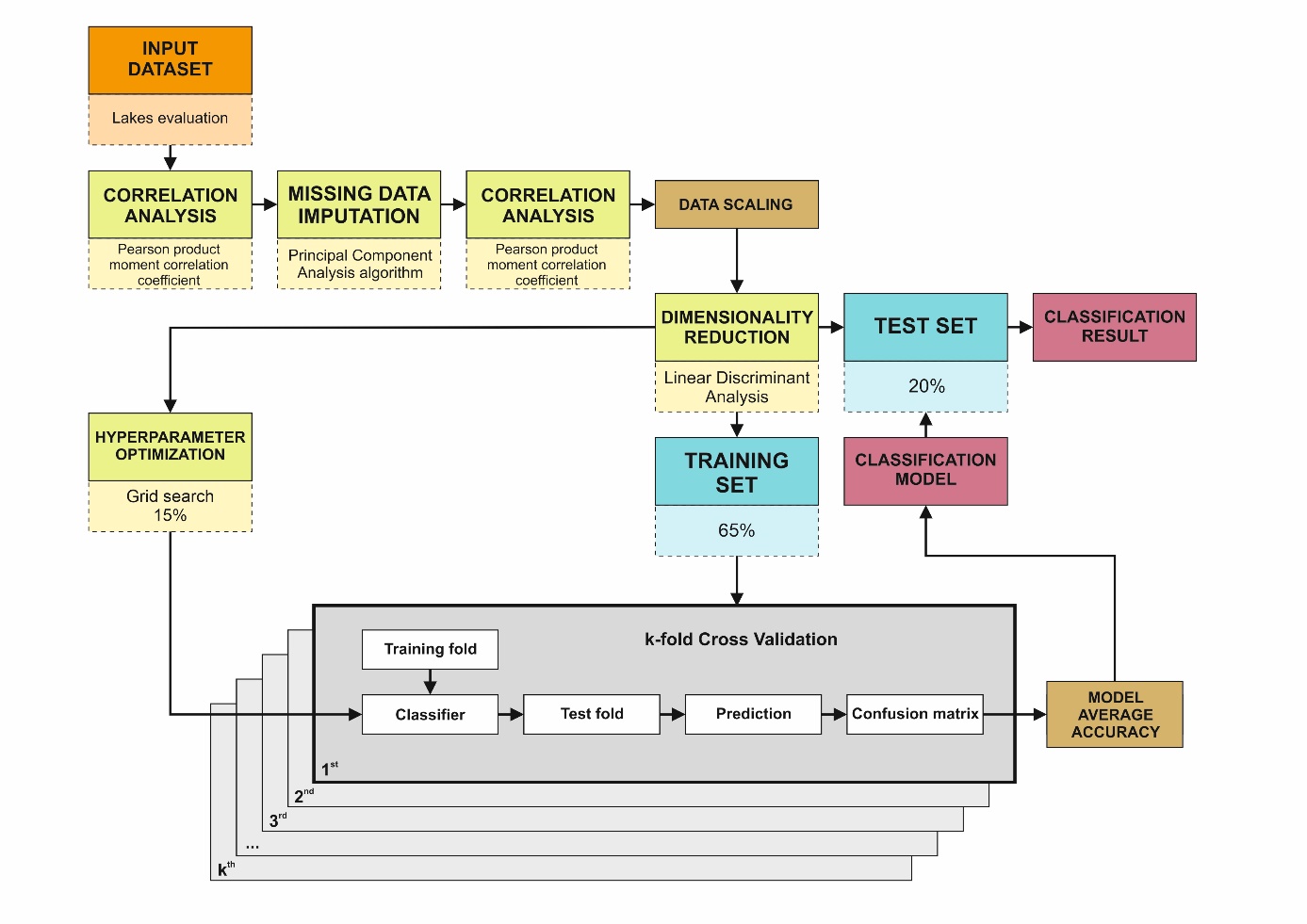


Figure 2. General schema of data pre-processing and modelling stages. Source: own.

**2.3. Correlation analysis**

Correlation analysis was performed to detect associations between individual measured parameters. Pearson product-moment Correlation coefficient (PPMCC) was used (Gooch, 2011). The linear dependence between each variables pair was measured as follows:

where:

**x** and **y** are the two vectors of length **n,**

**mx** and **my** correspond to the means of **x** and **y**, respectively;

The p-value (significance level) of the correlations was determined by calculating the t-value as follows:

The corresponding p-value was determined using t distribution table for n-2 degrees of freedom, confirming the significance of the correlation between a pair of variables if the p-value is less than 5%.

**2.4. Missing data imputation**

In the case of missing observations identified – the imputation of incomplete continuous datasets with the use of regularized iterative Principal Component Analysis (PCA) was performed (Grung & Manne, 1998). In the initial step of the algorithm (α = 0) – the missing observations are imputed with preparatory value, such as the mean of the variables. The matrix with imputted values is denoted as X0. Next, the M0 matrix is computed, which consist of vectors containing the mean of the X0 variables, repeated in each row of M0. In the second step, the PCA is performed on the completed dataset in the form of Singular Value Decomposition of (Xα-1 - Mα-1) in order to estimate Uα, Vα, and (ꓥα)1/2 parameters. While keeping the first F dimensions, the fitted matrix is built as follows:

where the variance of the noise is estimated as:

hence, the obtained imputed dataset is defined as:

where 1 is an all-ones matrix of size n x p.

As a result, the original values are preserved, and the missing data is replaced with the regularized fitted values. Mα is then updated from the newly completed matrix. The above steps are repeated as far as the change in the imputed matrix goes below a threshold:

with equal to, for example, 10-7.

The output of the above algorithm was used as input for the Linear Discriminant Analysis.

**2.5. Dimensionality reduction**

In a pre-processing step for classification modeling, a dimensionality reduction linear transformation technique was used in order to project an n-dimensional dataset of variables onto a lower-dimensional subspace k, where k ≤ n - 1. This method emphasises the class-separability of observations, which leads to avoidance of overfitting and reduction of computational effort (Cunningham & Ghahramani, 2015). LDA is a so-called supervised algorithm that computes the directions – linear discriminants – which represent the axes maximizing the separation between multiple observations (Gu, Li, & Han, 2011). The goal of LDA is achieved by searching for a linear combination of predictors (observations) that best separates the target classes. In the first step of LDA, the mean vectors vi are computed for m different classes, for each predictor where i = 1,2,3,…,n:

Afterwards, the two p x p dimensional matrices are computed - where p is the number of predictors – inter-class and inner-class matrixes. The inner-class matrix Sw is computed as:

for every class, where vi – the mean vector is computed as:

The inter-class matrix SB is computed as:

where M is the overall mean, vi is the sample mean vector, and Vi is sample size - each for the respective classes.

Next, the generalized eigenvalue problem is solved for the matrix in order to obtain the linear discriminants (Moler & Stewart, 1973). For v being an eigenvector of a Ω matrix :

Ωv = λv

where λ is the eigenvalue, and v is an eigenvector which has the same eigenvalue, because:

Ω(-v) = -vΩ = - λv = λ(-v)

hence eigenvalue calculation satisfies the equation:

The obtained eigenvectors indicate the directions of the distortion, while eigenvalues being scaling factors for the eigenvector describing the magnitude of distortion. The resulting eigenvectors form axes of a new subspace. The associated eigenvalues describe the informative clarity of new axes. However, the resulting eigenvectors describe only the new axis directions, since their unit length is equal to 1. In order to choose which eigenvectors would define the new lower-dimensional subspace, the corresponding eigenvalues are examined. The eigenvectors are ranked in descending order by eigenvalues, and the top k eigenvectors are selected. The eigenvectors with the lowest eigenvalues are considered less-informative about the data distribution. It should also be noted that the number of linear discriminants is at most c – 1, where c is the number of classes since the inter-class matrix SB is the sum of c matrices with rank ≤ 1. Based on the number of k informative eigenvectors, the n x k-dimensional matrix W is constructed. In the last step – the W matrix is used to transform the original observations onto the new subspace:

where X is a n x d-dimensional matrix representing the original observations, and Y is the transformed n x k-dimensional subspace.

**2.6. Support Vector Machine modelling**

For the task of developing a model that performed distinct classification of data within a lower-dimensional subspace defined previously, the kernel Radial Basis Function SVM (RBF kernel SVM) algorithm was used. Kernel methods are popular in machine learning (Hofmann, Schölkopf, & Smola, 2008; Gönen & Alpaydin, 2011). They are a robust method that can be used to expand the SVM algorithm to define non-linear decision boundaries. The key concept behind these methods is focused on the so-called kernels or kernel functions, which, under certain theoretical conditions of symmetry and positive definiteness, indirectly represent the inner product in high-dimensional space (Christensen, 2010). Replacing the original internal product in the input space with a positively defined kernel automatically stretches the algorithm to a non-linear separation in the input space (Figure 3).

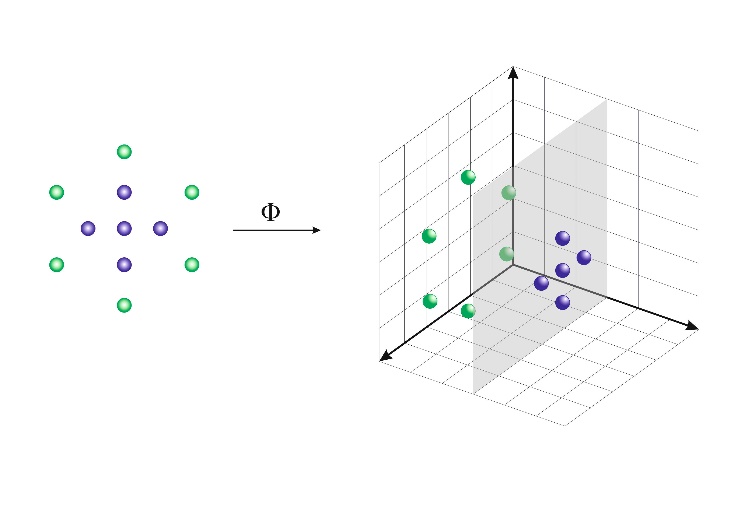


Figure 3. Generalized illustrartion of kernel function operation to repersent the inner product in higher dimensional space for separation purpose.

A function is called a kernel over . The main idea is to define the kernel K in such a way that, for any two points: is equal to the inner product of the vectors: and :

,

for some mapping to a *Hilbert* space called a function space (Cooper & Berberian, 1977). Since the inner product is a measure of the similarity between two vectors, is often viewed as a measure of similarity between the elements of the input space.

During the implementation of the SVM algorithm, the classification procedure is performed by the construction of hyperplanes (decision surfaces) in a multidimensional space that best separates cases of different class labels (Crisp & Burges, 2000). In order to obtain a desired (optimal) hyperplane, the method employs an iterative training algorithm, which is then used to minimize an error function. For a constant , a *Radial Basis Function* is the kernel defined over with:

,

with (*sigma*) acting as a tuning parameter denoting the smoothness of the decision boundary and controlling the model variance (Friedrichs & Igel, 2005). In terms of intuitive understanding – the sigma parameter defines the extent to which the influence of a single example of training reaches, with low values meaning ‘far’ and high values – ‘near’. The sigma parameter can be viewed as the opposite of the impact range of the samples chosen by the algorithm as support vectors.

**2.7. Hyperparameter optimization**

In the process of machine learning, the problem of choosing a set of optimal hyperparameters is refered to as optimization or tuning (Feurer & Hutter, 2019). A hyperparameter is a parameter whose value is used during the performance control of the learning process – it is external to the model and cannot be estimated from the dataset (Hutter, Hoos, & Leyton-Brown, 2014). One way to perform a hyperparameter optimization is via grid search (parameter sweep) process, which is a form of exhaustive searching through an indicated hyperparameter space of - in this particular case – SVM radial algorithm. For this work, the C and sigma parameters were optimized. The C parameter balances the right assignment of training instances against optimizing the margin of the decision function. With higher values of C, a narrower margin will be agreed if the decision function is more comfortable to identify all training points correctly. The lower C would promote a larger margin, therefore a more evident decision-making feature, at the cost of training precision (Hastie, Rosset, Tibshirani, & Zhu, 2004). In order to find the optimal hyperparameters, the R package ‘caret’ was used (Kuhn, 2008). During the process, an SVM modeling function was nested within the iterative grid search algorithm, which allowed for full cross over all parameter options, so that each hyperparameter value is tested at all other values of all parameters resulting in obtaining of optimal hyperparameter values. This step, presented in detail in section 3.3., allows for the selection of the most optimal hyperparameters from among the solutions available in the SVM function repository. The parameters selected in this way showed the highest efficiency in the next stage (k-Fold Cross-Validation), ensuring the maximization of the potential of the kernel-SVM model for the analyzed dataset.

**2.8. k-Fold Cross-Validation**

In order to estimate the skill of the SVM model, the cross-validation method was implemented. The k-Fold Cross Validation is commonly applied in machine learning to compare and select a model for a given problem while being relatively less-biased than other popular methods (Park & Kim, 2012). In this analysis, the k-Fold CV resampling procedure was used to estimate the skill of the model on the test dataset. The k parameter of k-Fold CV indicates the number of groups that a dataset is to be split into. During the procedure:

1. The dataset was shuffled randomly.
2. The dataset was split into k=10 groups.
3. For each of 10 groups:
   1. The one group is selected as a test set.
   2. The SVM linear model was fitted to the remaining nine groups – creating a training set.
   3. The SVM linear model was evaluated on the test set.
   4. The evaluation score was retained.
4. The skill of the SVM model was summarized using the sample of model evaluation scores.

The k parameter must be carefully chosen – a poorly selected k-value may result in a misleading notion of the model skill (Rodríguez, Pérez, & Lozano, 2010). In this particular case, the parameter of k is fixed to 10 – a value that has been found through previous research to generally result in a model skill estimate with relatively low bias (Brownlee, 2019). It is essential to acknowledge that each observation from the dataset was assigned to an individual group and remained in its group for the duration of the procedure. Therefore, each sample was allowed to be used in the test set 1 time and act as a training sample 1 - k times. The results of the *k-Fold CV* were finally summarized with the mean of the *SVM* model skill scores.

**3. Results**

**3.1. 2011 – 2015 campaign results**

The results indicate that a prevalent number of lakes assessed are in a moderate ecological condition with ~ 30% of the facilities with good condition represented, ~ 26% of the lakes with poor and bad conditions - ~ 12% each. High ecological status was detected for 50 ecosystems, i.e., ~ 10% of the studied sample (Figure 4). The monitoring results have been accompanied by information on classification uncertainty, which resulted from gaps in measurements for some objects. High classification uncertainty was found for 100 lakes, most of which were classified with good or high status (Figure 5). During the campaign, biological, physico-chemical and hydromorphological parameters and indexes were measured (examples in Figure 6) – for complete set of measurements please refer to Appendix B). Analysis of the completeness of the measurements showed deficiencies in five out of eight measured variables. The greatest deficiencies in the measurements were found for the Diatom Index for Lakes - 15%, in the case of the ESMI indicator 11% were missing, while for PMPL, Chlorophyll "a", and Visibility - it was less than 5% of missing measurements (Figure 7). In total, data deficiencies during the campaign amounted to 164 values, which accounted for ~ 5% of all measurements.



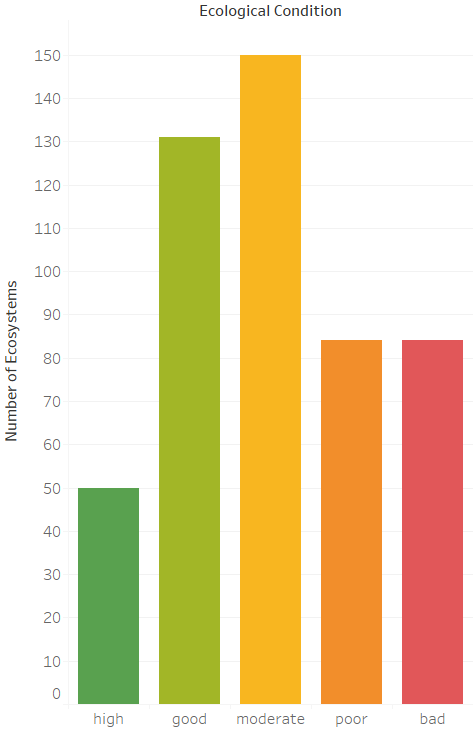
. 

Figure 4. Ecological status assessment results.

Figure 5. Uncertainty assessment results.

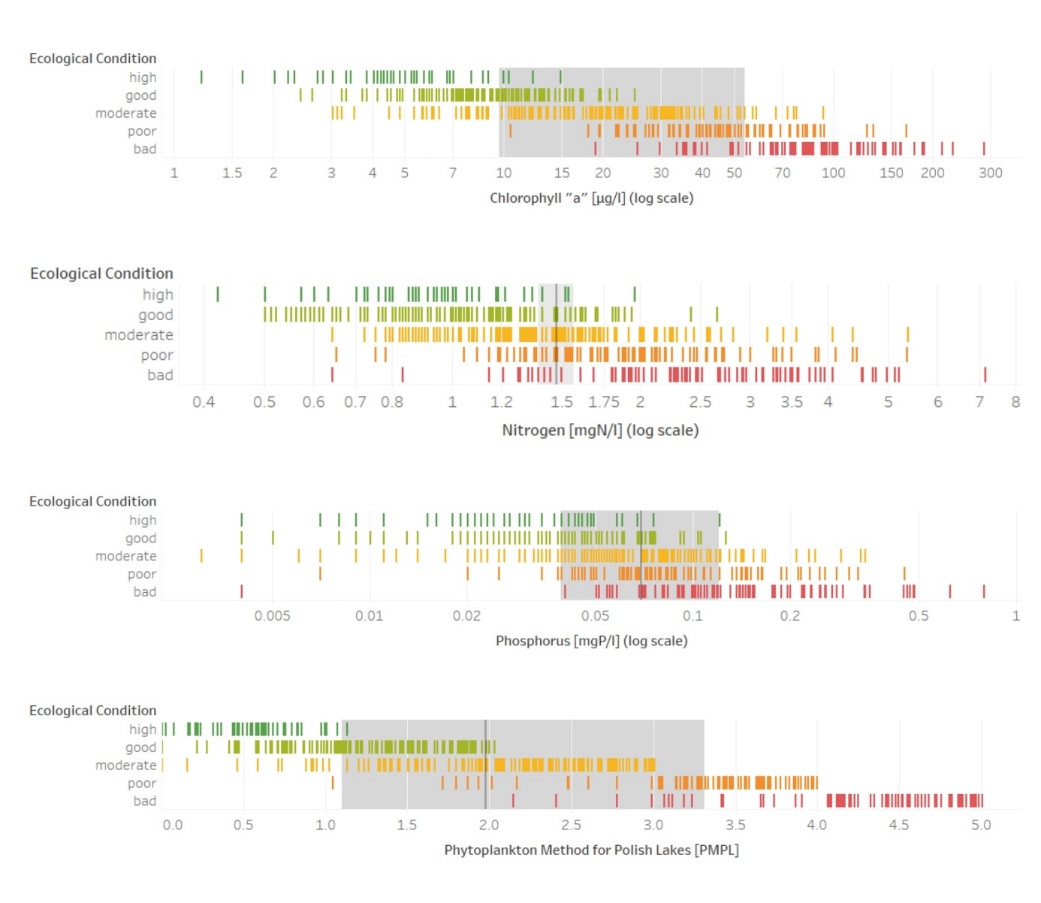


Figure 6. Parameters across ecological classes for Chlorophyll a, NItrogen, Phosphorus, and PMPL.

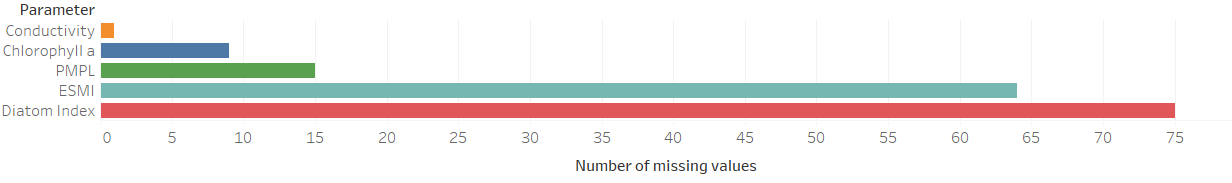


Figure 7. Missing values count per parameter.

**3.2.** **Missing Data Imputation and Correlation Analysis**

The identified missing observations were supplemented using the PCA algorithm. The number of dimensions for the analysis was estimated by cross-validation with the use of *estim\_ncpPCA* function from the *missMDA* package for R (Josse & Husson, 2016). The following parameters were chosen for the function arguments:

|  |  |  |
| --- | --- | --- |
| **Argument** | **Value** | **Description** |
| X | *Lakes* dataset | A data frame with missing values |
| ncp.min | 0 | Minimum number of components to test |
| ncp.max | 5 | Maximum number of components test |
| method | „Regularized” | Iterative PCA algorithm |
| method.cv | „gcv” | Generalized cross-validation |

Table 1. Parameters used for estim\_ncpPCA performance.

The number of components retained for the *PCA* was chosen based on *Mean Square Prediction Error* (*MSPE*) value, computed for each number of components, yelding results (Allen, 1971), as is shown in Table 2.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Components** | 0 | 1 | 2 | 3 | 4 | 5 |
| **MSPE** | 8808.373 | 9651.396 | 10439.058 | **3169.197** | 3524.744 | 5716.199 |

Table 2. MSPE values for particular componetns number.

The smallest *MSPE* value was obtained for a combination of 3 components – thus, for the following imputation algorithm, *ncp* argument was set to 3. The missing values were then estimated with the use of the *imputePCA* algorithm of the *missMDA* package (Bertsimas, Pawlowski, & Zhuo, 2018). The *PCA* was performed for the imputed dataset, and the resulting 2-dimensional vector map was compared to one created before the imputation process. The imputation procedure improved the combined explanatory skill of all component vectors. Within two leading dimensions, 68.35% of dataset variability was explained compared to 65.62% before imputation (Figure 8).

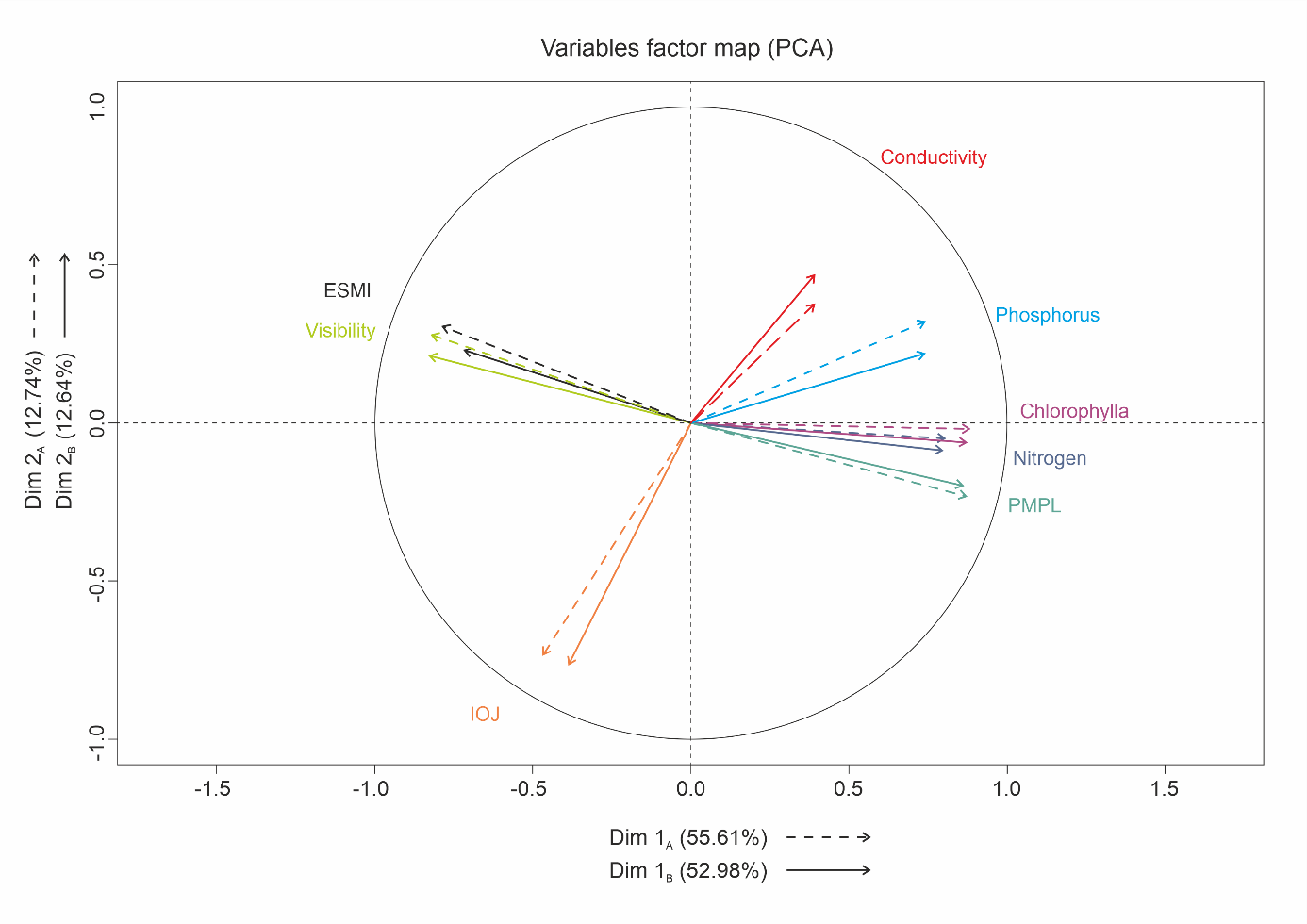


Figure 8. Factor map for variables set showing the explanatory skill prior to and after the imputation process.

Next, a comparison was made between the correlation results for the data before and after the data completion algorithm was applied (Figure 9). The results of PPMCC showed the existing correlation (negative and positive) between some of the measured variables. A strong or moderate downhill linear relationship was detected between PMPL - Visibility variables (-0.78), PMPL - ESMI (-0.69), and Chlorophyll „a” – Visibility (-0.65). Strong and moderate uphill relationships were detected for pairs PMPL – Chlorophyll „a” (0.78), Nitrogen – Chlorophyll „a” (0.72), Chlorophyll „a” – Phosphorus (0.67), and ESMI – Visibility (0.64). The results of the correlation analysis constituted a premise for the application of the dimensionality reduction technique (Mladenić, 2006; Khalid, Khalil, & Nasreen, 2014).

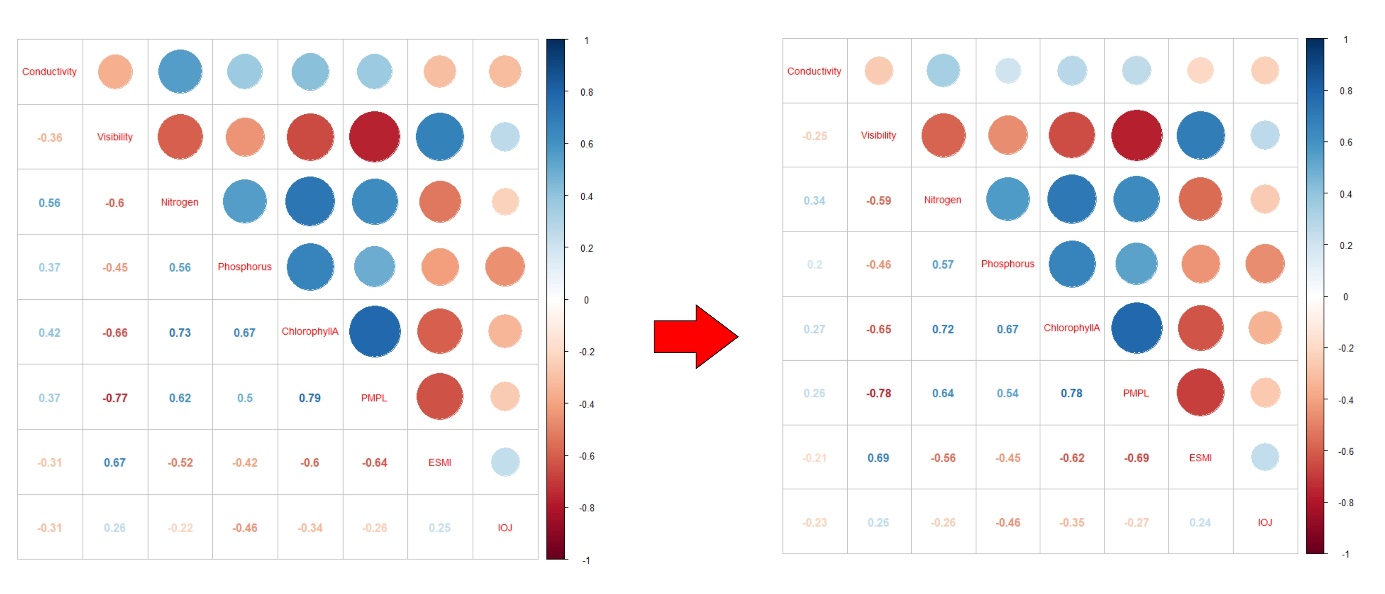


Figure 9. PPMCC results for dataset prior to and after the data imputation.

**3.3.** **Dimensionality reduction**

The training dataset was scaled before the dimensionality reduction (Nguyen & Holmes, 2019). Linear Discriminant Analysis (LDA) was used as a dimensionality reduction algorithm, due to its ability to emphasize the differences between groups given the grouping factor (Tharwat, Gaber, Ibrahim, & Hassanien, 2017). The LDA was performed with the use of the *MASS* package for R (Ye, 2007). The prior probabilities of class membership for observations were assumed from class counts in the original dataset (Table 3).

|  |  |  |
| --- | --- | --- |
| **class** | **prior probability** | **class count** |
| bad | 0.1683367 | 84 |
| poor | 0.1683367 | 84 |
| moderate | 0.3006012 | 150 |
| good | 0.2625251 | 131 |
| high | 0.1002004 | 50 |

Table 3. Prior probabilities for class membership derived from class counts.

The averages for each observed data group were determined in order to recognize a typical middle-value before LDA performance (Table 4).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **class** | **Conductivity** | **ESMI** | **Diatom Index** | **Nitrogen** | **Phosphorus** | **PMPL** | **Visibility** | **Chlorophyll „a”** |
| bad | 0.356 | -1.051 | -0.346 | 1.128 | 1.055 | 1.460 | -0.984 | 1.496 |
| poor | 0.328 | -0.654 | -0.240 | 0.438 | 0.293 | 0.787 | -0.696 | 0.440 |
| moderate | -0.051 | -0.063 | -0.091 | -0.122 | -0.126 | -0.127 | -0.139 | -0.270 |
| good | -0.247 | 0.728 | 0.244 | -0.607 | -0.509 | -0.782 | 0.685 | -0.641 |
| high | -0.350 | 1.146 | 0.616 | -0.674 | -0.555 | -1.343 | 1.446 | -0.764 |
|  |  |  |  |  |  |  |  |  |

Table 4. Middle values estimated for each variable group.

A transforming matrix was then developed by the algorithm (Table 5), which was used to transform observations to discriminant functions, normalized. Therefore covariance matrix is isotropic (diagonal with all elements on the diagonal equal) within groups of observations.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **1st LD** | **2nd LD** | **3rd LD** | **4th LD** |
| **Conductivity** | 0.02843 | -0.02241 | 0.35479 | 0.49875 |
| **ESMI** | -0.47358 | 0.28885 | 0.63065 | -0.77917 |
| **Diatom Index** | -0.01742 | 0.42344 | -0.06262 | 0.32642 |
| **Nitrogen** | 0.02070 | 0.09696 | -0.46705 | 0.56881 |
| **Phosphorus** | 0.10092 | 0.17338 | -0.52082 | 0.10191 |
| **PMPL** | 1.89904 | -0.01991 | 1.40100 | 0.40957 |
| **Visibility** | -0.06827 | 1.09310 | 0.37995 | 0.91133 |
| **Chlorophyll „a”** | 0.06319 | 1.35493 | -0.08710 | -0.93963 |

Table 5. Transformation matrix used to obtain normalized discriminant fuction values from observations.

The singular values (svd) were then obtained, which denoted the ratio of the *between* and *within*-group standard deviations on the linear discriminant variables. Then, svds were squared to obtain the canonical F-statistics, and the amount of between-group variance explained by each linear discriminant was computed (Table 6).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **svd** | **F-statistics** |  | **% variance explained** |
| **LD1** | 24.855200 | 617.780979 | 0.932079 | ~ 93.2% |
| **LD2** | 6.461864 | 41.755686 | 0.062999 | ~ 6.3% |
| **LD3** | 1.391552 | 1.936418 | 0.002921 | ~ 0.3% |
| **LD4** | 1.151296 | 1.325483 | 0.001999 | ~ 0.2% |

Table 6. Variance explanatory skill based on svd measures for each of Linear Discriminants.

The self-explanatory skill of the first dimension alone, which determined over 93% of dataset variance is presented in Figure 10. However, this ability does not seem to explain well the interactions between ‘good’ and ‘poor’ class discriminants and their neighbors. The visible discriminant value spreads for ‘good’ and ‘poor’ are 4.877 and 2.8 respectively. The results of supervised algorithm performance showed that the first two dimensions *LD1* and *LD2* of resultant subspace are capable of explaining most of the variance (over 99%) within the analysed dataset, except few situations visible in the graph, where prediction outcomes intercept between classes (Figure 11). The LDA result was then fed to a classification algorithm in order to create an optimized machine learning model focused on further elimination of inter-class overlaps.

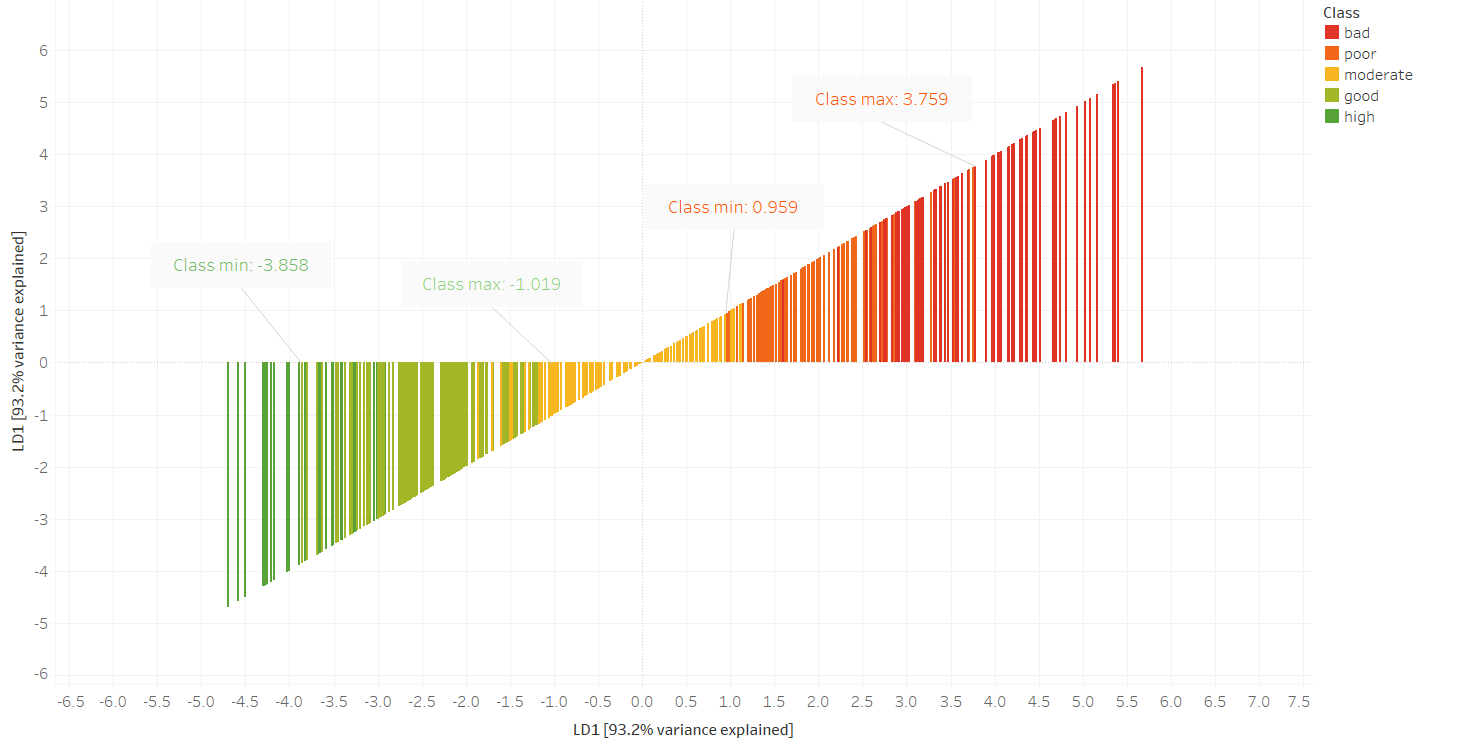
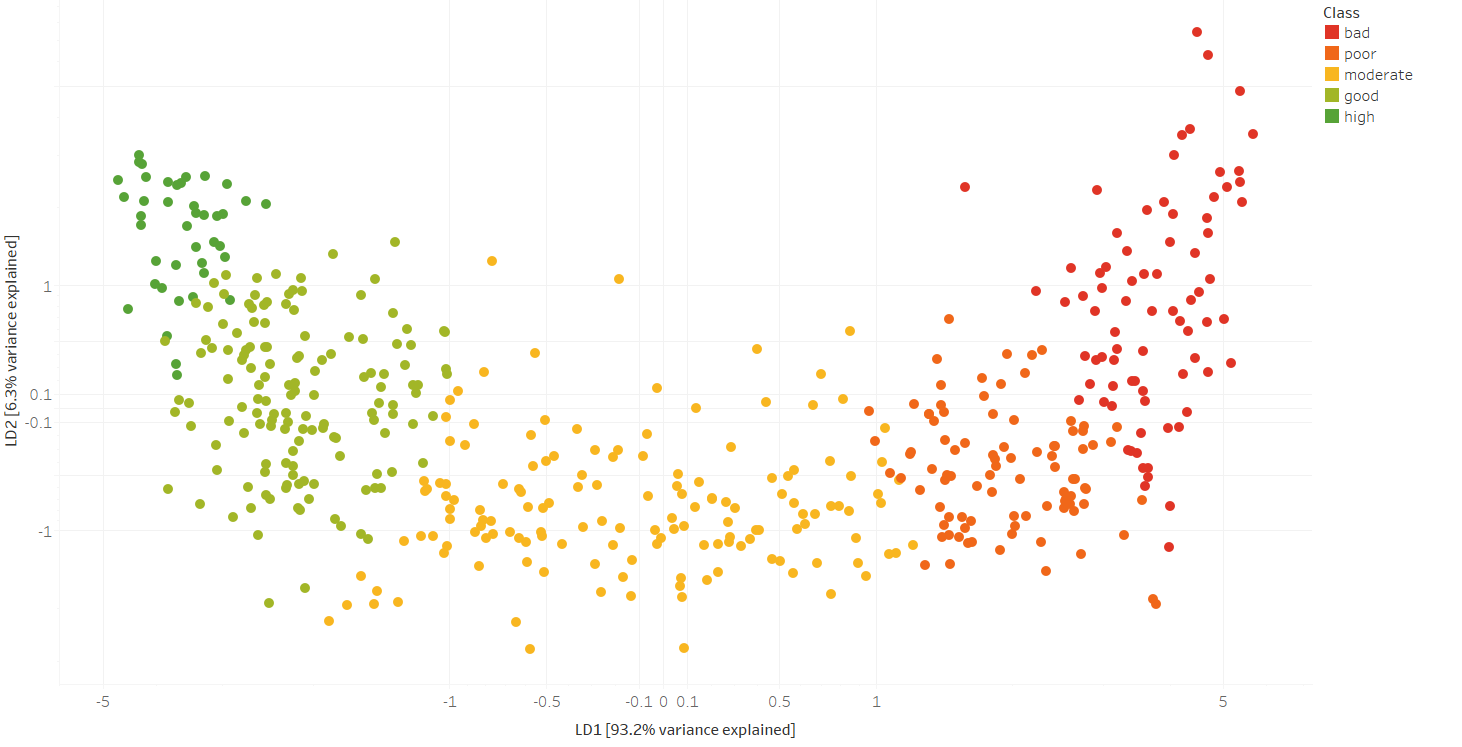


Figure 10. Self explanatory skill of 1st Linear Discriminant - minimum and miaximum values of classes 'good' and 'poor' are a signal of possible classifiaction issues in the future.

Figure 11. Explanatory skill of 1st and 2nd Linear Discriminants combined - inter-class overlaps are still visible, yet percent of variance explained reached ~99%.

**3.3.** **Classification Modeling with Hyperparameter Optimization**

Prior to modelling, the imputed dataset was split into the training test, and grid search sets, with split ratios of 65%, 20%, and 15% respectively. For the classification modeling, the SVM radial algorithm was used. In order to obtain optimal results, the grid search was performed, resulting in setting the best hyperparameter *‘C’* for a given classification task to the value of *1* with *sigma* value *2.31*, according to the tuning results of 25 repetitions (Table 7, Figure 12).

|  |  |
| --- | --- |
| model | Support Vector Machines Radial Basis Function Kernel |
| type | Classification |
| C-parameter (cost)– best tune | 1 |
| sigma – best tune | 2.310 |
| accuracy | 0.860 |
| kappa | 0.813 |
| resampling method | ‘bootstrap’ |
| resampling iterations | 25 |
| Leave Group Out CV parameter | 0.75 |

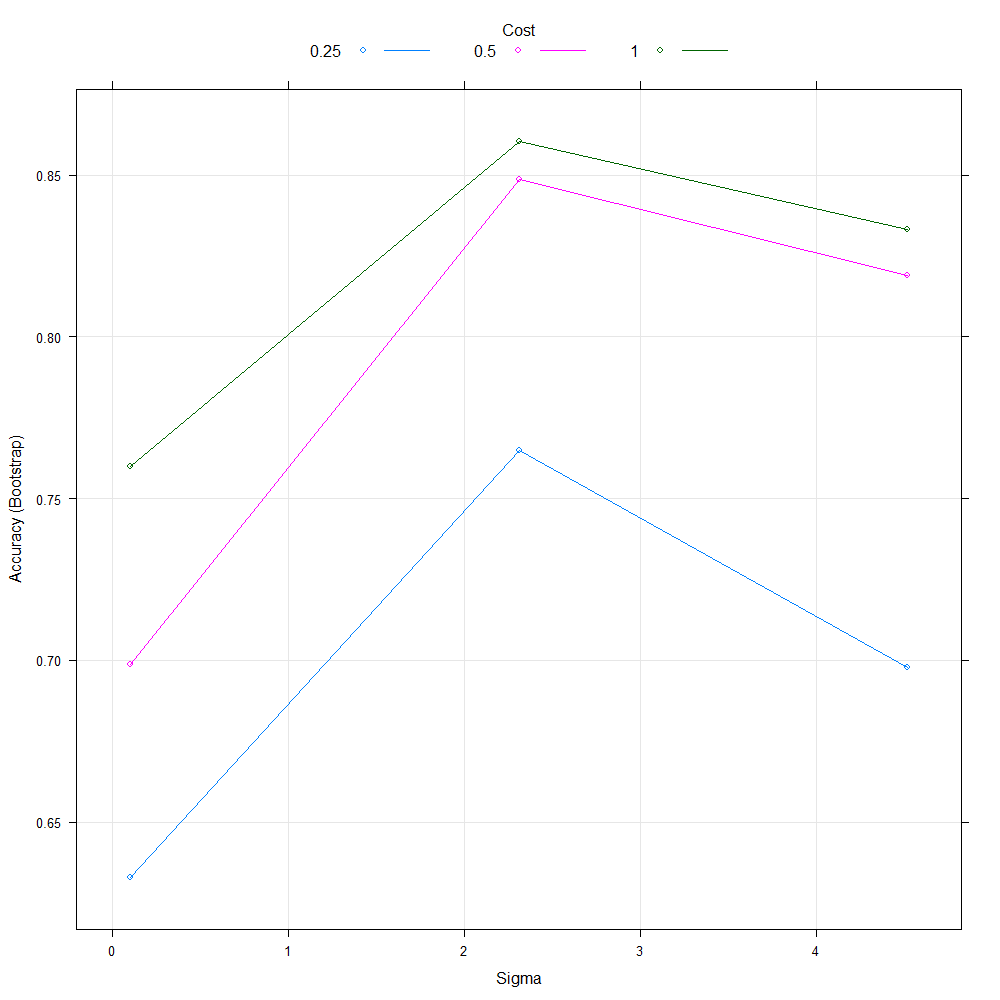
Table 7. SVM model initial parameters selected with grid search method.

Figure 12. Best tune of sigma parameter (2.31) obtained with Cost parameter of 1 - resulting in 0.86 fitting accuracy.

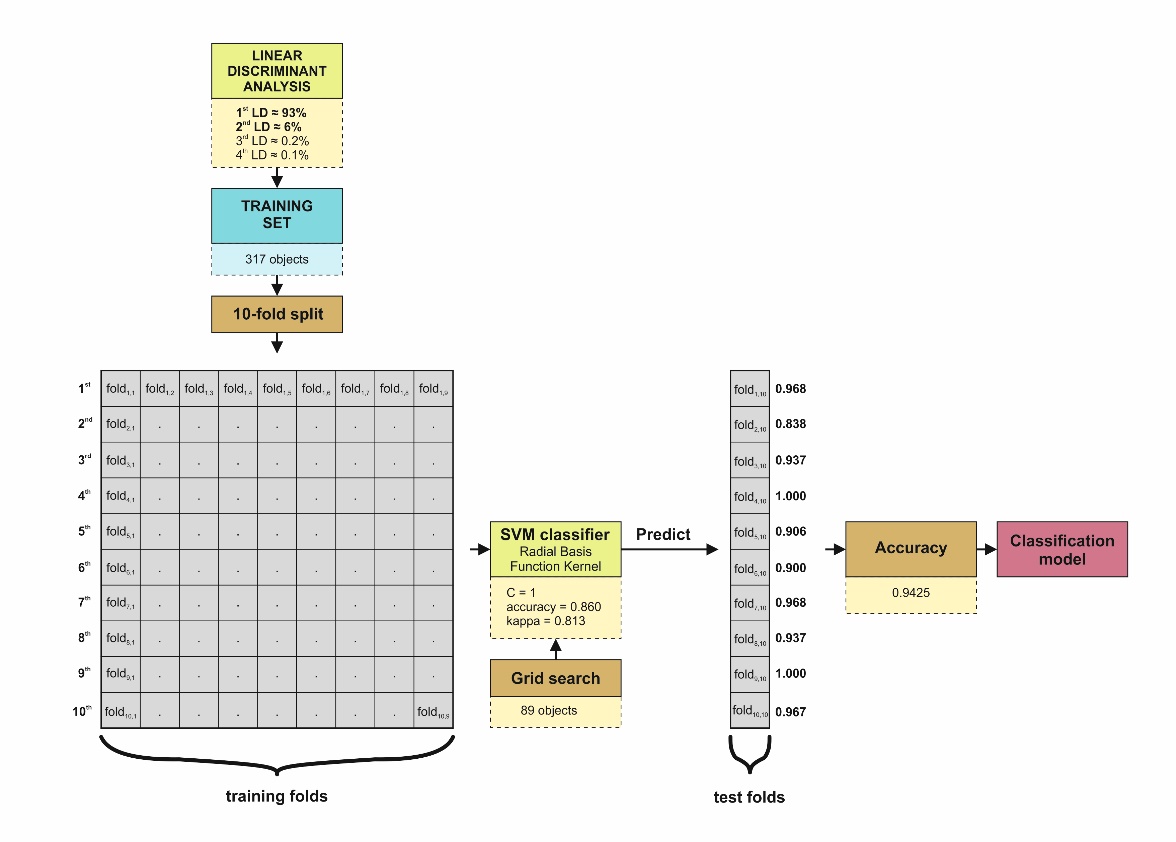
The tuned radial SVM function was then performed within the 10-fold cross-validation resampling procedure in order to obtain an evaluation of data samples (Figure 13). The training set was split into a 10x10 folds matrix, with the 10th column acting as a test set. An SVM classifier was tuned via grid search procedure and trained. Training accuracy was equal to 0.943. Next, the data from test folds were used to predict new outcomes, yelding ten accuracies ranging from 0.838 to 1.000. Overall mean accuracy of classification was equal to 0.9425. The obtained classification model was then evaluated on the previously created test set for analysis. During this procedure, the *caret* package for R was used (Kuhn, 2015). Most (9 out of 10) model iterations test results show more than 90% efficiency in object classification. Inter-class discrepancies were observed in the classification between bad - poor and moderate – poor, one error per each classification, resulting in overall four differences when related to original classification (Table 8). No misclassifications were found outside the directly neighbouring classes.

Figure 13. Process of SVM classifier development via k-fold testing matrice resulting in classification model with 0.94 performance accuracy.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Measured** | **Predicted** | | | | | |
|  | **bad** | **poor** | **moderate** | **good** | **high** |
| **bad** | 15 | 1 | 0 | 0 | 0 |
| **poor** | 0 | 17 | 1 | 0 | 0 |
| **moderate** | 0 | 1 | 28 | 0 | 0 |
| **good** | 0 | 0 | 1 | 27 | 0 |
| **high** | 0 | 0 | 9 | 0 | 9 |

Table 8. Confusion matrix of Measured-Predicted sets displaying classidication result with errors.

The accuracy was tested with 100 fold variants with k ranging from 10 to 500 to check for any changes in model performance. With k interval of 5, the best accuracy is obtained at k = 165 (0.95364), while worst behaviour is observed for k = 150 (0.92867), model stabilizes at about k = 320, reaching accuracy of 0.94239 (Figure 14).

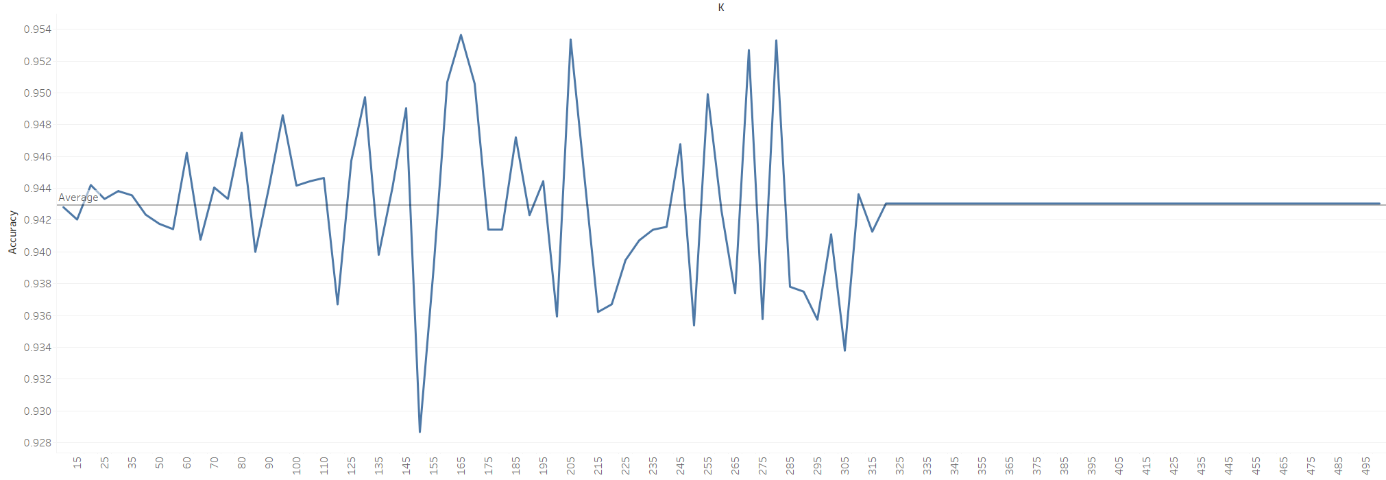


Figure 14. Accuracy changes for given k-folds - model stabilizes at k = 320 reaching 0.94239 arccuracy ratio.

**4. Discussion**

This study considered the explanatory ability of machine learning based ecological state classificator for lakes with reduced dimensionality based on measurements provided for 499 lakes in Poland. In the classifier, we made an attempt to recreate the classification procedure imitating the procedure in accordance with the principles of expert categorization (CIEP, 2015) and the WFD guidelines (European Environmental Agency, 2018) in order to achieve a tool to support future decision-making processes, as well as the detection of objects whose belonging to a given ecological state class should be particularly justified. When creating a classifier based on the non-linear SVM kernel approach, the dimensionality reduction technique (LDA) was used to reduce the set to two dimensions, which explain 99.5% of its internal variability. The model with the classification efficiency of 94.25% after ten standard iterations obtained stabilization after 320 runs at the level of 94.239% accuracy.

The final predictory skill was weakened by the presence of four classification errors. The same errors appear consistently in all of the test folds, for any proposed k. Even a robust kernel-based SVM model was unable to recreate observed grouping patterns. There are a number of possible reasons for this: (a) the unexplained variability of residual discriminants hides eigenvalues necessary for the correct classification of four objects; (b) There is an inconsequence in previous classification (e.x. due to high uncertainty), or (c) some remote factors exist which were taken into account - not included within the default grouping procedure (e.g. expert insights). Although we recognize the importance of lake-specific conditions, our approach is rather pragmatic in deriving class-membership changing parameters from relatively unspecific knowledge, which cannot be efficiently retraced without detailed habitat surveys.

In general, the goal of most models classifying the ecological state of ecosystems is universality, resistance to outliers, and optimal use of independent variables (Mouton, De Baets, & Goethals, 2010; Everaert et al., 2011). Moreover, it is just as important to cope with any ovelaps among classes resulting from classification mistakes, model shortcomings or existing uncertainties (Uusitalo, Lehikoinen, Helle, & Myrberg, 2015; Fernández et al., 2018). In this work, the dimensionality reduction of the set in conjunction with the non-linear classifier fulfilled the role of an effective tool, using patterns, without forcing interference in the stability of individual ecological status classes indicated in the WFD.

One of the main issues of ecological state classification was addressed in the research – that missing data often hamper grouping methods. The PCA imputation approach acted as an alternative to methods used by De’Ath (De’Ath & Fabricius, 2000), O’Hara (O’Hara, Arjas, Toivonen, & Hanski, 2002), and Ellington (Ellington et al., 2015) in ecological classification studies. We have chosen an approach which tends to minimize the absolute difference between correlations among variables prior to and after the imputation process while boosting the overall explanatory skill of developed discriminants, following the research of Dray (Dray & Josse, 2015) and Peres-Neto (Peres-Neto, Jackson, & Somers, 2003).

The use of dimensionality reduction on the analyzed dataset improved the overall readability of inter-class charactersitics. From the plotted LD1-LD2 results (Figure 11), the non-linear nature of class discriminants is recognizable, justifying the later use of SVM algorithm with radial kernel (Fernández-Delgado, Cernadas, Barro, & Amorim, 2014; Ustuner, Esetlili, Sanli, Abdikan, & Kurucu, 2016). The visually examined separability of particular classes allowed for spotting areas where the classification algorithm is likely to encounter problems. Despite the possibility of using the result of discriminant analysis as a classifier itself, as in Li (Li, Zhu, & Ogihara, 2006), we decided to use an additional non-linear classification algorithm to maximize the effectiveness of class separation as support vectors of division, following the encouraging results from Cao (Cao, Chua, Chong, Lee, & Gu, 2003) and Pal (Pal & Foody, 2010).

The pattern recognition based classification approach could become vital during model ensembling practices while dealing with incomplete, spatially-varying, or methodologically inconsequent datasets. In sets showing reproducible patterns it is easier to retrace object’s trajectories in time, between particular classes, and – what is also essential – within a given class. Thus, such approach of object observation does not require any class granularity changes until evident sub-classes emerge.

The concept of intra-class object classification gained popularity in research on dynamic systems modelling (Simovici & Djeraba, 2014; Szmidt, 2014). In the case of ecological state assessment, it is a default position for a lake being classified as belonging to a given eco-state class, while at the same time holding membership of some strength within the class. It is potentially beneficiary to observe the intra-class position of each object in order to deduce its tendencies (Slonneger, Zehna, & Johnson, 1974), for instance via answering the question: is the object a typical member of its class or does it demonstrate a shift towards any of neighboring classes? Once the tendencies of ecosystems are described in membership terms, it could be easier to explain why some of ecosystems fail to achieve desired ecological status. Such classification issues were widely discussed by Bennett (Bennett, Roca, Romero, & Alcoverro, 2011), Tarnita (Tarnita, Antal, Ohtsuki, & Nowak, 2009), Zienkiewicz (Zienkiewicz & Taylor, 2005), and Wallace (Wallace, 2007).

Non-linear classification algorithms are known to be robust for numerous pattern recognition tasks in applied ecology. Also, the use of dimensionality reduction techniques would be expected to improve the clarity and interpretability of multivariable datasets. Our results confirm this hypothesis, although the existance of misclassification posed the question on intra-class membership issues in sets. Classification errors were revealed for four subjects between poor-moderate-good classes. The obtained two-dimensional data set allowed to track the distribution and concentration of objects in each of the five ecological status classes. This demonstrates that even such simplification can improve our ability to recognize object diversity (dispersion) within ecological class, rising awareness of misclassifications, and that use of more sophisticated machine learning models may lead to assessment reproducibility improvement in future classification campaigns. An important element of the work was to draw attention to the possibility of effective reconstruction of even complex classification procedures. Additionally, due to reduction of dataset dimensions, the overall functionality of classification was maintained while enhancing the potential for visually-based data comprehension. As a result, the complexity of the observed data was reduced, and the reproducibility of classification method was structured within a model. It is particularly useful in the ecology of aquatic ecosystems, where between coarse resolution results and in-situ surveys many individual features are lost, especially when it comes to dealing with numerous measuring campaigns stretched in time.

**5. Conclusions**

Overall, the proposed framework successfully reproduced the ecological state assessment with visual recognition of patterns in measured parameters. Reducing the multidimensional set of observations to a two-dimensional space is convenient in terms of visualization. Furthermore, using the potential of latent variables made it possible to observe the behavior of individual lakes within their own ecological state classes. This is an extension of the analysis with valuable information for specialists undertaking the evaluation/monitoring of the effectiveness of actions increasing the ecological potential of ecosystems, as well as facilitating the decision-making process during the qualification of lakes for future ecological action programs. However, the target potential of the methodology will be unlocked when the model is applied to the results of previous measurement campaigns. Such approach can be used to re-classify objects from previous campaigns in order to acquire systematic object trajectories between and within each ecological class. Despite taking a step towards improving the usefulness of lake ecological state evaluation, we are aware of the need to formalize measures related to the position of objects in classes, for example for the purposes of structuring data in the report on water management plans, as well as harmonizing the results in water quality databases. Further developments are planned to improve the object membership indicators with use of Fuzzy SVMs and are focused on creation of effective, object-oriented membership assessment tool based on ecosystems ’individual tendency within its class.

**6. Literature**

Acreman, M. C., & Ferguson, A. J. D. (2010). Environmental flows and the European Water Framework Directive. *Freshwater Biology*. doi:10.1111/j.1365-2427.2009.02181.x

Alahuhta, J., Vuori, K. M., Hellsten, S., Järvinen, M., Olin, M., Rask, M., & Palomäki, A. (2009). Defining the ecological status of small forest lakes using multiple biological quality elements and palaeolimnological analysis. *Fundamental and Applied Limnology*. doi:10.1127/1863-9135/2009/0175-0203

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*. doi:10.1080/00401706.1971.10488811

Aoki, I. (2012). Lake Ecosystems. In *Entropy Principle for the Development of Complex Biotic Systems*. doi:10.1016/b978-0-12-391493-4.00007-x

Baho, D. L., Futter, M. N., Johnson, R. K., & Angeler, D. G. (2015). Assessing temporal scales and patterns in time series: Comparing methods based on redundancy analysis. *Ecological Complexity*. doi:10.1016/j.ecocom.2015.04.001

Benhadi-Marín, J. (2018). A conceptual framework to deal with outliers in ecology. *Biodiversity and Conservation*. doi:10.1007/s10531-018-1602-2

Bennett, S., Roca, G., Romero, J., & Alcoverro, T. (2011). Ecological status of seagrass ecosystems: An uncertainty analysis of the meadow classification based on the Posidonia oceanica multivariate index (POMI). *Marine Pollution Bulletin*. doi:10.1016/j.marpolbul.2011.06.016

Bennion, H., Kelly, M. G., Juggins, S., Yallop, M. L., Burgess, A., Jamieson, J., & Krokowski, J. (2014). Assessment of ecological status in UK lakes using benthic diatoms. *Freshwater Science*. doi:10.1086/675447

Bertsimas, D., Pawlowski, C., & Zhuo, Y. D. (2018). From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*.

Birk, S., Willby, N. J., Kelly, M. G., Bonne, W., Borja, a, Poikane, S., & Bund, W. Van De. (2013). Science of the Total Environment Intercalibrating classi fi cations of ecological status : Europe ’ s quest for common management objectives for aquatic ecosystems. *Science of the Total Environment*.

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. *Science of the Total Environment*. doi:10.1016/j.scitotenv.2013.03.037

Bode, M., Baker, C. M., Benshemesh, J., Burnard, T., Rumpff, L., Hauser, C. E., … Wintle, B. A. (2017). Revealing beliefs: using ensemble ecosystem modelling to extrapolate expert beliefs to novel ecological scenarios. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12703

Brownlee, J. (2019). A Gentle Introduction to k-fold Cross-Validation.

Cao, L. J., Chua, K. S., Chong, W. K., Lee, H. P., & Gu, Q. M. (2003). A comparison of PCA, KPCA and ICA for dimensionality reduction in support vector machine. *Neurocomputing*. doi:10.1016/S0925-2312(03)00433-8

Carvalho, L., Cortes, R., & Bordalo, A. A. (2011). Evaluation of the ecological status of an impaired watershed by using a multi-index approach. In *Environmental Monitoring and Assessment*. doi:10.1007/s10661-010-1473-9

Castellanos, A. A., Huntley, J. W., Voelker, G., & Lawing, A. M. (2019). Environmental filtering improves ecological niche models across multiple scales. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.13142

Chen, Q., Song, Z., Dong, J., Huang, Z., Hua, Y., & Yan, S. (2015). Contextualizing object detection and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:10.1109/TPAMI.2014.2343217

Chou, J. S., Ho, C. C., & Hoang, H. S. (2018). Determining quality of water in reservoir using machine learning. *Ecological Informatics*. doi:10.1016/j.ecoinf.2018.01.005

Christensen, O. (2010). Hilbert spaces. In *Applied and Numerical Harmonic Analysis*. doi:10.1007/978-0-8176-4980-7\_4

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.13256

Chrobak, K., Chrobak, G., & Kazak, J. K. (2020). The use of common knowledge in fuzzy logic approach for vineyard site selection. *Remote Sensing*. doi:10.3390/rs12111775

Cooper, J. L. B., & Berberian, S. K. (1977). Introduction to Hilbert Space. *The Mathematical Gazette*. doi:10.2307/3617432

Crisp, D. J., & Burges, C. J. C. (2000). A geometric interpretation of ν-SVM classifiers. In *Advances in Neural Information Processing Systems*.

Cunningham, J. P., & Ghahramani, Z. (2015). Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*.

De’Ath, G., & Fabricius, K. E. (2000). Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*. doi:10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

Desjardins-Proulx, P., Poisot, T., & Gravel, D. (2019). Artificial Intelligence for Ecological and Evolutionary Synthesis. *Frontiers in Ecology and Evolution*. doi:10.3389/fevo.2019.00402

Dörnhöfer, K., & Oppelt, N. (2016). Remote sensing for lake research and monitoring - Recent advances. *Ecological Indicators*. doi:10.1016/j.ecolind.2015.12.009

Dray, S., & Josse, J. (2015). Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*. doi:10.1007/s11258-014-0406-z

Ellington, E. H., Bastille-Rousseau, G., Austin, C., Landolt, K. N., Pond, B. A., Rees, E. E., … Murray, D. L. (2015). Using multiple imputation to estimate missing data in meta-regression. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12322

European Environmental Agency. (2018). *European waters Assessment of status and pressures 2018*. *Parents and Children Communicating with Society: Managing Relationships Outside of Home*. doi:/10.4324/9780203938607

Everaert, G., Boets, P., Lock, K., Džeroski, S., & Goethals, P. L. M. (2011). Using classification trees to analyze the impact of exotic species on the ecological assessment of polder lakes in Flanders, Belgium. *Ecological Modelling*. doi:10.1016/j.ecolmodel.2010.08.013

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*. doi:10.1117/1.JRS.11.015020

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. *Learning from Imbalanced Data Sets*. doi:10.1007/978-3-319-98074-4

Feurer, M., & Hutter, F. (2019). Hyperparameter Optimization. doi:10.1007/978-3-030-05318-5\_1

Flint, L. E., & Flint, A. L. (2012). Downscaling future climate scenarios to fine scales for hydrologic and ecological modeling and analysis. *Ecological Processes*. doi:10.1186/2192-1709-1-2

Forslund, a, & House, S. W. (2009). *Securing Water for Ecosystems and Human Well-being: The Importance of Environmental Flows*. *Swedish Water House Report*. doi:978-91-975872-4-2

Friedrichs, F., & Igel, C. (2005). Evolutionary tuning of multiple SVM parameters. *Neurocomputing*. doi:10.1016/j.neucom.2004.11.022

GIOŚ. (2015). Bank danych pomiarowych. Retrieved from https://powietrze.gios.gov.pl/pjp/archives

Gönen, M., & Alpaydin, E. (2011). Multiple kernel learning algorithms. *Journal of Machine Learning Research*.

Gooch, J. W. (2011). Pearson’s Product-Moment Correlation Coefficient. In *Encyclopedic Dictionary of Polymers*. doi:10.1007/978-1-4419-6247-8\_15318

Grizzetti, B., Lanzanova, D., Liquete, C., Reynaud, A., & Cardoso, A. C. (2016). Assessing water ecosystem services for water resource management. *Environmental Science and Policy*. doi:10.1016/j.envsci.2016.04.008

Grung, B., & Manne, R. (1998). Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*. doi:10.1016/S0169-7439(98)00031-8

Gu, Q., Li, Z., & Han, J. (2011). Linear discriminant dimensionality reduction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi:10.1007/978-3-642-23780-5\_45

Hamblin, S. (2013). On the practical usage of genetic algorithms in ecology and evolution. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12000

Hambright, K. D., & Berman, T. (2014). Water quality assessment. *Aquatic Ecology*. doi:10.1007/978-94-017-8944-8\_34

Hampton, S. E., Anderson, S., Bagby, S. C., & Gries, C. (2014). The Tao of Open Science for Ecology. *PeerJ PrePrints*. doi:10.7287/peerj.preprints.549v1

Harwood, T. D., Donohue, R. J., Williams, K. J., Ferrier, S., McVicar, T. R., Newell, G., & White, M. (2016). Habitat Condition Assessment System: a new way to assess the condition of natural habitats for terrestrial biodiversity across whole regions using remote sensing data. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12579

Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). The entire regularization path for the support vector machine. *Journal of Machine Learning Research*.

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *Annals of Statistics*. doi:10.1214/009053607000000677

Hooper, D. U., Chapin, F. S., Ewel, J. J., Hector, A., Inchausti, P., Lavorel, S., … Wardle, D. A. (2005). Effects of biodiversity on ecosystem functioning: A consensus of current knowledge. *Ecological Monographs*. doi:10.1890/04-0922

Hutter, F., Hoos, H., & Leyton-Brown, K. (2014). An efficient approach for assessing hyperparameter importance. In *31st International Conference on Machine Learning, ICML 2014*.

Jones-Farrand, D. T., Fearer, T. M., Thogmartin, W. E., Thompson, F. R., Nelson, M. D., & Tirpak, J. M. (2011). Comparison of statistical and theoretical habitat models for conservation planning: The benefit of ensemble prediction. *Ecological Applications*. doi:10.1890/10-1047.1

Josse, J., & Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software*. doi:10.18637/jss.v070.i01

Kaklauskas, A. (2014). The enterprise data lake: Better integration and deeper analytics. *PwC Technology Forecast: Rethinking Integration*. doi:10.1007/978-3-319-13659-2\_2

Keil, P., & Jetz, W. (2014). Downscaling the environmental associations and spatial patterns of species richness. *Ecological Applications*. doi:10.1890/13-0805.1

Kelly, F. L., Harrison, A. J., Allen, M., Connor, L., & Rosell, R. (2012). Development and application of an ecological classification tool for fish in lakes in Ireland. *Ecological Indicators*. doi:10.1016/j.ecolind.2012.01.028

Kelly, M. G., Birk, S., Willby, N. J., Denys, L., Drakare, S., Kahlert, M., … Poikane, S. (2016). Redundancy in the ecological assessment of lakes: Are phytoplankton, macrophytes and phytobenthos all necessary? *Science of the Total Environment*. doi:10.1016/j.scitotenv.2016.02.024

Khalid, S., Khalil, T., & Nasreen, S. (2014). A survey of feature selection and feature extraction techniques in machine learning. In *Proceedings of 2014 Science and Information Conference, SAI 2014*. doi:10.1109/SAI.2014.6918213

Kuhn, M. (2008). caret Package. *Journal Of Statistical Software*.

Kuhn, M. (2015). A Short Introduction to the caret Package. *R Foundation for Statistical Computing*.

Landi, P., Minoarivelo, H. O., Brännström, Å., Hui, C., & Dieckmann, U. (2018). Complexity and stability of ecological networks: a review of the theory. *Population Ecology*. doi:10.1007/s10144-018-0628-3

Lauer, F., & Bloch, G. (2019). Classification. In *Lecture Notes in Control and Information Sciences*. doi:10.1007/978-3-030-00193-3\_3

Li, T., Zhu, S., & Ogihara, M. (2006). Using discriminant analysis for multi-class classification: An experimental investigation. *Knowledge and Information Systems*. doi:10.1007/s10115-006-0013-y

Loucks, D. P., van Beek, E., Loucks, D. P., & van Beek, E. (2017). Water Resources Planning and Management: An Overview. In *Water Resource Systems Planning and Management*. doi:10.1007/978-3-319-44234-1\_1

Mantzouki, E., Campbell, J., Loon, E. Van, Visser, P., Konstantinou, I., Antoniou, M., … Ibelings, B. W. (2018). A European Multi Lake Survey dataset of environmental variables, phytoplankton pigments and cyanotoxins. *Scientific Data*. doi:10.1038/sdata.2018.226

Martín-López, B., Gómez-Baggethun, E., García-Llorente, M., & Montes, C. (2014). Trade-offs across value-domains in ecosystem services assessment. *Ecological Indicators*. doi:10.1016/j.ecolind.2013.03.003

Mayer, A., Winkler, R., & Fry, L. (2014). Classification of watersheds into integrated social and biophysical indicators with clustering analysis. *Ecological Indicators*. doi:10.1016/j.ecolind.2014.04.030

Mladenić, D. (2006). Feature selection for dimensionality reduction. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi:10.1007/11752790\_5

Moe, S. J., Haande, S., & Couture, R. M. (2016). Climate change, cyanobacteria blooms and ecological status of lakes: A Bayesian network approach. *Ecological Modelling*. doi:10.1016/j.ecolmodel.2016.07.004

Moler, C. B., & Stewart, G. W. (1973). An Algorithm for Generalized Matrix Eigenvalue Problems. *SIAM Journal on Numerical Analysis*. doi:10.1137/0710024

Morelli, F., & Tryjanowski, P. (2016). The dark side of the “redundancy hypothesis” and ecosystem assessment. *Ecological Complexity*. doi:10.1016/j.ecocom.2016.07.005

Morrissey, M. B. (2014). In search of the best methods for multivariate selection analysis. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.12259

Mouton, A. M., De Baets, B., & Goethals, P. L. M. (2010). Ecological relevance of performance criteria for species distribution models. *Ecological Modelling*. doi:10.1016/j.ecolmodel.2010.04.017

Muttil, N., & Chau, K. W. (2007). Machine-learning paradigms for selecting ecologically significant input variables. *Engineering Applications of Artificial Intelligence*. doi:10.1016/j.engappai.2006.11.016

Navarro, E., Caputo, L., Marcé, R., Carol, J., Benejam, L., Garcia-Berthou, E., & Armengol, J. (2009). Ecological classification of a set of Mediterranean reservoirs applying the EU Water Framework Directive: A reasonable compromise between science and management. *Lake and Reservoir Management*. doi:10.1080/07438140903238567

Nguyen, L. H., & Holmes, S. (2019). Ten quick tips for effective dimensionality reduction. *PLoS Computational Biology*. doi:10.1371/journal.pcbi.1006907

O’Hara, R. B., Arjas, E., Toivonen, H., & Hanski, I. (2002). Bayesian analysis of metapopulation data. *Ecology*. doi:10.1890/0012-9658(2002)083[2408:BAOMD]2.0.CO;2

Padisák, J., Borics, G., Grigorszky, I., & Soróczki-Pintér, É. (2006). Use of phytoplankton assemblages for monitoring ecological status of lakes within the water framework directive: The assemblage index. *Hydrobiologia*. doi:10.1007/s10750-005-1393-9

Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*. doi:10.1109/TGRS.2009.2039484

Park, C.-K., & Kim, D. G. (2012). K-FOLD Cross-Validation. *Progress in Neurological Surgery*. doi:10.1159/000331070

Parrott, L. (2002). Complexity and the limits of ecological engineering. *Transactions of the American Society of Agricultural Engineers*. doi:10.13031/2013.11032

Pasetto, D., Arenas-Castro, S., Bustamante, J., Casagrandi, R., Chrysoulakis, N., Cord, A. F., … Ziv, G. (2018). Integration of satellite remote sensing data in ecosystem modelling at local scales: Practices and trends. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.13018

Penning, W. E., Dudley, B., Mjelde, M., Hellsten, S., Hanganu, J., Kolada, A., … Ecke, F. (2008). Using aquatic macrophyte community indices to define the ecological status of European lakes. *Aquatic Ecology*. doi:10.1007/s10452-008-9183-x

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2003). Giving meaningful interpretation to ordination axes: Assessing loading significance in principal component analysis. *Ecology*. doi:10.1890/00-0634

Peterson, G. D., Douglas Beard, T., Beisner, B. E., Bennett, E. M., Carpenter, S. R., Cumming, G. S., … Havlicek, T. D. (2003). Assessing future ecosystem services: A case study of the Northern Highlands Lake District, Wisconsin. *Ecology and Society*. doi:10.5751/es-00557-070301

Piroddi, C., Teixeira, H., Lynam, C. P., Smith, C., Alvarez, M. C., Mazik, K., … Uyarra, M. C. (2015). Using ecological models to assess ecosystem status in support of the European Marine Strategy Framework Directive. *Ecological Indicators*. doi:10.1016/j.ecolind.2015.05.037

Poikane, S., Zampoukas, N., Borja, A., Davies, S. P., van de Bund, W., & Birk, S. (2014). Intercalibration of aquatic ecological assessment methods in the European Union: Lessons learned and way forward. *Environmental Science and Policy*. doi:10.1016/j.envsci.2014.08.006

Powers, S. M., & Hampton, S. E. (2019). Open science, reproducibility, and transparency in ecology. *Ecological Applications*. doi:10.1002/eap.1822

Qu, M., Bengio, Y., & Tang, J. (2019). GMNN: Graph markov neural networks. In *36th International Conference on Machine Learning, ICML 2019*.

Rask, M., Olin, M., & Ruuhijärvi, J. (2010). Fish-based assessment of ecological status of Finnish lakes loaded by diffuse nutrient pollution from agriculture. *Fisheries Management and Ecology*. doi:10.1111/j.1365-2400.2009.00685.x

Rodríguez, J. D., Pérez, A., & Lozano, J. A. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. doi:10.1109/TPAMI.2009.187

Schaumburg, J., Schranz, C., Hofmann, G., Stelzer, D., Schneider, S., & Schmedtje, U. (2004). Macrophytes and phytobenthos as indicators of ecological status in German lakes - A contribution to the implementation of the Water Framework Directive. *Limnologica*. doi:10.1016/S0075-9511(04)80003-3

Schneider, F. D., Fichtmueller, D., Gossner, M. M., Güntsch, A., Jochum, M., König-Ries, B., … Simons, N. K. (2019). Towards an ecological trait-data standard. *Methods in Ecology and Evolution*. doi:10.1111/2041-210X.13288

Schuwirth, N., Borgwardt, F., Domisch, S., Friedrichs, M., Kattwinkel, M., Kneis, D., … Vermeiren, P. (2019). How to make ecological models useful for environmental management. *Ecological Modelling*. doi:10.1016/j.ecolmodel.2019.108784

Simovici, D. A., & Djeraba, C. (2014). Sets, relations, and functions. In *Advanced Information and Knowledge Processing*. doi:10.1007/978-1-4471-6407-4\_1

Slonneger, K. R., Zehna, P. W., & Johnson, R. L. (1974). Elements of Set Theory. *The American Mathematical Monthly*. doi:10.2307/2319332

Society, B. E. (2017). A guide to reproducible code in ecology and evolution. *BES Guides to Better Science*. doi:10.1007/s11858-009-0177-5

Søndergaard, M., Jeppesen, E., Jensen, J. P., & Amsinck, S. L. (2005). Water Framework Directive: Ecological classification of Danish lakes. *Journal of Applied Ecology*. doi:10.1111/j.1365-2664.2005.01040.x

Søndergaard, M., Larsen, S. E., Jørgensen, T. B., & Jeppesen, E. (2011). Using chlorophyll a and cyanobacteria in the ecological classification of lakes. *Ecological Indicators*. doi:10.1016/j.ecolind.2011.03.002

Szmidt, E. (2014). Distances and similarities in intuitionistic fuzzy sets. *Studies in Fuzziness and Soft Computing*. doi:10.1007/978-3-319-01640-5\_1

Tarnita, C. E., Antal, T., Ohtsuki, H., & Nowak, M. A. (2009). Evolutionary dynamics in set structured populations. *Proceedings of the National Academy of Sciences of the United States of America*. doi:10.1073/pnas.0903019106

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications*. doi:10.3233/AIC-170729

UN-Water. (2017). The United Nations World Water Development Report 2017: Wastewater, The Untapped Resource. *Journal of Chemical Information and Modeling*. doi:10.1017/CBO9781107415324.004

Ustuner, M., Esetlili, M. T., Sanli, F. B., Abdikan, S., & Kurucu, Y. (2016). Comparison of crop classification methods for the sustainable agriculture management. *Journal of Environmental Protection and Ecology*.

Uusitalo, L., Lehikoinen, A., Helle, I., & Myrberg, K. (2015). An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environmental Modelling and Software*. doi:10.1016/j.envsoft.2014.09.017

Van Der Lee, G. E. M., Van Der Molen, D. T., Van Den Boogaard, H. F. P., & Van Der Klis, H. (2006). Uncertainty analysis of a spatial habitat suitability model and implications for ecological management of water bodies. *Landscape Ecology*. doi:10.1007/s10980-006-6587-7

Wallace, K. J. (2007). Classification of ecosystem services: Problems and solutions. *Biological Conservation*. doi:10.1016/j.biocon.2007.07.015

Wang, J., Nathan, R., Horne, A., Peel, M. C., Wei, Y., & Langford, J. (2017). Evaluating four downscaling methods for assessment of climate change impact on ecological indicators. *Environmental Modelling and Software*. doi:10.1016/j.envsoft.2017.06.016

Weyl, O. L. F., Ribbink, A. J., & Tweddlel, D. (2010). Lake Malawi: Fishes, fisheries, biodiversity, health and habitat. *Aquatic Ecosystem Health and Management*. doi:10.1080/14634988.2010.504695

Willcock, S., Martínez-López, J., Hooftman, D. A. P., Bagstad, K. J., Balbi, S., Marzo, A., … Athanasiadis, I. N. (2018). Machine learning for ecosystem services. *Ecosystem Services*. doi:10.1016/j.ecoser.2018.04.004

Xu, F. L., Tao, S., Dawson, R. W., Li, P. gang, & Cao, J. (2001). Lake ecosystem health assessment: Indicators and methods. *Water Research*. doi:10.1016/S0043-1354(01)00040-9

Xu, F. L., Zhao, Z. Y., Zhan, W., Zhao, S. S., Dawson, R. W., & Tao, S. (2005). An ecosystem health index methodology (EHIM) for lake ecosystem health assessment. *Ecological Modelling*. doi:10.1016/j.ecolmodel.2005.01.058

Yajima, H., & Derot, J. (2018). Application of the Random Forest model for chlorophyll-a forecasts in fresh and brackish water bodies in Japan, using multivariate long-term databases. *Journal of Hydroinformatics*. doi:10.2166/hydro.2017.010

Ye, J. (2007). Least squares linear discriminant analysis. In *ACM International Conference Proceeding Series*. doi:10.1145/1273496.1273633

Zhang, L. L., Liu, J. L., Yang, Z. F., Li, Y., & Yang, Y. (2013). Integrated ecosystem health assessment of a macrophyte-dominated lake. *Ecological Modelling*. doi:10.1016/j.ecolmodel.2012.07.029

Zienkiewicz, O. C., & Taylor, R. L. (2005). *The Finite Element Method Set*. *The Finite Element Method Set*.