INTERPRETATIVE MACHINE LEARNING AS A KEY IN RECOGNIZING THE VARIABILITY OF LAKES TROPHY PATTERNS

Jarosław Jasiewicz [1], Izabela Zawiska [2], Monika Rzodkiewicz [1], Michał Woszczyk [1]

[1] Institute of Geoecology and Geoinformatics, Adam Mickiewicz University, Poznań, Poland
[2] Institute of Geography and Spatial Organization, Polish Academy of Science, Warszawa, Poland

Manuscript received: February 14, 2022
Revised version: March 15, 2022

Jasiewicz J., Zawiska I., Rzodkiewicz M., Woszczyk M., 2022. Interpretative machine learning as a key in recognizing the variability of lakes trophy patterns. *Quaestiones Geographicae* 41(1), Bogucki Wydawnictwo Naukowe, Poznań, pp. 127–146. 11 figs, 1 table.

Abstract: The paper presents an application of interpretative machine learning to identify groups of lakes not with similar features but with similar potential factors influencing the content of total phosphorus – P_{tot}. The method was developed on a sample of 60 lakes from North-Eastern Poland and used 25 external explanatory variables. Selected variables are stable over a long time, first group includes morphometric parameters of lakes and the second group encompass watershed geometry geology and land use. Our method involves building a regression model, creating an explainer, finding a set of mapping functions describing how each variable influences the outcome, and finally clustering objects by 'the influence'. The influence is a non-linear and non-parametric transformation of the explanatory variables into a form describing a given variable impact on the modeled feature. Such a transformation makes group data on the functional relations between the explanatory variables and the explained variable possible. The study reveals that there are five clusters where the concentration of P_{tot} is shaped similarly. We compared our method with other numerical analyses and showed that it provides new information on the catchment area and lake trophy relationship.

Keywords: total phosphorus, interpretative machine learning, random forest, Masurian lakes

Corresponding authors: Jarosław Jasiewicz, jarekj@amu.edu.pl, Izabela Zawiska, izawiska@twarda.pan.pl

Introduction

A trophic state relates to the productivity of the lake (i.e., the availability of nutrients) and acts as one of the key characteristics of aquatic ecosystems. Natural factors largely control the trophic state of aquatic systems; however, they are also prone to anthropogenic disturbances. The latter are responsible for considerable changes in lakes during the Anthropocene. Therefore, much attention is devoted to tracing the long-term trophic evolution of lakes and understanding the controls on the trophic state of lakes in the present-day human-affected environment. Even though the trophic state is a long-term condition (Rodhe 1969, Schindler 1977), measurable indicators that define this state may vary periodically. Factors determining trophy indicators are roughly divided into a stable for a long time and changeable over a shorter period. The former include physiographic properties of the lake and its catchment, and the latter connects with seasonal processes

https://doi.org/10.2478/quageo-2022-0009
ISSN 2082-2103, eISSN 2081-6383

causing changes in the physical and biogeochemical properties of water. Ohle index, Schindler index, mean water depth, river network density, slope steepness, the share of endorheic areas, geological structure and land use in the catchment area are features regarded as controlling the trophy state (Bajkiewicz-Grabowska 2020). In-situ measurements of trophy indicators provide the best sources to build accurate models; however, such data are sparse, often incomplete, or access to it is restricted (Hollister et al. 2016). The information on the status of the trophy itself does not allow for a complex analysis of the responsible factors and thus, the reliable prediction of potential changes. With the rapid diffusion of geoscience and information technologies in the last decades (Chen et al. 2021), there is growing attention on numerical modelling in the many aspects of environmental changes. Thus, a possible solution for such a problem is predictive modelling that involves data obtained from public repositories like land cover/land use, lake basin geometry, and geology or directly calculated using GIS software.

Many papers describe the methods used to predict water trophies using watershed variables (Akbar et al. 2011, Benedini, Tsakiris 2013, Borics et al. 2013, Sun, Scanlon 2019, Gorgoglione et al. 2020). The efficiency of several multivariate analyses, including clustering, discriminant analysis, and principal component analysis, has been proven to reduce data complexity and detect intrinsic patterns in the underlying data. (Simeonov et al. 2010, Su et al. 2011, Li et al. 2017, 2018, Cui et al. 2019, Eliasz-Kowalska, Wojtal 2020). Such an approach has one considerable weakness: the detected patterns always show the internal differences of the explanatory variables, which does not always refer to the trophic state of the lakes. The linear multiple regression models (Jones et al. 2001, 2004, Beaulieu et al. 2014, Leach et al. 2018) provide insight into the relationship between explanatory variables and values of trophy indicators, but those methods are irrespective of the fact that most of the relationships are non-linear (Dormann et al. 2013, Huang et al. 2015).

New data acquisition techniques in geochemical surveys provide hundreds or thousands of observations described by tens or hundreds of features. When clarity of interpretation is more important than the model's accuracy, simple models such as linear models or regression trees (Froeschke, Froeschke 2011) usually provide sufficient insight into relationships between factors and the outcome at the expense of the prediction quality. Complex models cannot be directly explained because they are not easy to understand. Some of the learners, including random forest (RF) (Breiman 2001), provide a measure allowing to estimate the relative importance of variables used, thus identifying such a subset that influences the variation of trophy factors.

The abundance of the data provides new opportunities, but however, it is challenging to investigate and interpret the role of many environmental features (Dafforn et al. 2015). It is even more difficult to understand relationships between elements of the system and their influence on the outcome because of complex relationships inside the multidimensional data. Regression models are a natural approach in searching relations between explanatory and dependent variables. Simple methods do not provide interpretable results when variables are related to each other or mutually convoluted. Complex models like assembles (Hollister et al. 2016, Li et al. 2016) or neural networks (Li et al. 2015, Rocha et al. 2017, Gebler et al. 2021) are then the only solutions; however, such models cannot directly be used for interpretation because they are not easy to understand. Moreover, many variables are unrelated to the studied phenomenon and are usually removed based on researchers' experience or previous studies. Such removal, however, leads to the replication of the same, limited set of variables in subsequent studies (Goggin 1986, Harrell 2015).

There is no simple data analysis that would combine the advantages of supervised methods, like finding important variables and detecting relationships between explanatory variables and grouping – i.e. searching for new, possibly unknown patterns in the data. The first solution is optimisation: an iterative search of a combination of variables until the most optimal subset is found (Jasiewicz et al. 2021). An alternative solution is to create a non-linear regression model and then analyse it with interpretative machine learning (EML) (Molnar et al. 2020, Chen et al. 2021). Several tools have been recently proposed including Partial Dependency plots (Friedman 2001), local interpretable model-agnostic explanations (LIME) – Ribeiro et al. 2016), Learning Important

Features Through propagating activation differences (DeepLIFT) – (Shrikumar et al. 2017), moDel Agnostic Language and eXplanation (DALEX) – (Biecek 2018) and SHapley Additive exPlanations (SHAP) (Lundberg, Lee 2017). These methods replace original values of explanatory variables with functional relationships between the explanatory and explained variables; in simple words the influence of a given variable on the result. The latter means a function that operates on an original variable value and replaces it with variable influence on the outcome. This paper introduces a new term: the *variable influence*, providing a new insight into the relationship between the lake's surroundings and the value of trophic indicators. Moreover, *the influence* allows clustering the data, not by its original values denoted, but on the functional dependence between the explanatory and explained variables, creating a bridge between supervised and unsupervised learning.

The research presented in this paper aims to develop a solution allowing for clustering so that the resulting clusters minimise the dissimilarities inside the explanatory features and reduce the variation of the dependent variable inside the received clusters. The method was developed on a sample of 60 lakes from North-Eastern Poland. The lakes selected for this study are small and moderate in terms of their area and have relatively simple morphology but are sufficiently diverse to represent that geographical zone. In addition, the lakes were selected to obtain the strongest trophic gradient possible, expressed in terms of total phosphorus (P_{tot}). Although the method was designed for analysing the complex relationships between environmental variables that we believe impact the P_{tot} index, it can be easily applied to other complex systems. Thus, the data collection will be used as a case study to discuss the possibilities of the proposed method in practice.

Study area

The method was developed based on data collected from 60 lakes (Fig. 1) on the border of Suwalki and Masurian Lake District (SML). SML is an area of glacial and fluvioglacial origin, formed during the Pomeranian phase of the Weichselian glaciation between 24 k years and 19 k years BP (Marks 2012, Pochocka-Szwarc 2013). Dominant landforms include undulating morainic plateau with some hummocky and fluted till plains and washboard moraines (Weckwerth et al. 2019). Quaternary deposits are thick and contain typical components: tills, sands, silts, glaciofluvial gravels, and boulders. Lakes are the dominant and feature component of the SML landscape (Morawski 2005). All lakes are of glacial origin and are associated with the moraine plateau, inter-moraines and subglacial gutters. (Kondracki 2009, Pochocka-Szwarc 2013).

Variables

Lake water sampling and determination of P_{tot}

Epilimnion water samples were collected at the lake's deepest point 1 m below the water surface with UWITEC sampler. Each lake was sampled once, and the samples were taken during three field campaigns in summers 2018 (east/central sector in Fig. 1), 2019 (central/west sector in Fig. 1), and 2020 (east-central-west sector in Fig. 1). The selection of 1 m depth as a representation of surface water followed the methodology of Tandyrak et al. (2020). The deepest point is routinely regarded as representative for the whole lake (Tylmann et al. 2012, Hernández-Almeida et al. 2017, Apolinarska et al. 2020). Chemical analysis of P_{tot}, (μgL^{-1}) was done within a few days after collecting. P_{tot} was analysed spectrophotometrically using Nanocolor VIS; (Macherey-Nagel) with ammonium molybdate according to *PN-EN ISO 6878:2006P* after mineralisation with HNO_3 and H_2O_2 in UV Mineral 6.1. The repeatability of *P* determination expressed as a relative standard deviation (RSD) from duplicate measurements was between 0.3% and 5.5%. Analytical accuracy was estimated using certified reference materials (CRM 398–399: Major elements in seawater; ION 96: Hard river water from Grand River) and was between 87% and 93%. The content of phosphorus in the samples ranges from 0 μgL^{-1} to 70 μgL^{-1}, but the dominant values are below 20 μgL^{-1}.

Explanatory variables

The lake's trophy is influenced by catchment factors responsible for the supply of matter,

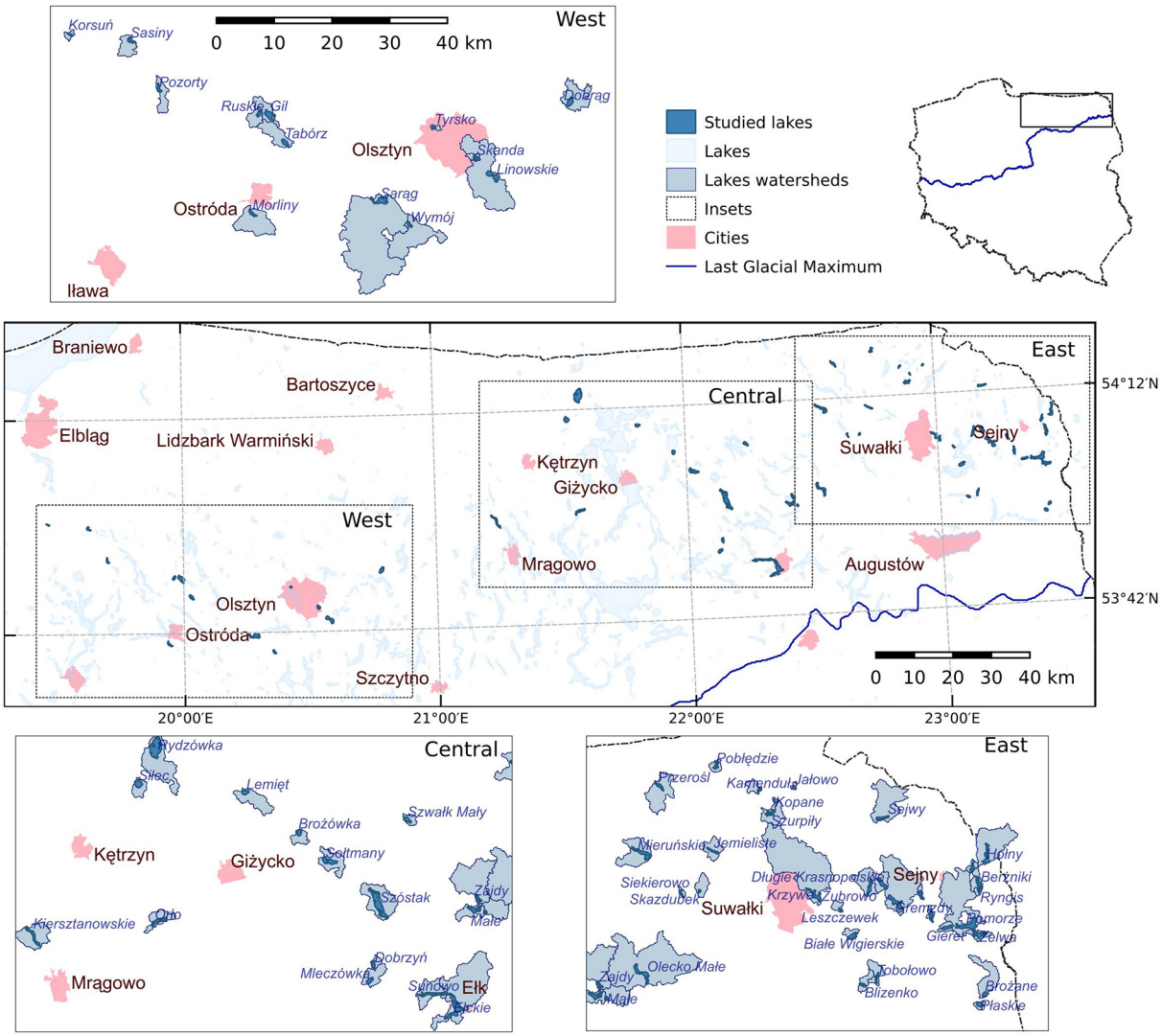

Fig. 1. Location of lakes and extent of the watershed on the study area.

including nutrients and the morphometric parameters of the lake, which mainly determine the resistance of the lake to the influence of the catchment area. In our work, we considered all the indicators that could be obtained. We collected 25 explanatory variables, having a potential effect on the P_{tot} concentration in the water. According to Bajkiewicz-Grabowska (2020), the first group includes the Ohle index, the type of lake water balance (flow rate), the density of the river network and average slope, the percentage of non-inflow areas in the direct catchment, and lithology land use. The second group encompasses average depth, the ratio of the lake volume to the shoreline length, the percentage of the hypolimnion in the lake volume, the Schindler index, the active bottom area, and the water exchange rate in the lake.

The flow rate was omitted due to the lack of information on the amount of flow and the river network density because the area of the studied catchments is too small to develop such a network. According to the authors' best knowledge, there are no measurement data on strictly hydrological processes, such as flow volume and water change time. Lange (1986) and Kalff (2001) suggest calculating those parameters using lake morphometry, but collected data already includes this information. We also omitted all factors that can be affected by the processes inside the lakes. Thus, the list of variables is limited only to stable variables over a long period; this eliminates factors that change seasonally (such as water temperature) or in short-term cycles (i.e., weather conditions).

The list of variables presented in Table 1 is divided into two groups. The first group includes morphometric parameters of lakes taken from The Atlas of Polish Lakes (Jańczak 1999) or ratios calculated directly from those features. The second group includes watershed geometry, morphometric parameters, lithology and land cover of the catchment area. In the first step, watersheds were delineated over the 30 m resolution Digital Elevation Model (DEM) DETD Level 2, (DEM in the rest of the paper) using GRASS GIS module r.stream.basins (Jasiewicz, Metz 2011). Finally, the geometry of the watershed was directly used to calculate the structure of their coverages and contribute to Schindler (1977) and Ohle (1956) ratios that define the relationship between lake and watershed geometry.

Table 1. Explanatory variables used in the study.

Variable	Abbreviation	Unit	Source
Elevation	ELEV	m a.s.l.	Jańczak 1999
Lake area	LARE	ha	Jańczak 1999
Lake capacity	LCAP	km^3	Jańczak 1999
Lake max depth	LDMX	m	Jańczak 1999
Lake average depth	LDAV	m	Jańczak 1999
Lake max length	LLEN	m	Jańczak 1999
Lake max width	LWID	m	Jańczak 1999
Lake shoreline length	PRIM	m	Jańczak 1999
Lake elongation	LELN	Ratio	*LLEN / LWID*
Lake capacity/ length ratio	LVAR	Ratio	LCAP / LLEN
Lake perim development	LPDV	Ratio	LLEN / sqrt (2 × π × LARE)
Lake exposition	LEXP	Ratio	LARE / LDAV
Watershed area	WARE	ha	Calculated
Mean slope	WSLP	%	Calculated
Height stddev	WHSD	m	Calculated
Urbanised	WURB	%	Calculated
Agriculture	WAGR	%	Calculated
Forests	WFRS	%	Calculated
Wetlands	WWET	%	Calculated
Sands	WSND	%	Calculated
Tills	WTLS	%	Calculated
Clay	WCLS	%	Calculated
Organic	WORG	%	Calculated
Schindler ratio	SR	Ratio	*WARE / LCAP*
Ohle ratio	OR	Ratio	WARE / LARE

The upper part contains variables describing lakes morphometry, the lower part watersheds parameters. Source 'Atlas' denote variables read from Jańczak (1999), source 'Calculated', means variables values were calculated using GIS software and spatial data. See text for details. Column 'Abbreviation' contains symbols used later in the text. Variables starting with L refer to lakes morphometry, variables starting with W- to catchment properties.

Information about the land cover, geology, and basic morphometry of watersheds surfaces was obtained using Corine Land Cover (CLC) 2018 (EEA 2018) and Geological Map of Poland (GMP) 1:500,000 (Marks et al. 2006) was used to calculate coverage properties, including land-cover/land-use and surface geology. Because both maps contain units with complex characteristics, they were simplified. CLC was reduced to first-level CLC units (urbanised, agriculture, forest, and wetlands, excluding given lake), while GMP to basic lithological units (tills, sands, clays, and organic). The analysed group of lakes is located within one geomorphological division of the last glaciation thus, the stratigraphic distinction of the lithological unit was neglected. Coverage variables are expressed as a percentage of a given unit in the watershed area (WARE), separately for CLC and GMP. The DEM was used to calculate geomorphometric features such as watersheds height standard deviation (WHSD) and watersheds mean slope (WSLP) inclination of the terrain in the watersheds.

Methods

Our preliminary observations show that each of the single collected variables is weakly related

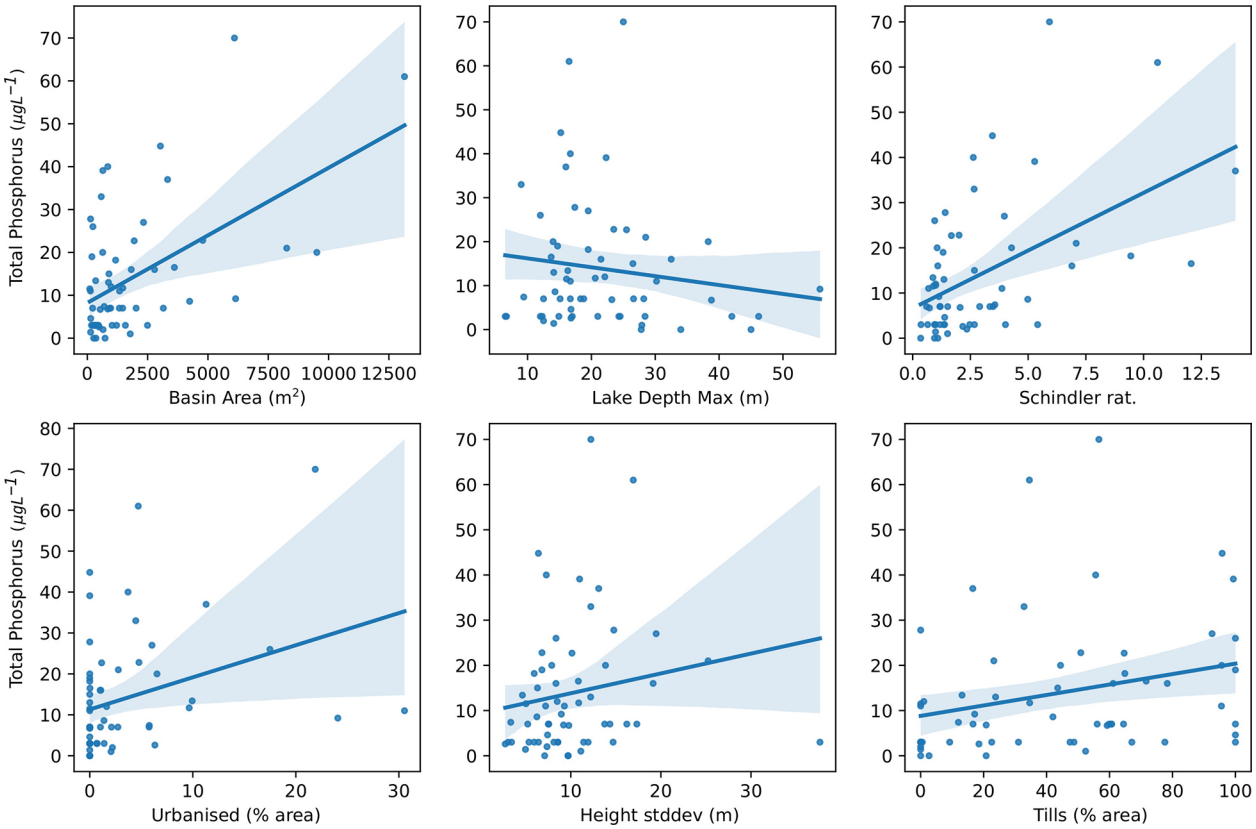

Fig. 2. Relations between P_{tot} and selected explanatory variables, see Table 1 for details.

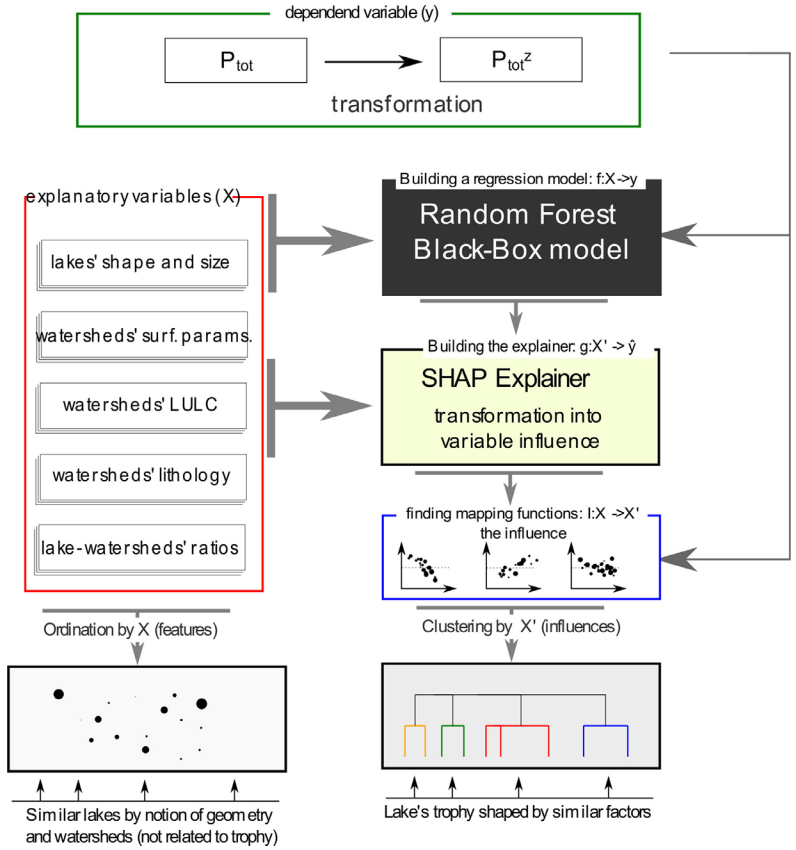

Fig. 3. The outline of methodology. See Section 'Methods' for details.

to the trophy of the lakes (Fig. 2). Such a situation precludes the creation of simple relations, such as trophy index variable. Our method involves four steps presented in Figure 3 and described below in detail: (1) Building a regression model: f: $X \rightarrow y$; (2) Building a simplified model g: $X' \rightarrow \hat{y}$ over the model, called the explainer; (3) finding a set of mapping functions: I: $X \rightarrow X'$, i.e. variable influences; (4) dissimilarity analysis and clustering objects by X'.

Building a regression model

The first step is to build an explanatory regression model, f: $X \rightarrow y$, where X is a set of explanatory variables (see also Table 1); f describes how the given variable X_i influences y; where y is an explained variable, here P_{tot}. We used a RF regression model (Breiman 2001), commonly used in many ecological and natural studies (Hollister et al. 2016, Bourel, Segura 2018, Leach et al. 2018, Li et al. 2018). The RF is a machine learning algorithm that grows a subset of the so-called weak predictors as shallow regression trees by bootstrapping samples of the training set. Those trees are non-parametric models; thus, the entire RF does not require a prior assumption about the variable distribution. It means in practice that RF accepts X in original form without the preceding transformation. Each tree grows recursively until it meets its stop criterion. At each step of growth, y is clustered in two child nodes to maximise y homogeneity inside these clusters. Then the best split on one of the X variables is selected. The RF is a bagging algorithm, which means that each tree grows on the independent subset of cases and variables. The importance of each variable depends only on its potential for reducing mean squared error between actual values of y and the outcome $\hat{y}$. For that reason, namely the random selection of variables, RF is more suitable for explanation than other machine learning algorithms.

Building the explainer

The second step of analysis includes the building of the explainer. Explainer g: $X' \rightarrow \hat{y}$ is a transformation of a previously trained complex model (here RF) into its interpretable approximation (Lundberg, Lee 2017), where X' is a transformed X, and $\hat{y}$ predicted values of y. All mentioned explainers assign an influence to each explanatory variable for a particular prediction – i.e., single case. The explanation process starts from the prediction when all values are set to their means. Next, for successive variables, their original values are restored. If variables are not independent, what happens almost always is that variables are restored is the order that matters. If two variables in a model are correlated, the first analysed variable will explain its more significant part of the model's variability, while the role of the second variable will remain depreciated. If variables are restored in the opposite order, the influence of those variables will also be different. For that reason, we decided on the core part of SHAP – a game theory concept of Shapely values (Shapely 1953), because this solution is not sensitive to the order of the variables selection. A unique advantage of the SHAP explainer is that it averages the influence across all possible orderings for a given prediction. In this way, the influence of the variable represents the mean change in the model prediction when conditioning on the given variable. An additional standard deviation of the change informs how a given variable is robust against variable ordering.

Finding variable influence

As a result, a vector of new values X' now describes the data and represents the influence of the given feature on the final prediction. The influence is a mapping function, I: $X \rightarrow X'$, where I denotes non-parametric mapping function, called the influence. In that way, X' is a set of mean Shapely numbers that thus describe how variables influence the outcome. X' can be positive, negative or indifferent (Fig. 4). The influence of each variable can be presented in the form of an influence plot, where the *x-axis* contains original values and y the range of the influences. Each case is represented by a single dot at x and y. The I: $X \rightarrow X'$ is in close correspondence with partial dependence plots (PDP) – (Friedman 2001), such that for each specific variable, influence values arrange along the PDP line. If a dependent variable y is standardised, i.e., mean is at 0 and values are represented in units of y standard deviation, both the PDP lines and the influence values acquire

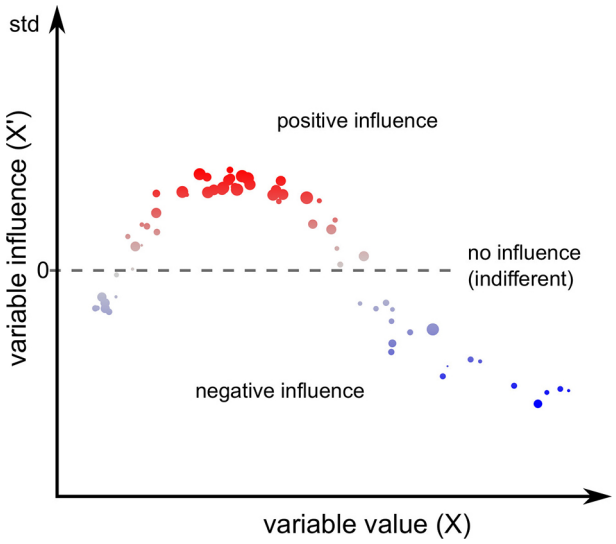

Fig. 4. The concept of mapping variable values into influence. The size of dots simulates the values of the dependent variable.

particular property: all X' values are scaled in the same range, relative to the variable importance. The $I: X \rightarrow X'$ transformation scales the new values to the same scope as y. Thus, 0 means that the given variable does not influence $\hat{y}$ for the given case. Values other than zero, positive or negative, determine the increasing or decreasing $\hat{y}$, in units of P_{tot} standard deviation.

The influence plots (Fig. 4) show the relations between X and X' and detects sections where the influence is positive, negative, or indifferent (no influence). Such an approach extends the notion of variable importance (Jones et al. 2004, Håkanson 2005, Genuer et al. 2010, Leach et al. 2018) and provides new insights into the relations between the studied complex system and the factors that shape it. If the plot identifies the result of clustering, it also allows for identifying sets of cases for which the variable is significant (in the form of positive or negative influence) and cases for which the variable is not. The range of variability of individual variables determines the scale of the impact. The greater the difference in values, the more significant a given variable's role in the explanatory model.

Dissimilarity analysis and clustering objects

The influence plots provide information on two levels. The first is the variable level, and it describes how changes of the variable impact the outcome in the given range of variable values. Moreover, the range of the *y-axis* (influence) is proportional to variable importance. The second is the case (individual object) level– the X'_i is calculated for each case (i.e., lake) separately and describes how each factor with a given value contributes to the value of the $\hat{y}$.

Clustering is a process of searching for similarities between natural objects and separating them into smaller yet consistent groups. Sometimes, preliminary clustering is used to improve the regression models (Kocev et al. 2020). Such clusters, however, only reflect the variability inside the set of explanatory variables; thus, the relationship between independent and dependent variables cannot be inputted into the unsupervised model. On the other hand, regression models by themselves cannot provide unknown information from the training data and cannot cluster the data by discovering their features independently. In that way, the zero-mean and relative to y values of X' are essential for further clustering: the most influencing variables with the highest range of X' contribute most to the dissimilarity between objects, and the impact of the minor influencing variables is weak or negligible. It means that preceding arbitrary variable selection is unwanted, and the clusters will include the distribution of the dependent variable.

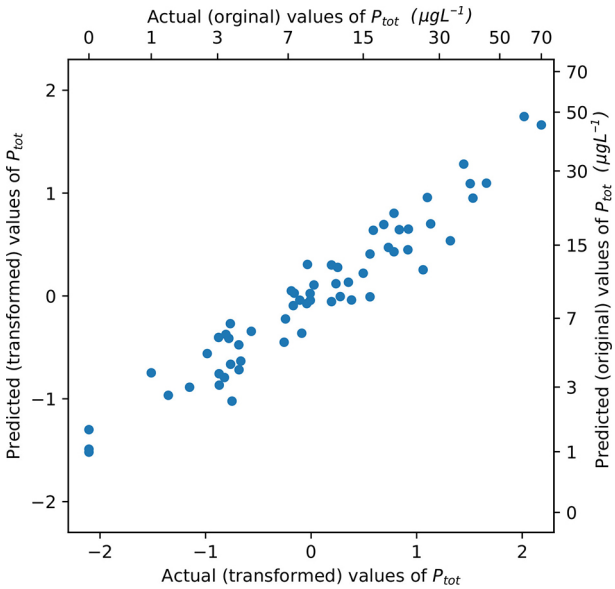

Fig. 5. Relation between actual values of dependent variables and the outcome of the model P_{tot}.

Results

Model quality

Because RF does not require any assumptions about data distribution, that part of the data remained in its original form. The explained P_{tot} variable was transformed into normal-like distribution with Yeo and Johnson's (2000) power transformation. Power transformation stabilises variance and transforms data into Z-score form. Such a transformation is necessary to correctly estimate the influence of individual variables in a uniform unit, i.e., in proportion to the explained variable's standard deviation. Forasmuch the primary goal of the model is an

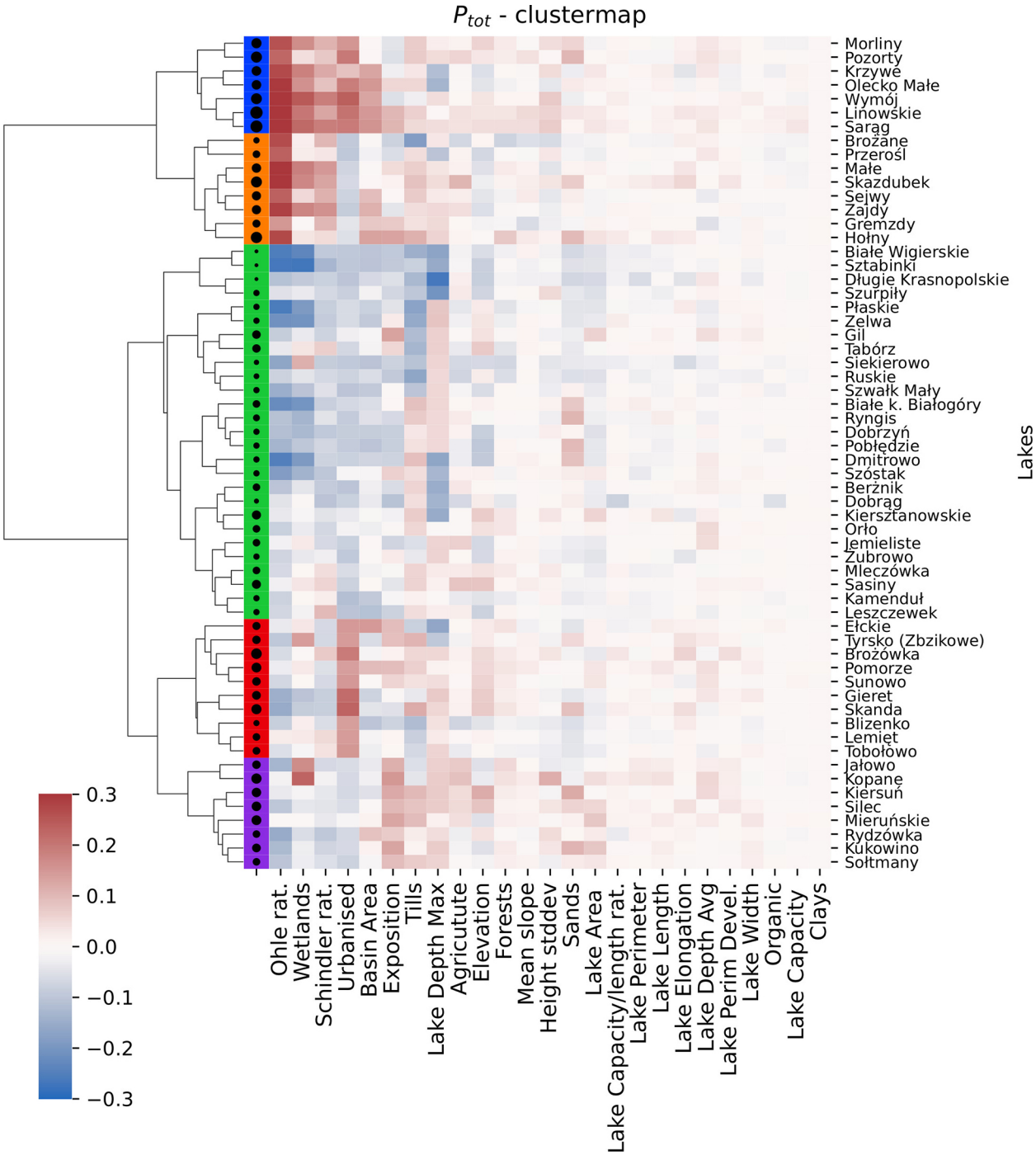

Fig. 6. Relation between clusters and variable influence. The red-white-blue gradient denotes the influence of the given variable. The colours marking the clusters are used in the same way in the other figures.

analysis of existing data set, not a prediction on a new data; the model was trained and tested on the entire data set with specific hyper-parameters: the depth of the trees was reduced to 3, and the number of trees to 50 and the number of cases and variables selected to train each tree was reduced to 0.3. In the result, the root mean square error (RMSE) of the model was higher than for the best set of tuned parameters, but such structure of the learner guarantees that the role of less significant variables will not be omitted. RF is a stochastic algorithm, so we have trained 3000 candidates and selected the best fitted, with the lowest achieved RMSE. The stochastic nature of the RF model causes the results of each execution to differ slightly; nevertheless, the list of the most influential variables remains the same.

The $X' \rightarrow \hat{y}$ is the basis for the reasoning, namely the influence describes $\hat{y}$ not y, so the quality of conclusions is a derivative of the quality of the prediction. This is the main limitation of this method. The relation $y \sim \hat{y}$ depends on the information about y carried by X. If X does not contain key variables for modelling y, the model has low performance and the error of $y \sim \hat{y}$ is the main part of the uncertainty of conclusions. Moreover, such a model only reveals a statistical relationship between the variables and the outcome, which does not yet imply a physical dependency.

Therefore, the first step is to assess the quality of the model. Figure 5 shows the relationship between the actual values of P_{tot} to the values modelled by the RF model. The quality of the model is moderate. The correlation between actual and outcome is very high (R^2 = 0.92), the RMSE value is 0.37 of transformed variable standard deviation. It means that collected variables do not describe approximately one-third of the standard deviation. The remaining part of the P_{tot} variability is most likely the result of processes inside the ecosystems of individual lakes.

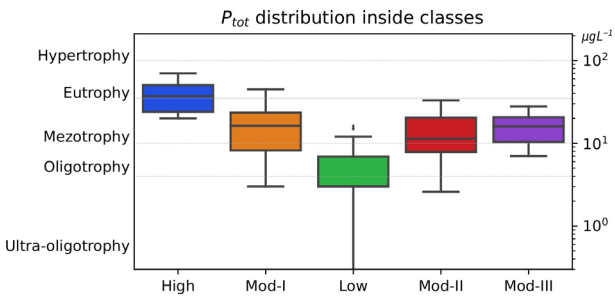

Fig. 7. Variation of P_{tot} in clusters. X-labels denote the name of the class: High trophy, Moderate trophy (Mod.-I, Mod.-II, Mod.-III), and Low trophy.

Figure 6 presents the result of clustering in the form of a cluster map, where $x - axis$ contains all explanatory variables ordered by their importance, e.g. contribution to explainer and $y - axis$ is ordered by clustering results. The dendrogram was cut at 0.85 providing five classes of P_{tot} concentration: one with high values, three with moderate values, and one, the largest, containing lakes with the lowest values (Fig. 6). The cluster containing lakes with the highest values of P_{tot} was labelled as High; three Moderate clusters as Mod.-I, Mod.-II, Mod.-III; and the last as Low. Figure 7 also shows that the quartile range partially refers to the classification of OECD, respectively, such as high to eutrophic-mesotrophic, moderate to mesotrophic, and low to oligotrophic and ultra-oligotrophic, but some lakes do not fit fully to OECD scheme.

The role of individual variables

Figure 8 allows one to analyse $I: X \rightarrow X'$ for each variable and each lake. A partial dependency plot is a tool that allows for tracking the generalised impact of variable values on the outcome, but Shapely numbers show the individual impact of each variable for each case (i.e., lake). $I: X \rightarrow X'$ corresponds to PDP, but these relationships get weaker as the variables' importance decreases, so with the least important variables, the relationship is barely noticeable. The correspondence between variable values and their influence is not always linear, and its dynamics presents new knowledge unavailable with other machine learning (ML) methods but often concordant with intuition and common sense.

The $I: X \rightarrow X'$ plots show that these relationships can be very different. The X' of several variables: Ohle rat. (OR), Schindler rat. (SR), Urbanised (WURB) and Basin Area (WARE) or Lake Depth Max (LDMX) shows a bimodal distribution, with a clear threshold value. Other variables, like Wetlands (WWET), Tills (WTLS), or Agriculture (WAGR) show linear dependency or have an inflection point. The X' of the last ten variables is minimal, within the range

of 0.02–0.05 of P_{tot} standard deviation, and thus their internal structure has a weak interpretative value. Especially for the last four variables, there is no functional relationship between the value of its feature and the effect on the $\hat{y}$. While the linear $I: X \rightarrow X'$ are easy to follow, non-linear relationships require more attention. Although a detailed analysis of this phenomenon is beyond the scope of this paper, it should be noted that the occurrence of threshold values or at least changes in the trend requires the searching for natural processes with similar characteristics. Surface runoff can be such a process, primarily when runoff is caused by short-duration intense storms (Kandel et al. 2004). Also, Guan et al. (2016) noticed the bimodal nature of minor and major rainfall events.

The structure of clusters

We decided to use hierarchical clustering (AHP) with euclidean dissimilarity and Ward linkage since such parameters gave the most interpretative clusters and allowed for easy exploration of possible sub-clusters. The first three steps, i.e.: the transformation of the explanatory model f into explainer g and then the explanatory variables X into the influence of X' is automatic and does not require parameters. Since X' is in the z-score-like form, the scope of each variable is proportional to its role. Thus, X' can be used directly to calculate the dissimilarity between lakes without prior scaling, and dissimilarity matrix d subjects to the clustering process. The selection of a number of classes is an entirely arbitrary decision made after analysing the dendrogram structure.

The clustering process reveals the main novelty of the proposed method. A typical clustering process minimises differences within clusters and maximises differences between them. For this reason, each grouping process requires the prior selection of a limited number of variables, which are preferably not correlated with each other. In our method, the clusters do not mean lakes with similar characteristics, but rather a group of objects where the content of P_{tot} in the water is shaped in a similar way. The latter results follow directly from the fact that the clustering process uses functional dependencies between the lakes features and the content of P_{tot} in the water. As a result, the coherence of the value of the dependent variable is a natural feature of the generated clusters. The most interesting fact is the way each of these classes is explained. The structure clusters show that values in moderate clusters have a similar trophy level but this value results from different processes. It is also noteworthy that arranging the variables according to their importance reveals that the first nine variables have a real influence on P_{tot}, while the last ten has no real impact on this trophy index.

The role of variables in explaining the P_{tot} value can be traced both for classes and individually for each lake. To illustrate the five clusters, the most representative lakes were selected, where the selection criterion was the P_{tot} and the smallest difference between y and $\hat{y}$ for a given cluster. The SHAP plot (Fig. 9) clearly shows the impact of each feature on modelled P_{tot}.

High trophy class

The high value of the P_{tot} index is the effect of the synergy of the five most significant variables: OR, WWET, SR, WURB, and WARE. There is a trace, positive role of LDMX, elevation (ELEV) and Hights Stddev (WHSD), lake exposition (LEXP) and WTLS's negative role. The remaining variables do not affect the trophy index.

Moderate trophy I class

The value of the P_{tot} index in this class results from the synergy of the first three variables: OR, WWET and SR, and the secondary role of LEXP, WTLS, and LDMX. Mainly the WURB, ELEV, and WHSD values have a negative effect. The influence of the other variables is mutually exclusive.

Moderate trophy II class

This class also includes major mesotrophic lakes, but the P_{tot} value results from the positive impact of the U variable and the synergy of less significant variables: LEXP, WTLT, LDMX, ELEV, and Forests (WFRS). The values of OR and SR have a negative effect, but they cannot balance the positive impact of the other variables. In the analysed case of Lake Sunowo, the positive influence of the variables Lake Area (LARE) and Lake Depth Average (LDAV) is also visible, but this is not a feature of the whole class, but only this case.

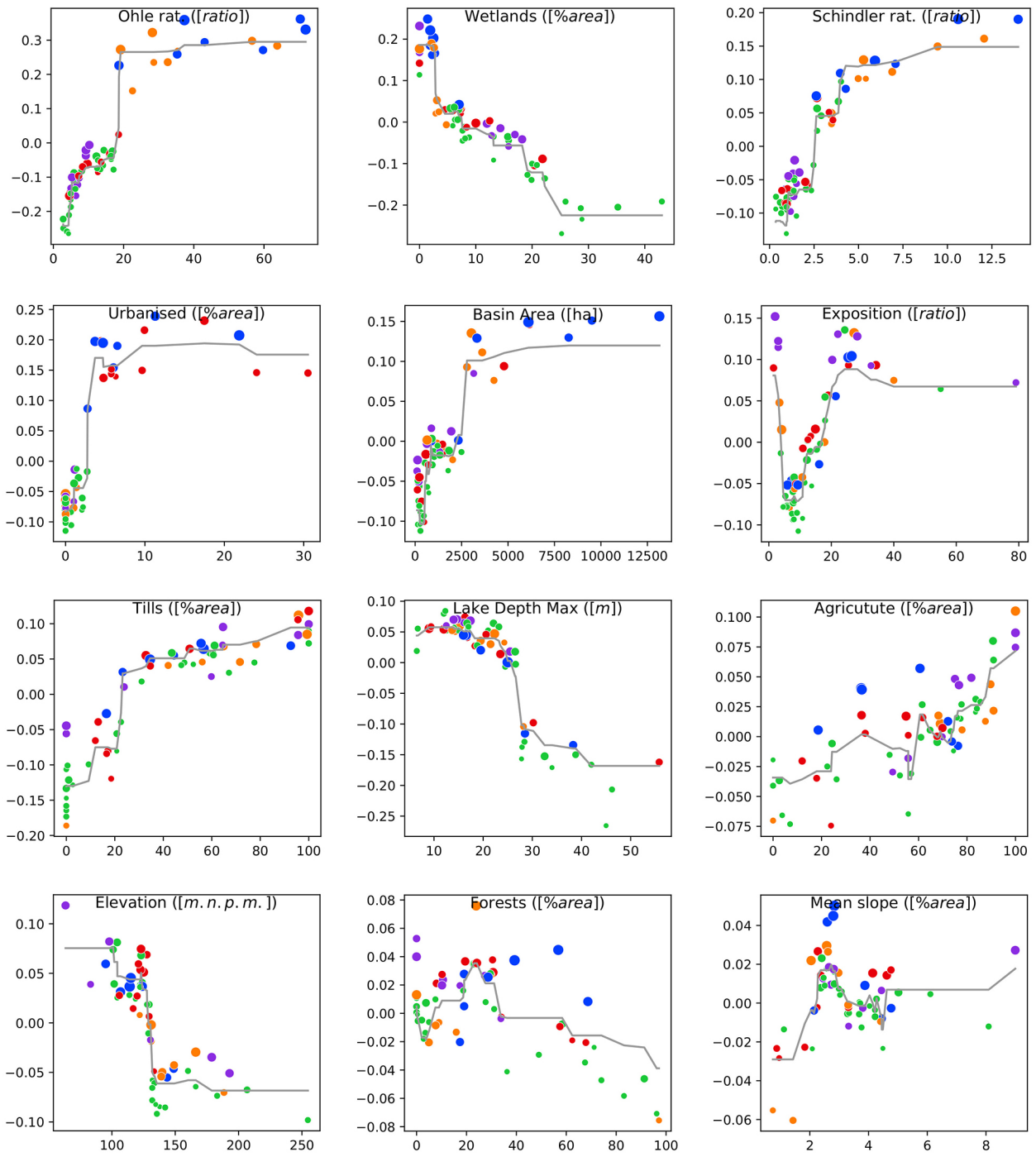

Fig. 8. The influence plots for each variable. The X-axis contains the original values of the variable, Y-axis contains the influence. Colours in legend denote clusters (see Fig. 7), size of dots value of P_{tot}. Partial Dependency Plot (PDP) is marked by a light grey line.

Moderate trophy III class

In the case of this class, the P_{tot} value is the effect of the positive influence of variables with a lower influence: LEXP, WTLS, LDMX, Sands (WSND), and ELEV also have a significant influence on the P_{tot} value in the case of Lake Kiersuń, which is also not a feature of the whole class. The four most influential variables in this class have a negative impact, but the impact is not significant.

Low trophy class

The last class, including lakes with the lowest trophic index values, results from a strong negative impact of the six most significant variables:

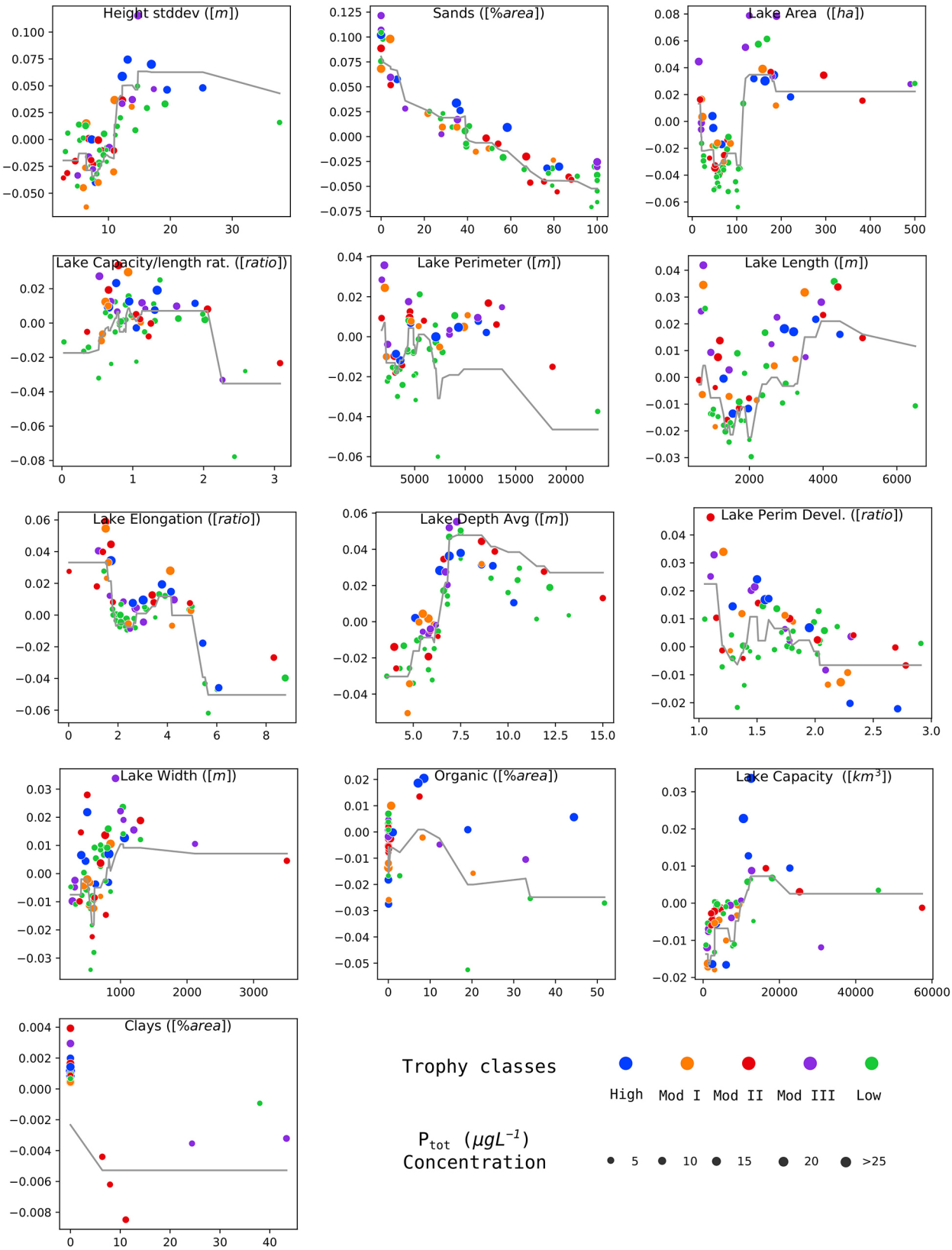
Height stddev ([m])
Sands ([%area])
Lake Area ([ha])
Lake Capacity/length rat. ([ratio])
Lake Perimeter ([m])
Lake Length ([m])
Lake Elongation ([ratio])
Lake Depth Avg ([m])
Lake Perim Devel. ([ratio])
Lake Width ([m])
Organic ([%area])
Lake Capacity ([km³])
Clays ([%area])
Trophy classes
High
Mod I
Mod II
Mod III
Low
P_tot (μgL⁻¹) Concentration
5
10
15
20
>25

Fig. 8. cont.

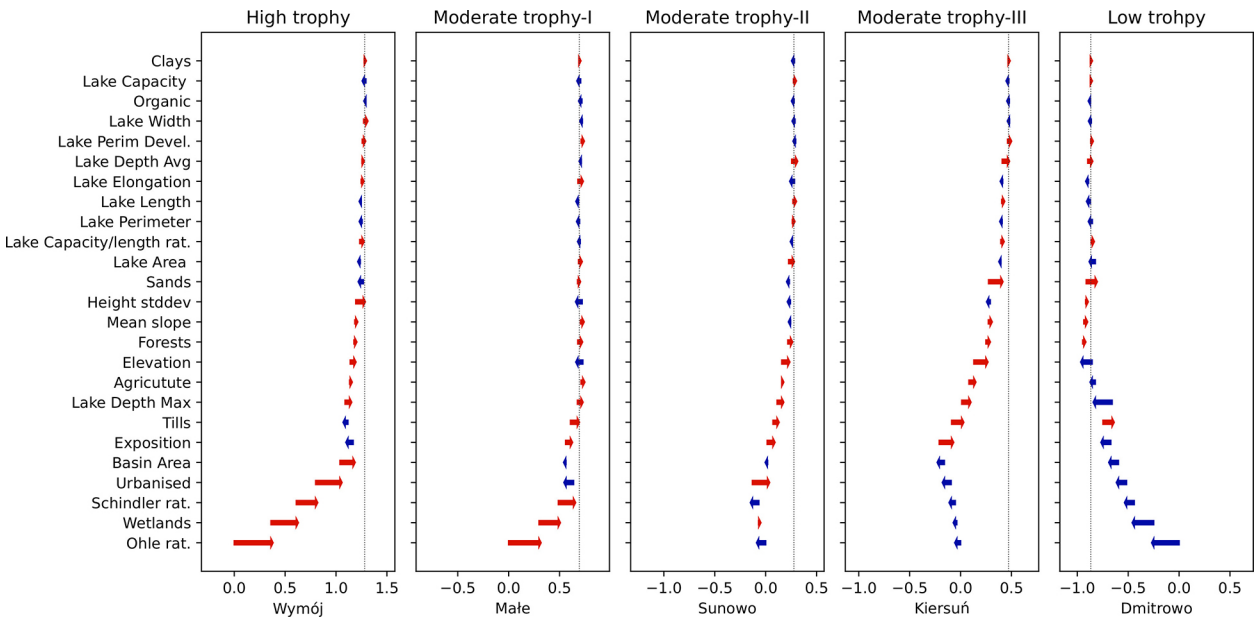

Fig. 9. The SHAP (SHapley Additive exPlanations) plots (Lundberg, Lee 2017) for the five most representative lakes for each class. X-axis presents influence in units of P_{tot} standard deviation scale is the same for each subplot, but ranges are different. Length and direction of arrows denote the scale and influence orientation (negative or positive) brought by a given variable.

OR, WWET, SR, WURB, LARE, and LEXP. The positive impact of WTLS, LDM, and WSND is noted only in selected cases, and in the analysed case of Dmitrowo lakes, the effect of LDMX is negative.

Regardless of the dissimilarity of patterns related to specific values of the P_{tot} index, all cases in each trophy group share features. When the concentration of P_{tot} is high, it results from a positive synergy of the most influential variables; the remaining variables have only a trace effect. Lakes with a moderate concentration of P_{tot} are positively influenced by several variables – different in different groups, and few variables with a negative impact slightly reduce this positive influence. The low concentration of P_{tot} is an effect of the negative synergy of the most important variables, slightly balanced by the influence of less essential features.

Discussion

Variable importance

The proposed method allows for a comprehensive assessment of the influence of selected variables on the share of P_{tot} in water. However, it should be examined how this approach differs from other methods used in lake studies, namely multiple linear regression (LR) (Su et al. 2011, Staehr et al. 2012) and RF (Genuer et al. 2010). Those methods are commonly used to select significant variables (Li et al. 2016, 2017, Bourel, Segura 2018). From existing methods of LRs, we used ElasticNet, a LR model extended by regularisation, a technique that adds a penalty to the model parameters when the model complexity increases. In simple words, when the coefficients are either very high or very low, ElasticNet eliminates those features from the model and shows no relationship between a given explanatory variable and the response variable. The final model is described then, only by those variables that explain the main trends of the model, while variables introducing the noise are eliminated. The RF model is the core of the presented method, so the order of variable importance is identical to our method. The value of importance indicates so-called impurity (or Gini) importance, a normalised total reduction of the error brought by that variable. Thus, if the selected variables significantly reduce the error, but only for selected cases, the role of such a variable may be overestimated.

The RF importance describes only the importance of variables without the information about the direction: whether the variable

generally increases or decreases the modelled value. Compared to RF, the LR model provides more information because it also includes the signs of coefficients. After removing least-informative variables, the sign shows whether the variable increases or decreases the outcome and the coefficient value is somehow related to the intensity of this factor.

Figure 10 shows the evaluation results of the importance of the variables for both models. Values of linear coefficient (slopes) of the LR model and Gini importance of RF cannot be compared by numbers, but both methods indicate more or less the same subset of variables. The value of the coefficient sign follows the orientation of the influence plots in Figure 8. Nevertheless, the variables indicated by LR show rather the strength of the general relationship between a given explanatory variable and the explained variable. Thus, linear models may overestimate the role of variables for which there is a linear relationship (even very weak) with the explained variable at the expense of the variables whose influence is significant but not linear.

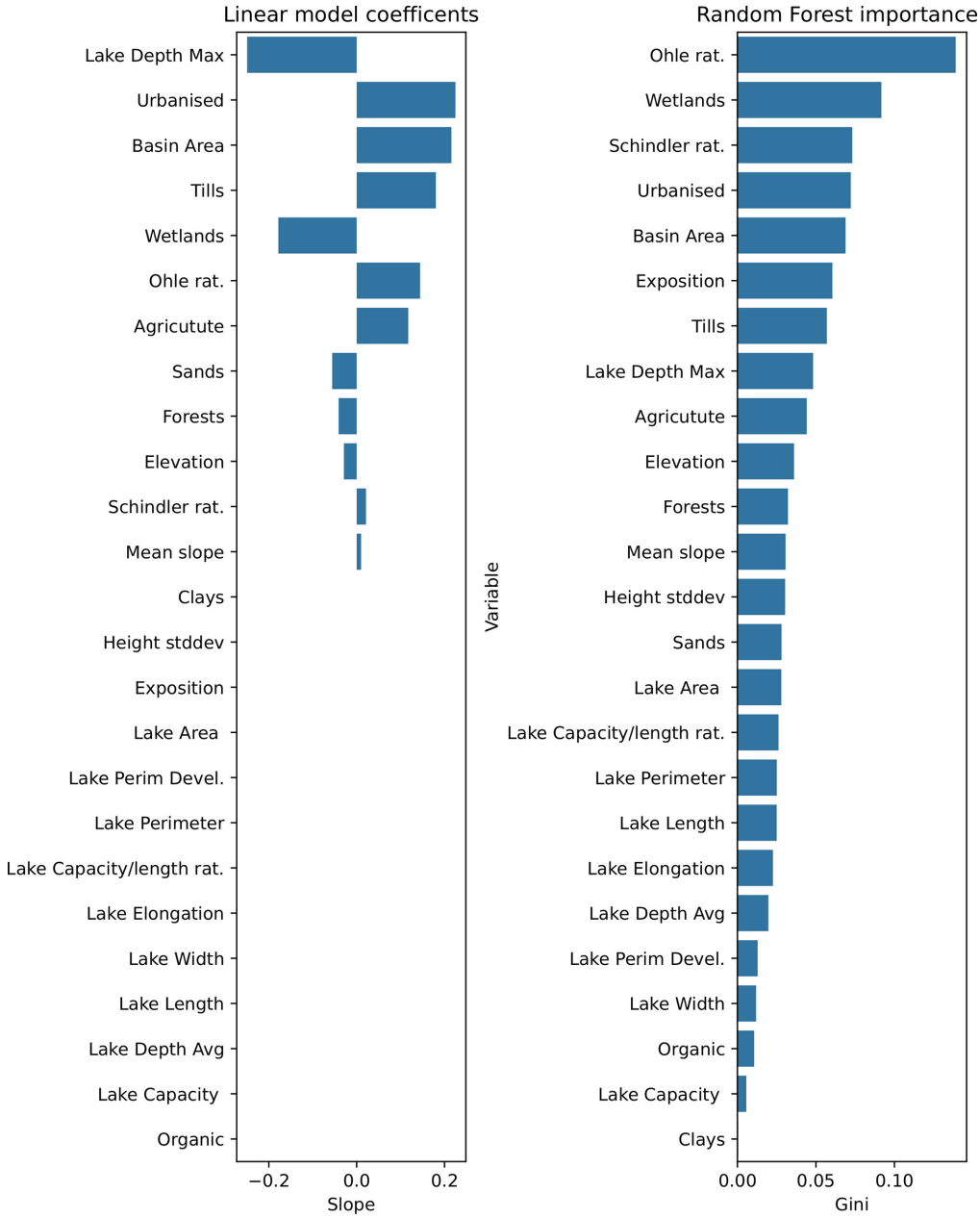

Fig. 10. The variable importance estimated using multiple linear regression (ElasticNet) and random forest. See text for details.

Analysis of dissimilarity inside original and transformed data

The goal of clustering is to identify stable groups in a dataset. The purpose of the comparison is to check whether grouping only the original explanatory variables will allow to identify groups of lakes with a similar trophy. As the clustering technique is not crucial for the entire method, multidimensional scaling (MDS) was used for comparative analysis. The MDS is an ordination technique, a form of non-linear dimensionality reduction that maps distances between objects in original multidimensional spaces into lower-dimensional space positions, preserving original dissimilarities as much as possible (Cox, Cox 2000). The d was calculated for z-scored X variables and in our method d' using X' directly. Due to a large number of variables (high dimensionality), and to avoid the curse of dimensionality,

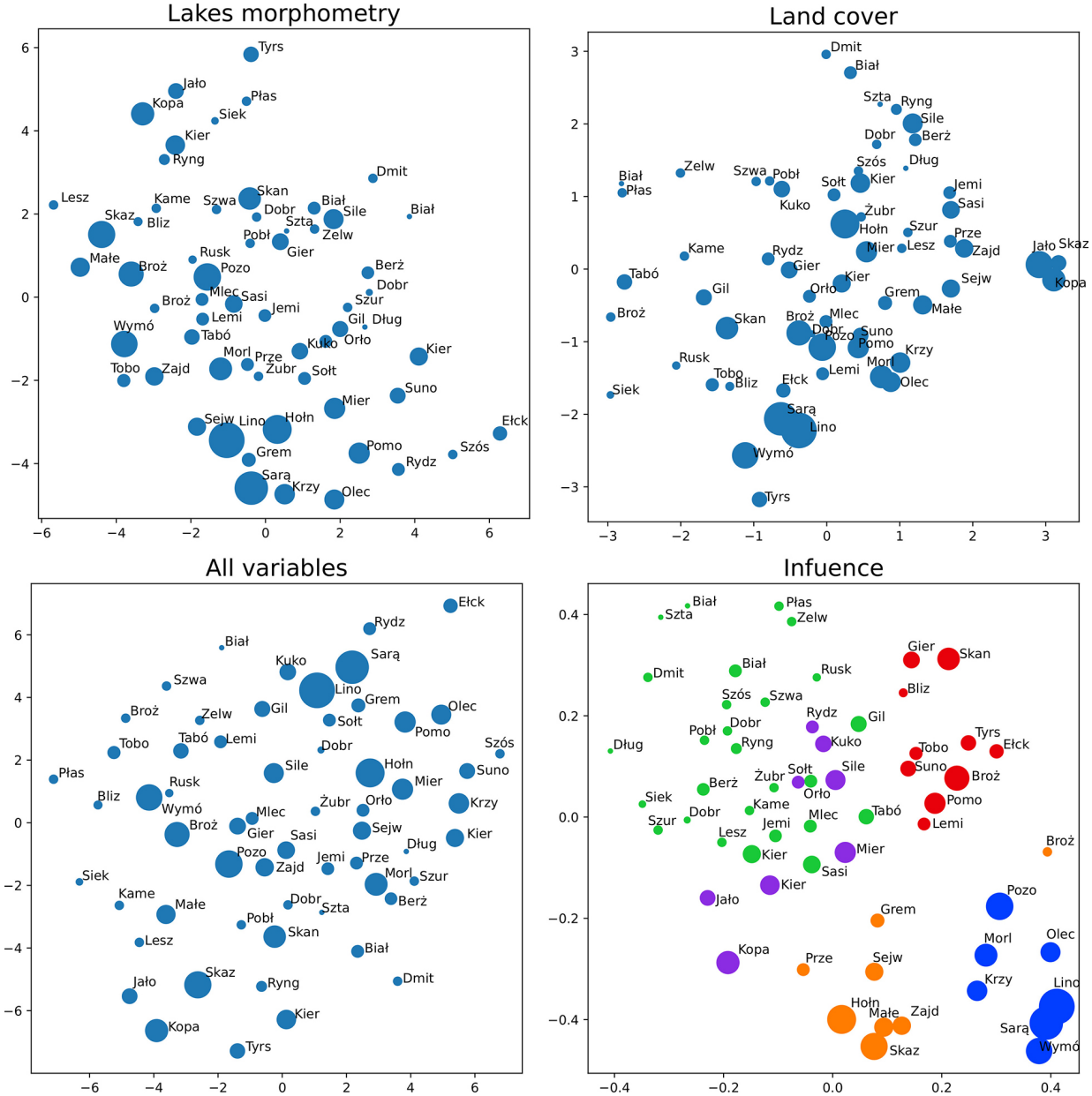

Fig. 11. Visualisation of dissimilarities between lakes in a form of MDS (multidimensional scaling). The axes of the plot have no units. Lakes morphometry presents dissimilarity for a group of morphometric features (see Table 1); Land cover presents dissimilarity for four land cover variables (urbanised, agriculture, forests, wetlands); 'All variables' plate uses all 25 variables; Influence presents dissimilarity between lakes in a space of variables influence. For colours see Figure 7.

we used the L1 ('Cityblock') metric, which is preferable for high dimensional applications to the default L2 ('Euclidean') metric (Aggarwal et al. 2001). Figure 11 shows city block dissimilarity between lakes and the value of the P_{tot} index.

Three cases were analysed: an ordination using all variables and two ordinances for a subset of variables describing land cover and lakes' morphometry. The ordination with all variables shows an amorphous structure and does not reveal P_{tot} variability gradient. The ordinance performed for both subsets also does not show separate clusters, but there were gradients of trophic changes along the horizontal dimension: negative for morphometry and positive for land cover. The appearance of these gradients is related to the role of LDMX and WARE variables in the subgroup of morphometric variables, and WURB and WWET in the land cover group. The ordinance based on d' shows distinct clusters with similar values of P_{tot} indices. Only Low trophy class and Mod.-III class are not separated. This lack of separation is a consequence of the structure of both clusters. In the Mod-III class, the five most important variables have a negative impact on the P_{tot} index as in the Low-trophy class. However, in the latter class, this impact is definitely more substantial.

Conclusions

The paper proposes a new, complex method of data analysis, classified as EML. This method allows for a detailed visual analysis of complex nonlinear regressors and introduces a new concept: variable influence. The latter is a transformation of the set of explanatory variables into a form describing the influence of a given variable on the modelled feature. Such a transformation makes it possible to group data based on the functional relations between the explanatory variables and the explained variable instead of the variation in the explanatory variables only. This is also the limitation of the method because its quality depends on the performance of the model. The method was developed to explain the P_{tot} index for the group of glacial lakes in north-eastern Poland. As part of the analysis, complex nonlinear factors shaping the P_{tot} of individual lakes were detected. On this basis, lakes were grouped into five clusters showing similar values of trophy. In each of the classes, the trophy is the effect of synergy between different groups of factors. The method of LR and RF was compared, and it was shown that it combines the advantages of both – the proposed method allows to precisely determine the real impact of each variable and the relationship between the explanatory and explained variable.

Nonlinear relationships between the variables and the value of the impact are related to nonlinear natural processes – for example, the bimodal distribution of rainfall intensity. However, this problem requires a separate, dedicated research. The cluster analysis showed that the studied lakes could be divided into several clusters, where the P_{tot} value is shaped similarly. This means that there is no single pattern, on how the watershed influences the content of P_{tot} but rather a few repeating patterns representing the studied phenomenon of the trophy value. There are five classes, one for lakes with high and low trophies and three classes with medium trophies, where the P_{tot} index is shaped differently in each. Such a conclusion can have potential value for protecting and managing limnic environments. Although the method has been developed for the problems of lake ecology, its application seems to be more comprehensive and can be applied wherever complex; multivariate numerical models can be used.

Acknowledgments

This work was founded by Polish National Science Center No.2016/23/D/ST10/03071 Czego możemy nauczyć się od wioślarek (Cladocera)? Wykorzystanie zbioru testowego i nowoczesnych metod statystycznych do rekonstrukcji zmian środowiska.

Authors' contribution

JJ: Conceptualisation, Data curation, Investigation, Formal analysis, Methodology, Software, Validation, Writing – original draft, Writing – review and editing, Visualization. IZ: Conceptualisation, Data curation, Investigation, Writing – review and editing, Project administration, Funding acquisition. MR: Data curation, Investigation, Writing – review and editing. MW: Investigation; Writing – review and editing.

References

Aggarwal C.C., Hinneburg A., Keim D.A., 2001. On the surprising behavior of distance metrics in high dimensional space. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*: 420–434. DOI 10.1007/3-540-44503-x_27.

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* 39: 916–924. DOI 10.1002/clen.201100050.

Apolinarska K., Pleskot K., Pełechata A., Migdałek M., Siepak M., Pełechaty M., 2020. The recent deposition of laminated sediments in highly eutrophic Lake Kierskie, Western Poland: 1 year pilot study of limnological monitoring and sediment traps. *Journal of Paleolimnology* 63: 283–304. DOI 10.1007/s10933-020-00116-2.

Bajkiewicz-Grabowska E., 2020. Geoecosystems of Polish Lakes. In: Korzeniewska E., Harnisz M. (eds), *Polish River Basins and Lakes – Part I. The handbook of environmental chemistry*, vol. 86. Springer, Cham. DOI 10.1007/978-3-030-12123-5_3.

Beaulieu, M., Pick, F., Palmer, M., Watson, S., Winter, J., Zurawell, R., Gregory-Eaves, I., 2014. Comparing predictive cyanobacterial models from temperate regions. *Canadian Journal of Fisheries and Aquatic Sciences* 71: 1830–1839. DOI 10.1139/CJFAS-2014-0168/SUPPL_FILE/CJFAS-2014-0168SUPPLC.PDF.

Benedini M., Tsakiris G., 2013. *Water quality modelling for rivers and streams*. Springer, p 233. DOI 10.1007/978-94-007-5509-3.

Biecek P., 2018. DALEX: explainers for complex predictive models in r. *The Journal of Machine Learning Research* 19: 3245–3249.

Borics G., Nagy L., Miron S., Grigorszky I., László-Nagy Z., Lukács B.A., G-Tóth L., Várbíró G., 2013. Which factors affect phytoplankton biomass in shallow eutrophic lakes? *Hydrobiologia* 714: 93–104. DOI 10.1007/S10750-013-1525-6/FIGURES/3.

Bourel M., Segura A.M., 2018. Multiclass classification methods in ecology. *Ecological Indicators* 85: 1012–1021. DOI 10.1016/J.ECOLIND.2017.11.031.

Breiman L., 2001. Random forests. *Machine Learning* 45: 5–32. DOI 10.1023/A:1010933404324.

Chen V., Li J., Kim J.S., Plumb G., Talwalkar A., 2021. Interpretable machine learning. *Queue* 19: 28–56. DOI 10.1145/3511299.

Cox T., Cox M., 2000. *Multidimensional scaling. 2nd edition*. Chapman and Hall/CRC, p 328. DOI 10.1201/9780367801700.

Cui H., Ou Y., Wang L., Wu H., Yan B., Han L., Li Y., 2019. Identification of environmental factors controlling phosphorus fractions and mobility in restored wetlands by multivariate statistics. *Environmental Science and Pollution Research* 26: 16014–16025. DOI 10.1007/s11356-019-05028-x.

Dafforn K.A., Johnston E.L., Ferguson A., Humphrey C., Monk W., Nichols S.J., Simpson S.L., Tulbure M.G., Baird D.J., 2015. Big data opportunities and challenges for assessing multiple stressors across scales in aquatic ecosystems. *Marine and Freshwater Research* 67: 393–413. DOI 10.1071/MF15108.

Dormann C.F., Elith J., Bacher S., Buchmann C., Carl G., Carré G., Marquéz J.R., Gruber B., Lafourcade B., Leitão P.J., Münkemüller T., Mcclean C., Osborne P.E., Reineking B., Schröder B., Skidmore A.K., Zurell D., Lautenbach S., 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography* 36: 27–46. DOI 10.1111/J.1600-0587.2012.07348.X.

EEA 2018. Corine land cover (CLC) 2018, version 2020-20u1. Online: https://land.copernicus.eu/pan-european/corine-land-cover/clc2018 (accessed: XXX).

Eliasz-Kowalska M., Wojtal A.Z., 2020. Limnological characteristics and diatom dominants in lakes of Northeastern Poland. *Diversity* 12: 1–16. DOI 10.3390/d12100374.

Friedman J.H., 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29: 1189–1232.

Froeschke J.T., Froeschke B.F., 2011. Spatio-temporal predictive model based on environmental factors for juvenile spotted seatrout in Texas estuaries using boosted regression trees. *Fisheries Research* 111: 131–138. DOI 10.1016/j.fishres.2011.07.008.

Gebler D., Kolada A., Pasztaleniec A., Szoszkiewicz K., 2021. Modelling of ecological status of Polish lakes using deep learning techniques. *Environmental Science and Pollution Research* 28: 5383–5397. DOI 10.1007/s11356-020-10731-1.

Genuer R., Poggi J.M., Tuleau-Malot C., 2010. Variable selection using random forests. *Pattern Recognition Letters* 31: 2225–2236. DOI 10.1016/j.patrec.2010.03.014.

Goggin M.L., 1986. The "Too Few Cases/Too Many Variables" problem in implementation research. *The Western Political Quarterly* 39: 328. DOI 10.2307/448302.

Gorgoglione A., Gregorio J., Ríos A., Alonso J., Chreties C., Fossati M., 2020. Influence of land use/land cover on surface-water quality of Santa Lucia River, Uruguay. *Sustainability (Switzerland)* 12. DOI 10.3390/su12114692.

Guan M., Sillanpää N., Koivusalo H., 2016. Storm runoff response to rainfall pattern, magnitude and urbanization in a developing urban catchment. *Hydrological Processes* 30: 543–557. DOI 10.1002/HYP.10624.

Håkanson L., 2005. The importance of lake morphometry and catchment characteristics in limnology – Ranking based on statistical analyses. *Hydrobiologia* 541: 117–137. DOI 10.1007/s10750-004-5032-7.

Harrell F.E., 2015. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. Springer, New York, p 582. DOI 10.1007/978-3-319-19425-7.

Hernández-Almeida I., Grosjean M., Gómez-Navarro J.J., Larocque-Tobler I., Bonk A., Enters D., Ustrzycka A., Pi otrowska N., Przybylak R., Wacnik A., Witak M., Tylmann W., 2017. Resilience, rapid transitions and regime shifts: Fingerprinting the responses of Lake Zabińskie (NE Poland) to climate variability and human disturbance since AD 1000. *The Holocene* 27: 258–270. DOI 10.1177/0959683616658529.

Hollister J.W., Milstead W.B., Kreakie B.J., 2016. Modeling lake trophic state: A random forest approach. *Ecosphere* 7: 1–14. DOI 10.1002/ecs2.1321.

Huang J., Gao J., Zhang Y., 2015. Combination of artificial neural network and clustering techniques for predicting phyto plankton biomass of Lake Poyang, China. *Limnology* 16: 179–191. DOI 10.1007/S10201-015-0454-7/TABLES/5.

Jańczak J., 1999. *The Atlas of Polish Lakes, vol. 3 Masurian Lakes and the Southern Part of Poland*. Bogucki Wydawnictwo Naukowe, Poznań.

Jasiewicz J., Metz M., 2011. A new GRASS GIS toolkit for Hortonian analysis of drainage networks. *Computers and Geosciences* 37: 1162–1173. DOI 10.1016/j.cageo.2011.03.003.

Jasiewicz J., Niedzielski P., Krueger M., Hildebrandt-Radke I., Michałowski A., 2021. Elemental variability of prehistoric ceramics from postglacial lowlands and its implications for emerging of pottery traditions – an example from the pre-roman iron age. *Journal of Archaeological Science: Reports* 39: 103177.

Jones J.R., Knowlton M.F., Obrecht D.V., Cook E.A., 2004. Importance of landscape variables and morphology on nutrients in Missouri reservoirs. *Canadian Journal of Fisheries and Aquatic Sciences* 61: 1503–1512. DOI 10.1139/F04-088.

Jones K.B., Neale A.C., Nash M.S., Van Remortel R.D., Wickham J.D., Riitters K.H., O'Neill R.V., 2001. Predicting nutrient and sediment loadings to streams from landscape metrics: A multiple watershed study from the United States Mid-Atlantic Region. *Landscape Ecology* 16: 301–312. DOI 10.1023/A:1011175013278.

Kandel D.D., Western A.W., Grayson R.B., Turral H.N., 2004. Process parameterization and temporal scaling in surface runoff and erosion modelling. *Hydrological Processes* 18: 1423–1446. DOI 10.1002/HYP.1421.

Kallf J., 2001. *Limnology: inland water ecosystems*. Prentice Hall, New Jersey, p 592.

Kocev D., Ceci M., Stepišnik T., 2020. Ensembles of extremely randomized predictive clustering trees for predicting structured outputs. *Machine Learning* 109: 2213–2241. DOI 10.1007/S10994-020-05894-4/FIGURES/14.

Kondracki J., 2009. *Geografia regionalna Polski*. Wydanie trzecie, Wydawnictwo Naukowe PWN, Kraków.

Lange W., 1986. *Fizyczno-limnologiczne uwarunkowania tolerancji systemów jeziornych Pomorza*. Zeszyty Naukowe UG Rozprawy i monografie nr 79, Gdańsk, 3–177.

Leach T.H., Beisner B.E., Carey C.C., Pernica P., Rose K.C., Huot Y., Brentrup J.A., Domaizon I., Grossart H.P., Ibelings B.W., Jacquet S., Kelly P.T., Rusak J.A., Stockwell J.D., Straile D., Verburg P., 2018. Patterns and drivers of deep chlorophyll maxima structure in 100 lakes: The relative importance of light and thermal stratification. *Limnology and Oceanography* 63: 628–646. DOI 10.1002/lno.10656.

Li B., Yang G., Wan R., Dai X., Zhang Y., 2016. Comparison of random forests and other statistical methods for the prediction of Lake water level: A case study of the Poyang Lake in China. *Hydrology Research* 47: 69–83. DOI 10.2166/nh.2016.264.

Li B., Yang G., Wan R., Hörmann G., Huang J., Fohrer N., Zhang L., 2017. Combining multivariate statistical techniques and random forests model to assess and diagnose the trophic status of Poyang Lake in China. *Ecological Indicators* 83: 74–83. DOI 10.1016/j.ecolind.2017.07.033.

Li T., Li S., Liang C., Bush R.T., Xiong L., Jiang Y., 2018. A comparative assessment of Australia's Lower Lakes water quality under extreme drought and post-drought conditions using multivariate statistical techniques. *Journal of Cleaner Production* 190: 1–11. DOI 10.1016/j.jclepro.2018.04.121.

Li W., Zhang Y., Cui L., Zhang M., Wang Y., 2015. Modeling total phosphorus removal in an aquatic environment restoring horizontal subsurface flow constructed wetland based on artificial neural networks. *Environmental Science and Pollution Research* 22: 12347–12354. DOI 10.1007/S11356-015-4527-2/TABLES/2.

Lundberg S.M., Lee S.I., 2017. A unified approach to interpreting model predictions. arXiv, 1–10. Online: https://github.com/slundberg/shap (accessed ??.??.????).

Marks L., 2012. Timing of the Late Vistulian (Weichselian) glacial phases in Poland. *Quaternary Science Reviews* 44: 81–88. DOI 10.1016/j.quascirev.2010.08.008.

Marks L., Ber A., Gogo Lek, W., Piotrowska K., 2006. *Geological map of Poland 1:500000*. Państwowy Instytut Geologiczny, Warszawa.

Molnar C., Casalicchio G., Bischl B., 2020. Interpretable machine learning – a brief history, state-of-the-art and challenges. In: *Hands-on machine learning with R*, 417–431. DOI 10.1007/978-3-030-65965-3_28.

Morawski W., 2005. Warmińska prowincja paleogeograficzna plejstocenu (północno-wschodnia Polska). *Przeglad Geologiczny* 53: 477–488.

Ohle W., 1956. Bioactivity, production, and energy utilization of lakes. *Limnology and Oceanography* 1: 139–149. DOI 10.4319/lo.1956.1.3.0139.

Pochocka-Szwarc K., 2013. Some aspects of the last glaciation in the Mazury Lake District (north-eastern Poland). *Acta Palaeobotanica* 53: 3–8. DOI 10.2478/acpa-2013-0001.

Ribeiro M.T., Singh S., Guestrin C., 2016. Why Should I Trust You? In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, NY, USA. 1135–1144. DOI 10.1145/2939672.2939778.

Rocha J.C., Peres C.K., Buzzo J.L.L., de Souza V., Krause E.A., Bispo P.C., Frei F., Costa L.S., Branco C.C., 2017. Modeling the species richness and abundance of lotic macroalgae based on habitat characteristics by artificial neural networks: a potentially useful tool for stream biomonitoring programs. *Journal of Applied Phycology* 29: 2145–2153. DOI 10.1007/s10811-017-1107-5.

Rodhe W., 1969. Crystallization of eutrophication concepts in northern Europe. In: *Eutrophication: causes, consequences, correctives*. National Academy of Sciences, Washington: 50–64.

Schindler D.W., 1977. Evolution of phosphorus limitation in lakes. *Science* 195: 260–262. DOI 10.1126/science.195.4275.260.

Shapely L.S., 1953. A value of n-person games. In: Kuhn H., Tucker A. (eds.) *Contribution to the theory of games* II. Princeton University, Princeton, 307–317.

Shrikumar A., Greenside P., Kundaje A., 2017. Learning important features through propagating activation differences. In: *34th International Conference on Machine Learning*, ICML 2017, 4844–4866. arXiv:1704.02685.

Simeonov V., Simeonova P., Tsakovski S., Lovchinov V., 2010. Lake water monitoring data assessment by multivariate statistics. *Journal of Water Resource and Protection* 2: 353–361. DOI 10.4236/jwarp.2010.24041.

Staehr P.A., Baastrup-Spohr L., Sand-Jensen K., Stedmon C., 2012. Lake metabolism scales with lake morphometry and catchment conditions. *Aquatic Sciences* 74: 155–169. DOI 10.1007/s00027-011-0207-6.

Su S., Li D., Zhang Q., Xiao R., Huang F., Wu J., 2011. Temporal trend and source apportionment of water pollution in different functional zones of Qiantang River, China. *Water Research* 45: 1781–1795. DOI 10.1016/J.WATRES.2010.11.030.

Sun A.Y., Scanlon B.R., 2019. How can big data and machine learning benefit environment and water management: A survey of methods, applications, and future directions. *Environmental Research Letters* 14(7): 073001. DOI 10.1088/1748-9326/ab1b7d.

Tandyrak R., Grochowska J., Parszuto K., Augustyniak R., Łopata M., 2020. Environmental conditions in polish

lakes with different types of catchments. In: Korzeniewska E., Harnisz M. (eds), *Polish River Basins and Lakes – Part I. The handbook of environmental chemistry*, vol 86. Springer, Cham. 119–138.

Tylmann W., Szpakowska K., Ohlendorf C., Woszczyk M., Zolitschka B., 2012. Conditions for deposition of annually laminated sediments in small meromictic lakes: a case study of Lake Suminko (northern Poland). *Journal of Paleolimnology* 47: 55–70. DOI 10.1007/s10933-011-9548-3.

Yeo I.N., Johnson R.A., 2000. A new family of power transformations to improve normality or symmetry. *Biometrika* 87: 954–959. DOI 10.1093/biomet/87.4.954.

Weckwerth P., Wysota W., Piotrowski J.A., Adamczyk A., Krawiec A., Dąbrowski M., 2019. Late Weichselian glacier outburst floods in North-Eastern Poland: landform evidence and palaeohydraulic significance. *Earth-Science Reviews* 194: 216–233. DOI 10.1016/j.earscirev.2019.05.006.