Author(s): Ärje, Johanna; Kärkkäinen, Salme; Meissner, Kristian; Iosifidis, Alexandros; Ince, Türker; Gabbouj, Moncef; Kiranyaz, Serkan

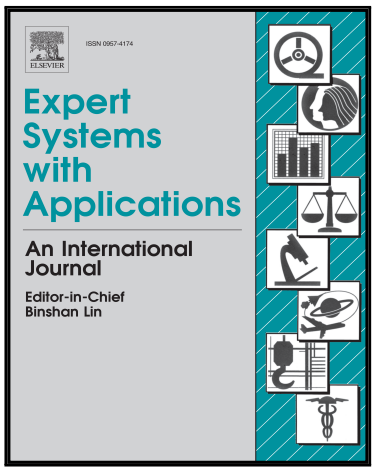
Title: The effect of automated taxa identification errors on biological indices

Year: 2017

Version:

Please cite the original version:

Ärje, J., Kärkkäinen, S., Meissner, K., Iosifidis, A., Ince, T., Gabbouj, M., & Kiranyaz, S. (2017). The effect of automated taxa identification errors on biological indices. Expert Systems with Applications, 72, 108-120. https://doi.org/10.1016/j.eswa.2016.12.015



The effect of automated taxa identification errors on biological indices

Johanna Ärje, Salme Kärkkäinen, Kristian Meissner,
Alexandros Iosifidis, Türker Ince, Moncef Gabbouj, Serkan Kiranyaz

PII: S0957-4174(16)30689-3
DOI: 10.1016/j.eswa.2016.12.015
Reference: ESWA 11024

To appear in: *Expert Systems With Applications*

Received date: 20 May 2016
Revised date: 8 December 2016
Accepted date: 9 December 2016

Please cite this article as: Johanna Ärje, Salme Kärkkäinen, Kristian Meissner, Alexandros Iosifidis, Türker Ince, Moncef Gabbouj, Serkan Kiranyaz, The effect of automated taxa identification errors on biological indices, *Expert Systems With Applications* (2016), doi: 10.1016/j.eswa.2016.12.015

Highlights

- We study the effects of classification errors from 11 different classifiers on 14 biological indices.
- We assess the reasons why certain indices are more sensitive to classification errors.
- We compare the classifiers based on their ultimate use in the biomonitoring of macroinvertebrates.

The effect of automated taxa identification errors on biological indices

Johanna Ärje[a], Salme Kärkkäinen[a], Kristian Meissner[b], Alexandros Iosifidis[c], Türker Ince[d], Moncef Gabbouj[c] and Serkan Kiranyaz[e]

December 13, 2016

[a] Department of Mathematics and Statistics, University of Jyvaskyla, Finland.
P.O. Box 35 (MaD), 40014 University of Jyvaskyla, Finland
email: johanna.arje@jyu.fi
salme.karkkainen@jyu.fi

[b] Freshwater Centre, Finnish Environment Institute, SYKE, Jyväskylä Office, Jyväskylä, Finland.
kristian.meissner@ymparisto.fi

[c] Department of Signal Processing, Tampere University of Technology, Finland.
alexandros.iosifidis@tut.fi, moncef.gabbouj@tut.fi

[d] Electrical & Electronics Engineering Department, Izmir University of Economics Turkey.
turker.ince@izmirekonomi.edu.tr

[e] Electrical Engineering, College of Engineering, Qatar University, Qatar.
serkan.kiranyaz@tut.fi

Abstract

In benthic macroinvertebrate biomonitoring systems, the target is to determine the status of ecosystems based on several biological indices. To increase cost-efficiency,

computer-based taxa identification for image data has recently been developed. Taxa identification errors can, however, have strong effects on the indices and thus on the determination of the ecological status. In order to shift the biomonitoring process towards automated expert systems, we need a clear understanding on the bias caused by automation. In this paper, we examine eleven classification methods in the case of macroinvertebrate image data and show how their classification errors propagate into different biological indices. We evaluate 14 richness, diversity, dominance and similarity indices commonly used in biomonitoring. Besides the error rate of the classification method, we discuss the potential effect of different types of identification errors. Finally, we provide recommendations on indices that are least affected by the automatic identification errors and could be used in automated biomonitoring.

Keywords: Biomonitoring; Classification error; Diversity: Error propagation; Identification; Similarity.

1 Introduction

In biomonitoring, reliable taxa identification is an important prerequisite for subsequent index calculation. Diversity, richness, dominance and similarity indices are often used in aquatic biomonitoring to determine the status of waterbodies (e.g. Birk *et al.*, 2012). In order to calculate indices, samples of biological indicator groups such as benthic macroinvertebrates are collected and the individuals in the samples are identified to taxa. However, when taxa identification errors are made, these errors may affect the statistical properties of the estimated indices. This can result in incorrect ecological status predictions that can further propagate into unnecessary mitigation measures or even prevent needed mitigation measures (Haase *et al.*, 2010).

The ever decreasing research and monitoring funding calls for new and more efficient ways of monitoring and sample processing. To improve the cost-efficiency of the monitoring process, it needs to be automated. This can be achieved by building an expert system that automatically identifies samples to taxa, calculates biological indices based on their abundances and provides the user with an assessment of the ecological status of the sampled waterbody. Before the process can be automated, we need a clear understanding on the possible bias involved with automation. However, the common approach followed when designing an expert system involves the selection of its building blocks based on an absolute performance metric. For the case of the classification block, this metric is usually the absolute classification rate on a pre-defined test set. While this approach, indeed, leads to a good selection in the cases where a classifier is clearly superior compared to its competition, it might lead to a bad selection without taking into account the bias introduced by the tested data (Ali *et al.*, 2017).

To address the high costs of the identification step in benthic macroinvertebrate biomonitoring, researchers have explored e.g. citizen-science monitoring (Dickinson *et al.*, 2012) and automated identification methods (e.g. Blaschko *et al.*, 2005; Culverhouse *et al.*, 2006; Lytle *et al.*, 2010; Kiranyaz *et al.*, 2011; Ärje *et al.*, 2013; Joutsijoki *et al.*, 2014). However, such approaches may introduce additional bias and variation into biological indices calculated from samples due to identification errors. Indeed, Gardiner *et al.* (2012) noted that

misidentification in citizen-science monitoring results in overestimation of species richness (Magurran, 2004) and Simpson's diversity (Simpson, 1949).

The goal of this study is to empirically investigate the statistical properties of biological indices when the automated identification of individuals contains misidentifications. Similar studies have been done in remote sensing for landscape pattern indices (Chen *et al.*, 2010; Shao *et al.*, 2001; Wickham *et al.*, 1997), e.g. mean patch size, total edge and contagion index. Shao *et al.* (2001) included Shannon's and Simpson's diversity indices in their study but concentrated on variation caused by classification errors rather than bias. Scardi *et al.* (2008) studied an expert system based on multimetric indices of fish assemblages but did not take into account possible identification errors on the fish species. Understanding the bias and extra variation caused by identification errors is a prerequisite step in shifting towards automated biomonitoring and ecological status assessments. Recent studies in the field of expert systems have highlighted the importance of measuring not only absolute accuracy but also quality. Biswas and Biswas (2017); Piltan and Sowlati (2016) and Carayannis *et al.* (2016) have developed multi-metric criteria to measure performance based on quality, while Abdar *et al.* (2017) have used sensitivy, specificity and other statistics of the confusion matrix to assess the quality of classifiers. In remote sensing, Chen *et al.* (2010) have used the bias caused in index values due to classification errors as a measure of quality. However, to our knowledge, there exist no comprehensive studies on the quality of classification as performance measure in automated biomonitoring.

We consider several commonly derived biological indices i) describing richness, i.e. species richness (Magurran, 2004), Margalef's diversity (Clifford and Stephenson, 1975) and Chao's estimator of the absolute number of species in an assemblage (Chao, 1984), ii) describing diversity, i.e. Shannon index (Shannon and Weaver, 1963) and Simpson's index (Simpson, 1949), iii) describing evenness and dominance, i.e. Shannon evenness (Pielou, 1969, 1975), Simpson's evenness (Smith and Wilson, 1996) and Berger-Parker index (Berger and Parker, 1970), and iv) describing similarity of two assemblages, i.e. Sørensen index (Sørensen, 1948), percent model affinity index (Renkonen, 1938; Novak and Bode, 1992), Canberra metric (Lance and Williams, 1967), Euclidian similarity (Clifford and Stephenson, 1975), Morisita-Horn index (Horn, 1966) and Jaccard similarity (Jaccard, 1901). The similarity indices compare the similarity of species distributions in two conditions, e.g. reference and monitored conditions in aquatic systems. Richness, diversity and dominance indices are calculated for a single species distribution, i.e. for a monitored sample.

In the current work, we are especially interested in estimating the error propagation of indices that use computer-based taxa identification from image data. In automated identification, the task is to classify n images of individuals belonging to c classes using features extracted from the images (e.g. width, height, mean grey value, etc). Various classification methods can be used in automated identification (see e.g. Hastie *et al.*, 2009; Duda *et al.*, 2001). However in all approaches, the classifiers are trained with a training data of known identity (i.e. the gold standard). Subsequently, optimal parameter values are selected based on classification error of a validation data and the final error rate is evaluated with an independent test data set. Often, the best classifier is the one having lowest error rate. Besides error rate, we can also estimate a confusion matrix which provides the probabilities of different correct and incorrect classifications. When considering the estimation of the indices, the confusion matrix is of great interest as its properties affect the amount of bias and variation

propagated. We perform a simulation study to showcase the effects of different types of confusion matrices on error propagation. We acknowledge that there are other sources of bias but in this paper we focus on bias due to classification errors.

Using a benthic macroinvertebrate image data, we illustrate the effect of classification errors on biological indices. We use eleven classifiers: Random Bayes Array (RBA, Ärje *et al.*, 2013), Support Vector Machines (SVM, KSVM, Cortes and Vapnik, 1995), Random Forest (RF, Breiman, 2001), Linear Discriminant Analysis (LDA, Hastie *et al.*, 2009), Radial Basis Function Network (RBFN, Haykin, 2009; Kiranyaz *et al.*, 2011), Multilayer Perceptron (MLP, Haykin, 2009; Kiranyaz *et al.*, 2009), Reference Discriminant analysis + nearest neighbor (KRDA, Iosifidis *et al.*, 2014a), Graph Embedded Extreme Learning Machine (GEELM, Iosifidis *et al.*, 2015), Graph Embedded Kernel Extreme Learning Machine (GEKELM, Iosifidis *et al.*, 2014b) and Naïve Bayes (NB, Hastie *et al.*, 2009). Some of these methods have been evaluated with the same image data in Ärje *et al.* (2013) with small changes. However, the target of the current work is to compare the statistical properties of estimated indices using the results of these eleven classifiers. In the comparisons, we use simulation-based results. Finally, we provide some recommendations on which of the indices are least biased by classification errors and could thus be used in automated biomonitoring.

2 On biological indices and their properties

In this section, we first describe the set-up for data collection, second, the considered indices with respect to the given set-up and third, the modified set-up in the case of misclassification is outlined.

2.1 The set-up

Mathematically, let $\{\omega_1, \ldots, \omega_c\}$ be the finite set of c classes such that p_h is the probability of class ω_h in a monitored situation and q_h is the probability of class ω_h in a reference situation. For simplicity, we assume that a random sample of counts $\mathbf{X} = (X_1, \ldots, X_c)$ is drawn from a multinomial distribution $M(n, \mathbf{p})$, where n is sample size and $\mathbf{p} = (p_1, \ldots, p_c)$ the probabilities of interest. Then, the natural estimator of p_h is $\hat{p}_h = X_h/n$, a maximum likelihood estimator for $h = 1, \ldots, c$. Similarly, the random sample $\mathbf{Y} = (Y_1, \ldots, Y_c)$ of size m is drawn from a multinomial distribution $M(m, \mathbf{q})$, where a natural estimator for the values of $\mathbf{q} = (q_1, \ldots, q_c)$ is $\hat{q}_h = Y_h/m$.

Below, we present the indices (Table 1), give the references for the statistical properties, if known, and further outline some practical details. The ranges of the indices are used in the comparison of index behavior in Section 4.

We tested three richness indices: 1) species richness (S, Magurran, 2004), 2) Chao's estimator of the absolute number of species in an assemblage (S_{Chao}, Chao, 1984) and 3) Margalef's diversity (D_{Mg}, Margalef, 1958; Clifford and Stephenson, 1975). Smith and Grassel (1977) studied the theoretical mean and variance of S. Using those results, the same properties of D_{Mg} could easily be derived. Chao (1987) derived variance for $S_{Chao} = S + F_1^2/2F_2$. Due to cases $F_2 = 0$, we use instead the formula in Table 1 by Magurran and McGill (2010).

We also study the effect of classification errors on two diversity indices: 4) Shannon's index (H', Shannon and Weaver, 1963) and 5) Simpson's index (D_x, Simpson, 1949). Tong (1983) presents some distributional properties for H' assuming multinomial distribution. Paninski (2003) studies nonparametric estimation of H' and gives an overview on its bias and variance.

Further, we study three evenness/dominance indices in our analyses: 6) Shannon evenness (J', Pielou, 1969, 1975), 7) Simpson's evenness ($E_{1/D}$, Smith and Wilson, 1996) and 8) Berger-Parker index (d, Berger and Parker, 1970). J' is a scaled version of H' that measures evenness instead of diversity.

We study the effect of classification errors on six similarity indices: 9) Sørensen similarity (QS, Sørensen, 1948), 10) percent model affinity index (PMA, Renkonen, 1938; Novak and Bode, 1992), 11) Canberra metric ($1 - CM$, Lance and Williams, 1967), 12) Euclidian similarity ($1 - D^2_{Eucl}$, Clifford and Stephenson, 1975), 13) Morisita-Horn index (C_λ, Horn, 1966) and 14) Jaccard similarity coefficient (J, Jaccard, 1901). Theoretical properties of the PMA in the case of multinomial distribution are presented in Ärje *et al.* (2016) and the references therein. For the calculation of $1 - CM$, classes with zero abundancies in both samples are left out. Janson and Vegelius (1981) studied the asymptotical standard error of J. Further, Albatineh and Niewiadomska-Bugaj (2011) discovered the expectation for corrected form of the index. C_λ has a maximum value not equal to one but 'about one' (Horn, 1966).

To our knowlegde, the properties of the other diversity, evenness, dominance and similarity indices have only been studied with simulation experiments (e.g. Magurran, 2004; Smith, 2002).

2.2 The effect of classification errors on indices

The classification of objects performed by either human or machine may include errors which affect the values of indices calculated from classified samples. Let us formulate the set-up as proposed by Healy (1981) and Fortier (1992). The confusion matrix A of a specified classification procedure is assumed to be known. Its element $a_{hh'}$ is the probability of classifying an object into the class h when originating from the class h'. Further, $\sum_h a_{hh'} = 1$ and $a_{hh'} \geq 0, h, h' = 1, \ldots, c$. Then, the probability of an object to be classified to the class h is

$$\tilde{p}_h = \sum_{h'=1}^{c} a_{hh'} * p'_h.$$

The interesting consequence is that the allocated counts $\tilde{X}_1, \ldots, \tilde{X}_c$ of size n are drawn from a multinomial distribution $M(n, \tilde{\mathbf{p}})$ instead of $M(n, \mathbf{p})$, respectively $\tilde{\mathbf{Y}} \sim M(m, \tilde{\mathbf{q}})$. As the distribution of the allocated counts is biased, the identification errors may propagate into the expected values of the indices causing bias in the index values.

In this paper, we do not comment on the properties of the indices *per se* but restrict our analyses to study the error propagation into the indices due to classification errors as follows. Using a general notation of index I with correct classification and index $\tilde{I}$ with incorrect classification, we concentrate on the proportional bias defined as follows

$$\%bias = \frac{E(\tilde{I}) - E(I)}{|\max I - \min I|}, \tag{1}$$

where the expectations are Monte Carlo estimates. Similar proportional bias has been used by Chen *et al.* (2010) to study error propagation in remote sensing. The %bias provides a measure of the biological significance of the bias and enables us to compare the bias in different biological indices over a range of taxa distributions. Similarly, we study the effect of classification errors on the variation of the biological indices as follows

$$\%sd = \frac{sd(\tilde{I}) - sd(I)}{|\max I - \min I|}, \tag{2}$$

where the standard deviations are Monte Carlo estimates.

Table 1: Biological indices used for analyses and their ranges.

Index	Formula	min	max
Richness			
1) Species richness	$S_x = \sum_{h=1}^{c} I(X_h > 0)$	0	c
2) Chao's estimator	$S_{Chao,x} = S_x + \frac{F_{1,x}(F_{1,x}-1)}{2(F_{2,x}+1)}$, where $F_{1,x} = \sum_{h=1}^{c} I(X_h = 1)$ and $F_{2,x} = \sum_{h=1}^{c} I(X_h = 2)$	0	$(c^2 - c + 2)/2$ if $n > c$ $(c^2 + c)/2$ if $n = c$
3) Margalef's diversity	$D_{Mg,x} = \frac{S_x - 1}{\log n}$	0	$(c-1)/\log n$
Diversity			
4) Shannon index	$H'_x = -\sum_{h=1}^{c} \hat{p}_h \log \hat{p}_h$	0	$\log c$
5) Simpson's index	$D_x = \sum_{h=1}^{c} \hat{p}_h^2$	$1/c$	1
Evenness/dominance			
6) Shannon evenness	$J'_x = H'_x / \log S_x$	0	1
7) Simpson's evenness	$E_{1/D,x} = \frac{1/D_x}{S_x}$	0	1
8) Berger-Parker index	$d_x = \max(\mathbf{X})/n$	$1/c$	1
Similarity			
9) Sørensen similarity	$QS = \frac{2S_{xy}}{S_x + S_y}$, where $S_{xy} = \sum_{h=1}^{c} I(X_h > 0 \wedge Y_h > 0)$	0	1
10) Percent model affinity index	$PMA = 1 - \frac{1}{2}\sum_{h=1}^{c} \lvert\hat{p}_h - \hat{q}_h\rvert$	0	1
11) Canberra metric	$1 - CM = 1 - \frac{1}{S_x + S_y - S_{xy}} \sum_{h=1}^{c} \frac{\lvert X_h - Y_h\rvert}{(X_h + Y_h)}$	0	1
12) Euclidian similarity	$1 - D^2_{Eucl} = 1 - \sum_{h=1}^{c} (\hat{p}_h - \hat{q}_h)^2$	-1	1
13) Morisita-Horn index	$C_\lambda = \frac{2\sum_{h=1}^{c} X_h Y_h}{(D_x + D_y) n m}$	0	1
14) Jaccard similarity coefficient	$J = \frac{S_{xy}}{S_x + S_y - S_{xy}}$	0	1

3 Materials and methods

3.1 Data

To study the effects of identification errors on biological indices, we use two datasets. The first data is a benthic macroinvertebrate image data set with 6814 individual images of 33 lotic taxa and two lentic gastropod taxa. Lotic specimens were collected during research projects of the Finnish Environment Institute and the national freshwater biomonitoring program in Finland, whereas lentic specimens were collected by the department of Biological and Environmental Sciences at the university of Jyväskylä. The taxonomic identities of the specimens were verified by three taxonomic experts and are considered to be true (i.e. form the gold standard). The macroinvertebrates were batch imaged onto a computer one taxa at a time using VueScan(c) software (http://www.hamrick.com/, Phoenix, Arizona, USA) with an HP Scanjet4850 flatbed scanner at an optical resolution of 2400 d.p.i. The images were normalized to the same intensity range and color balance. The specimens were segmented from these batches to their individual images and from each image, a total of 64 geometric and color scale features were extracted. The feature extraction was done with ImageJ (Rasband, 1997-2010). Detailed information on the features and taxa used can be found in Ärje *et al.* (2013).

The second data set is abundance data of benthic macroinvertebrates gathered during the national freshwater biomonitoring program 2006–2013 in Finland. The monitoring program includes a total of total 12 stream types (small, medium and large or extra large peatland and woodland streams for northern and southern Finland separately). For details, see Aroviita *et al.* (2012). For each stream type, there are reference streams that are considered to be in near natural condition unaltered by human-induced stressors and non-reference streams considered to be impacted by human actions. The second data set comprises a total of 149 taxa. We restrict our analysis to taxa that are present in both data sets and combine some taxa into groups to obtain equal taxa lists (i.e. 32 taxa) for both the image data and the monitoring data. The taxa list and info of combined taxa can be found in the appendix (Table 7).

3.2 Classification

We first use the image data for taxa identification, i.e. classification. The data is divided 10 times into training (33,33 %), validation (33,33 %) and testing (33,33 %) sets. Each classifier is first trained with the training data and the validation data is utilized to find the optimal parameter values. Then, training and validation data are combined and used to train the classifier with the chosen parameter values. Finally, we evaluate the classifier with the test data. This procedure is repeated 10 times, once with each data split. The error rate of a classifier is then calculated as the average classification error from these 10 repetitions. Similarly, we obtain the confusion matrix of a classifier as the average from the 10 repetitions.

We explore the effects of misclassifications with eleven different classifiers: Naïve Bayes (NB, Hastie *et al.*, 2009), Linear Discriminant Analysis (LDA, Hastie *et al.*, 2009), Random Forest (RF, Breiman, 2001), Random Bayes Array (RBA, Ärje *et al.*, 2013), Support Vec-

tor Machines (SVM, KSVM, Cortes and Vapnik, 1995), Reference Discriminant analysis + nearest neighbor (KRDA, Iosifidis *et al.*, 2014a), Graph Embedded Extreme Learning Machine (GEELM, Iosifidis *et al.*, 2015), Graph Embedded Kernel Extreme Learning Machine (GEKELM, Iosifidis *et al.*, 2014b), Multilayer Perceptron (MLP, Haykin, 2009; Kiranyaz *et al.*, 2009) and Radial Basis Function Network(RBFN, Haykin, 2009; Kiranyaz *et al.*, 2011). Some of the classifiers are known to perform poorly with the macroinvertebrate image data (Ärje *et al.*, 2013) but are included as examples to fully explore to gradient of error propagation.

NB and LDA are Bayesian classifiers (e.g. Hastie *et al.*, 2009) that both assume that features are normally distributed and which classify observations according to the highest posterior probability. LDA assumes that all classes have a common covariance matrix whereas NB that features are independent from each other.

RF (Breiman, 2001) is a collection of random decision trees. For each tree, the classifier takes a bootstrap sample of the training data. For each node in a tree, RF randomly selects a subset of M features and chooses the one that best separates the data based on entropy. RF builds k trees and uses voting to get the final class predictions for the test data.

RBA (Ärje *et al.*, 2013) is an implementation of RF for quadratic discriminant analysis (QDA) which is a generalization of LDA that allows arbitrary covariance matrices. RBA forms a collection of random QDAs. Each QDA classifier is trained using a bootstrap sample of the training data and M randomly selected features. RBA consists of k random QDAs. It uses either voting, posterior weighted voting, averaged posterior probabilities, or highest average rank to determine the final class predictions of the test data. RBA can also be used to evaluate the importance of the features, which can thereon be used as weights when sampling the features for each random QDA. Here we used averaged posterior probabilities to make the final class decision.

SVM (Cortes and Vapnik, 1995) is a non-probabilistic binary classifier that determines the hyperplane separating the two classes with maximal margin. Non-linear decision functions are obtained by exploiting the kernel trick, which inherently maps the input data to a feature space of high dimensions. The determination of the optimal hyperplane separating the two classes in this high-dimensional feature space corresponds to the determination of a non-linear decision function in the input space. Multi-class classification is obtained by combining multiple binary classifiers. In this paper we employ the One-Versus-Rest combination scheme. KSVM is an extension of SVM that uses a radial basis function kernel.

KRDA (Iosifidis *et al.*, 2014a) is an extension of Kernel Discriminant Analysis (KDA) that tries to overcome the assumption of the latter concerning the optimal representation of each class. KDA employs the class mean for class representation, assuming that the classes in the feature space are unimodal and follow Gaussian distributions. However, since these two assumptions are usually not valid in many real world problems, class representation by the class mean is suboptimal. KRDA overcomes this problem by determining both the optimal class representation and data projection.

GEELM (Iosifidis *et al.*, 2015) is an algorithm for Single-hidden Layer Feedforward Neural (SLFN) network training that exploits geometric data relationships. GEELM first nonlinearly maps the data from the input space to a high-dimensional feature space based on random weights. Then a regularized regression problem is solved. The regularization term in this process is designed in order to exploit geometric data (or class) relationships encoded

in an intrinsic and a penalty graph. In our experiments we employed the graphs used in Local Fisher Discriminant Analysis (LFDA Sugiyama, 2007, ,).

GEKELM is a kernel extension of GEELM. The main idea of GEKELM is that the networkâĂŹs hidden layer can be formed by a very large (even infinite) number of neurons. In this case, the ELM network is similar to an infinite neural network in which the training data similarities are encoded in a kernel matrix (Iosifidis *et al.*, 2014b). GEKELM trains such a network by also exploiting geometric data (or class) relationships encoded in an intrinsic and a penalty graph. For GEKELM we also employ the LFDA graphs.

MLPs (Haykin, 2009; Kiranyaz *et al.*, 2009) are feed-forward, fully-connected Artificial Neural Networks (ANNs), which can be described as directed graphs where each node is performing some activation function to its inputs and forwarding the result to the input of other neurons in the adjacent layer. MLPs may contain one or more layers of hidden neurons. In this work, for all experiments, a conventional back-propagation training rule with a global adaptation of the learning rate (with initial value of 0.001) is used and both shallow (single hidden layer of 32 neurons) and deep (two hidden layers of 64 and 32 neurons respectively) MLP configurations are considered.

RBFN (Haykin, 2009; Kiranyaz *et al.*, 2011) is another well-known feed-forward, fully-connected ANN type which can approximate any solution space or function as a sum of N RBFs (such as Gaussian functions) in a single hidden layer. For training of RBFN, given the specified maximum number of hidden neurons and the spread parameter of each Gaussian neuron, for each epoch a hidden layer neuron is added to minimize training Mean-Squared Error (MSE) below specified target level. For each data partition, the spread parameter is chosen to minimize the validation data classification error. Both shallow (64 hidden neurons) and deep (384 hidden neurons) RBFN configurations are considered.

3.3 Simulation study

We study the effect of classification errors on the richness, diversity, evenness and dominance indices in each of the 12 river types for both, reference and non-reference compositions, resulting in a total of 24 different taxa distributions. We use the reference and non-reference streams as the two conditions being compared with the similarity indices. In biomonitoring, the reference condition is often considered to be a known (i.e. fixed) target distribution. Therefore, we study the error propagation of the similarity indices in two cases. In the first case, the reference sample is assumed to be known, i.e. correctly identified by several human experts, and the non-reference sample is classified using the aforementioned classifiers. In the second case, both samples are classified using the classifiers and may contain classification errors.

First, we draw 1000 samples from multinomial distributions, $\mathbf{X} \sim M(n, \mathbf{p})$ for non-reference streams and respectively, $\mathbf{Y} \sim M(n, \mathbf{q})$ for reference streams. The taxa distributions $\mathbf{p}$ and $\mathbf{q}$ are weighted averages over one river type's non-reference and reference stream monitoring samples using sample sizes as weights. We calculate the values of all richness, diversity, evenness, dominance and similarity indices, denoted by I. As a result, we obtain an empirical distribution of each index I, called the correct distribution below. Second, we draw 1000 samples from multinomial distributions $\tilde{\mathbf{X}} \sim M(n, \tilde{\mathbf{p}})$ and $\tilde{\mathbf{Y}} \sim M(n, \tilde{\mathbf{q}})$, where $\tilde{\mathbf{p}} = A\mathbf{p}$, $\tilde{\mathbf{q}} = A\mathbf{q}$ and A is the average confusion matrix of a classifier. Using the allocated counts,

we calculate the values of each index, denoted by $\tilde{I}$, and the obtained empirical distribution is called the allocated distribution of the index I. Finally, we compare the distributions of the correct and allocated index values to see how the different indices are affected by misclassifications. In this work, we restrict sample sizes to $n = m = \{200, 500, 1000\}$.

4 Results

Considering solely classification error, the best classifier is GEKELM and the worst MLP (see Table 2). However, we are more interested in the end result, i.e. how index values affecting decision making are biased due to classification errors. Below we discuss the results summarized over all river types, i.e. 24 different taxa distributions for the richness, diversity, evenness and dominance indices and 12 different taxa distribution pairs for the similarity indices. As an example, Fig. 1, 2 and 3 show the results for the most common river type of the monitoring data, medium-sized non-reference peatland streams in southern Finland. All following tables are ordered based on the classification errors of Table 2.

To evaluate the severity of error propagation to biological indices, we concentrate on the proportional bias in Eq. 1. Table 3 shows the average proportional bias for the diversity, richness, evenness and dominance indices over all river types, i.e. 24 different species distributions. As the sign of the bias can be different among the classifiers even in one river type and different for the same classifier in different river types, in Table 3 the average is taken over absolute proportional bias. With our parameters ($c = 32, n = 500$), S_{Chao} has a very high maximum value which is reached if there is one large class with the majority of observations and all other classes have a single observation in them. As this is a highly unlikely scenario, we calculate the %bias in S_{Chao} proportional to the range of S, which is c, instead of $|\max S_{Chao} - \min S_{Chao}|$.

From Table 3, it is evident that richness indeces 1)-3) S, S_{Chao} and D_{Mg} are sensitive to classification errors. For these indices, even the best classifiers result in approximately 20 %bias. All three indices are based on presence/absence data and linked to the number of species, which may well be the cause of their sensitivity. This is due to the fact that even one misclassified observation can bring a new taxa into the calculation and cause overestimation in the number of taxa. This conclusion is also supported by Fig. 1 as the allocated index value distributions for S, S_{Chao} and D_{Mg} are biased upwards for all classifiers.

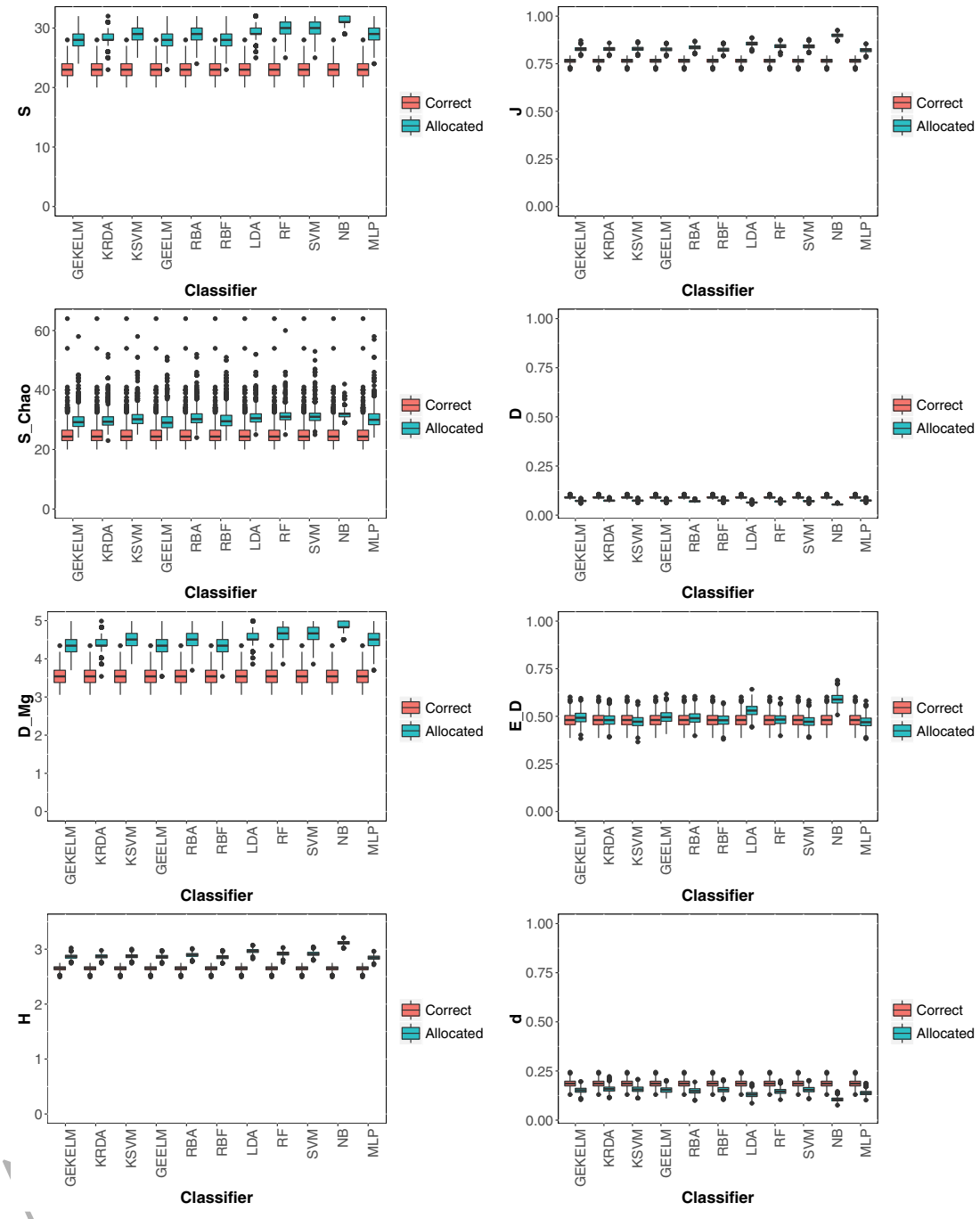


Figure 1: The effect of classification errors on richness, diversity, evenness and dominance indices for medium-sized non-reference peatland streams in southern Finland. Here, $\mathbf{X} \sim M(500, \mathbf{p})$. The red boxplots represent the correct index value distributions. The blue boxplots represent the allocated index value distributions.

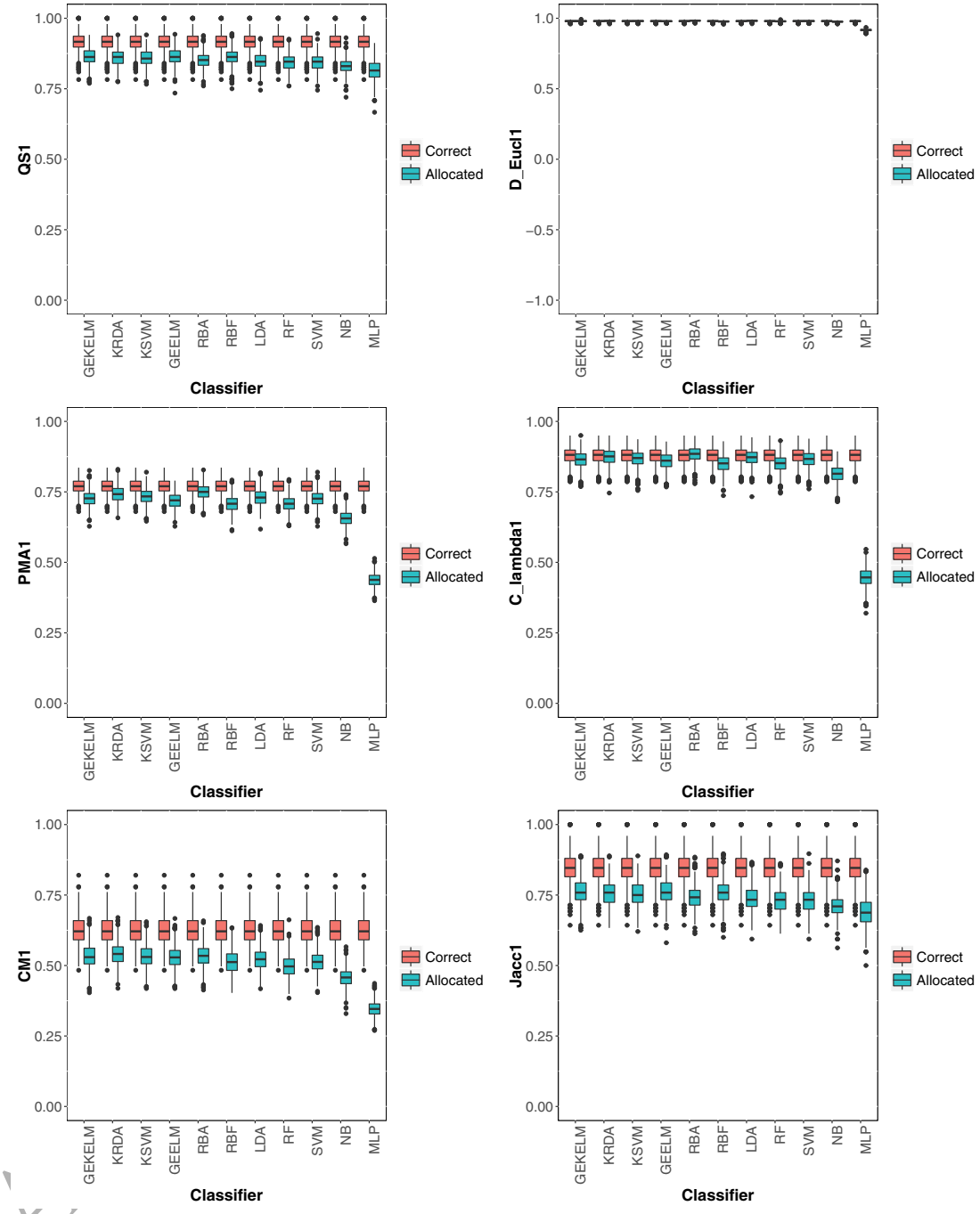


Figure 2: The effect of classification errors on similarity indices for medium-sized non-reference peatland streams in southern Finland when the reference sample is assumed to be known. Here, $\mathbf{X} \sim M(500, \mathbf{p})$ and $\mathbf{Y} \sim M(500, \mathbf{q})$. The red boxplots represent the correct index value distributions. The blue boxplots represent the allocated index value distributions.

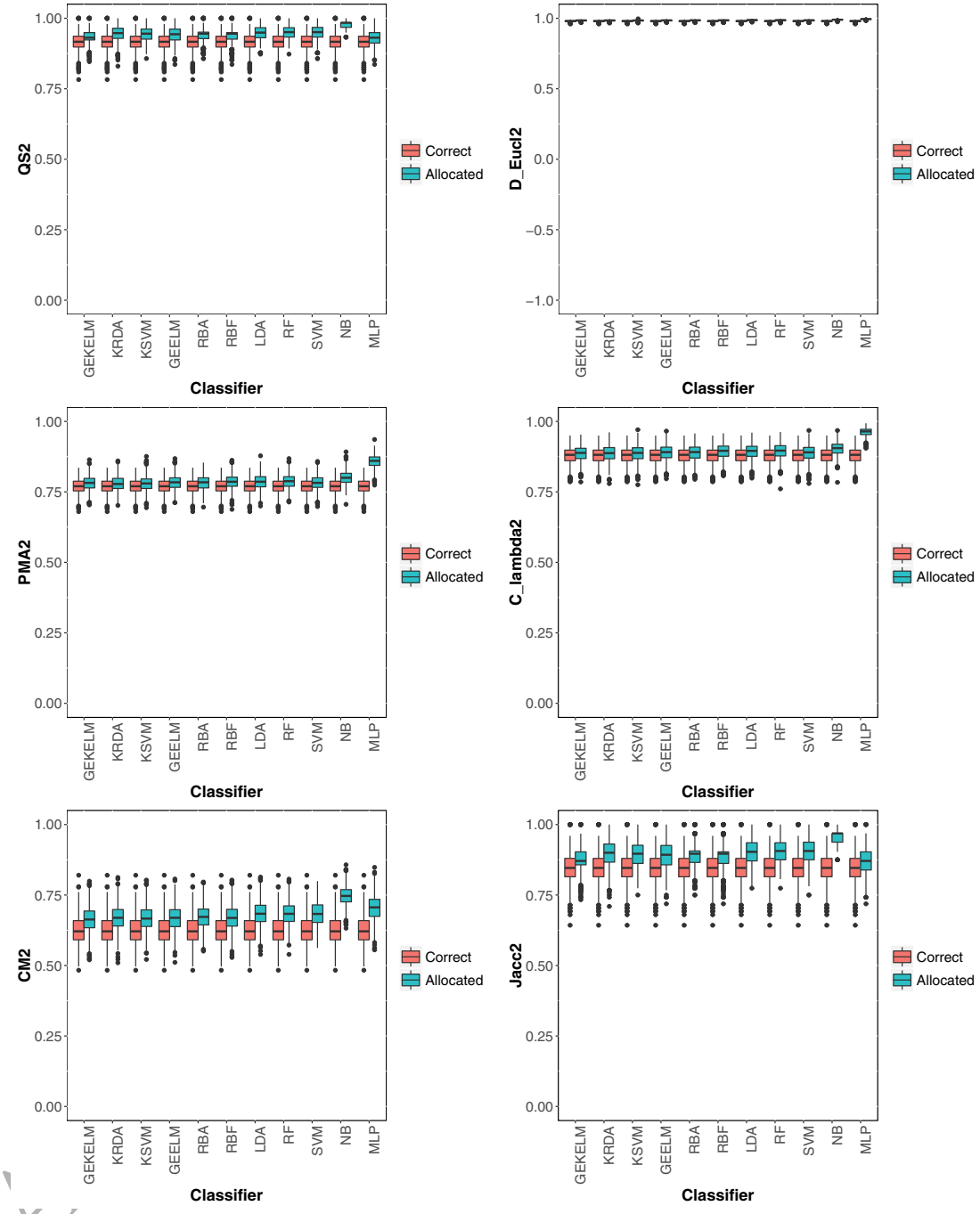


Figure 3: The effect of classification errors on similarity indices for medium-sized non-reference peatland streams in southern Finland when both samples may contain classification errors. Here, $\mathbf{X} \sim M(500, \mathbf{p})$ and $\mathbf{Y} \sim M(500, \mathbf{q})$. The red boxplots represent the correct index value distributions. The blue boxplots represent the allocated index value distributions.

Table 2: Classification errors using 66,6/33,3 split for training and test data. The classification errors are averages from 10 runs, standard deviation for classification errors is presented in parenthesis.

	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
CE	0.159 (0.006)	0.161 (0.008)	0.167 (0.006)	0.169 (0.007)	0.190 (0.008)	0.194 (0.007)	0.229 (0.008)	0.240 (0.008)	0.245 (0.007)	0.514 (0.009)	0.892 (0.015)

Table 3: Average proportional bias for diversity, richness, evenness and dominance indices with sample size $n = 500$. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
S	0.18 (0.06)	0.18 (0.06)	0.22 (0.07)	0.17 (0.06)	0.20 (0.06)	0.19 (0.07)	0.23 (0.07)	0.23 (0.08)	0.27 (0.07)	0.30 (0.08)	0.27 (0.09)
S_{Chao}	0.20 (0.07)	0.21 (0.07)	0.25 (0.08)	0.19 (0.07)	0.22 (0.07)	0.22 (0.07)	0.24 (0.08)	0.25 (0.08)	0.29 (0.08)	0.29 (0.09)	0.29 (0.09)
D_{Mg}	0.18 (0.06)	0.18 (0.06)	0.22 (0.07)	0.17 (0.06)	0.20 (0.06)	0.19 (0.07)	0.23 (0.07)	0.23 (0.08)	0.27 (0.07)	0.30 (0.08)	0.27 (0.09)
H'	0.07 (0.03)	0.07 (0.03)	0.08 (0.03)	0.07 (0.04)	0.07 (0.03)	0.07 (0.04)	0.10 (0.04)	0.09 (0.05)	0.10 (0.04)	0.14 (0.05)	0.11 (0.08)
J'	0.07 (0.03)	0.07 (0.03)	0.08 (0.03)	0.07 (0.04)	0.07 (0.03)	0.07 (0.04)	0.10 (0.04)	0.09 (0.05)	0.10 (0.04)	0.14 (0.05)	0.11 (0.08)
D	0.03 (0.03)	0.03 (0.02)	0.03 (0.02)	0.03 (0.03)	0.02 (0.02)	0.03 (0.03)	0.04 (0.03)	0.03 (0.03)	0.04 (0.03)	0.05 (0.03)	0.05 (0.05)
$E_{1/D}$	0.04 (0.02)	0.03 (0.02)	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)	0.05 (0.02)	0.03 (0.02)	0.04 (0.02)	0.04 (0.02)	0.05 (0.03)	0.08 (0.07)
d	0.05 (0.04)	0.03 (0.03)	0.04 (0.04)	0.05 (0.04)	0.03 (0.02)	0.05 (0.05)	0.05 (0.04)	0.05 (0.04)	0.05 (0.05)	0.07 (0.05)	0.08 (0.08)

Table 4: Average proportional bias for similarity indices with sample size $n = 500$, when only one of the two samples may contain classification errors. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
QS	0.06 (0.04)	0.06 (0.04)	0.06 (0.05)	0.06 (0.04)	0.06 (0.04)	0.06 (0.04)	0.07 (0.05)	0.07 (0.05)	0.08 (0.05)	0.09 (0.06)	0.11 (0.06)
PMA	0.03 (0.02)	0.02 (0.02)	0.03 (0.02)	0.03 (0.02)	0.02 (0.02)	0.04 (0.03)	0.03 (0.02)	0.04 (0.03)	0.04 (0.02)	0.07 (0.05)	0.33 (0.13)
$1 - CM$	0.06 (0.04)	0.05 (0.04)	0.06 (0.05)	0.06 (0.04)	0.05 (0.04)	0.06 (0.05)	0.06 (0.05)	0.07 (0.06)	0.07 (0.06)	0.08 (0.06)	0.20 (0.09)
$1 - D^2_{Eucl}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.05 (0.03)
C_λ	0.03 (0.03)	0.02 (0.02)	0.02 (0.03)	0.04 (0.04)	0.02 (0.02)	0.05 (0.05)	0.03 (0.03)	0.04 (0.04)	0.05 (0.04)	0.07 (0.04)	0.46 (0.19)
J	0.08 (0.06)	0.08 (0.06)	0.09 (0.07)	0.08 (0.06)	0.08 (0.06)	0.09 (0.06)	0.10 (0.07)	0.10 (0.07)	0.11 (0.07)	0.12 (0.08)	0.15 (0.08)

Table 5: Average proportional bias for similarity indices with sample size $n = 500$, when both samples are classified and may contain classification errors. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
QS	0.05 (0.04)	0.06 (0.04)	0.06 (0.04)	0.05 (0.04)	0.05 (0.04)	0.05 (0.04)	0.07 (0.04)	0.07 (0.04)	0.08 (0.05)	0.10 (0.05)	0.09 (0.06)
PMA	0.05 (0.03)	0.04 (0.02)	0.04 (0.03)	0.05 (0.03)	0.04 (0.03)	0.06 (0.03)	0.05 (0.03)	0.07 (0.04)	0.06 (0.04)	0.11 (0.06)	0.22 (0.10)
$1 - CM$	0.08 (0.04)	0.09 (0.03)	0.10 (0.03)	0.08 (0.04)	0.09 (0.04)	0.09 (0.04)	0.12 (0.04)	0.11 (0.04)	0.13 (0.04)	0.19 (0.05)	0.23 (0.08)
$1 - D^2_{Eucl}$	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.01)	0.03 (0.02)
C_λ	0.04 (0.04)	0.03 (0.03)	0.04 (0.03)	0.05 (0.04)	0.03 (0.03)	0.06 (0.05)	0.05 (0.04)	0.06 (0.05)	0.05 (0.05)	0.09 (0.07)	0.22 (0.14)
J	0.07 (0.06)	0.09 (0.06)	0.09 (0.06)	0.08 (0.06)	0.08 (0.06)	0.08 (0.06)	0.11 (0.06)	0.10 (0.06)	0.12 (0.07)	0.16 (0.07)	0.14 (0.09)

The rest of the diversity, evenness and dominance indices have proportional bias 10 % or under (Table 3), at least with the better classifiers that have classification errors under 30 %. Actually, even for the poorly performing classifiers, NB (ce>51%) and MLP (ce>89%), the error propagation into the biological indices is surprisingly small. The reason for this is that, the calculation of these indices is based on taxa proportions instead of counts. Therefore few individual misclassifications have less influence on the index values, at least with reasonably large sample size. $E_{1/D}$ seems to have a slightly larger %bias than D because it is proportional to S and therefore affected more by extra species. Note that the proportional bias for H' and J' is identical, as the latter is a scaled version of the former. The Berger-Parker index, d, depends only on the most common taxa in the sample so it may have high %bias in river types where the most common taxa is one with a higher classification error rate. However, this problem can be overcome since biologists are likely to choose classification methods that identify the most common taxa of a sampling site well.

According to Tables 4 and 5, none of the similarity indices are as sensitive to classification errors as the richness indices based on presence/absence data (Table 3). For similarity indices, the quality of the classification method has a more clear impact as MLP produces severe %bias in the index values when compared to the other classifiers. However, not taking MLP into account, all of the similarity indices have proportional bias mostly under 10 % (Table 4). QS and J are based on presence/absence data but are much less biased than S, S_{Chao} and D_{Mg}. This may be because in QS and J the number of species affects both the numerator and denominator. Extra species due to misclassifications thus increase both the number of common taxa and the number of observed taxa and therefore do not increase the final index value as much. Euclidian similarity, $1 - D^2_{Eucl}$, and PMA index have very similar formulas, yet $1 - D^2_{Eucl}$ has smaller proportional bias than the PMA index. Unlike PMA, $1 - D^2_{Eucl}$ is affected by how the observations are distributed in non-common classes, giving a larger distance if the observations in the non-common classes are distributed evenly. Therefore Euclidian similarity has range $[-1, 1]$, compared to the range of the PMA $[0, 1]$.

The proportional bias increases the most for the Canberra metric, $1 - CM$, when both samples are classified (see Table 5), compared to the case when the reference sample is assumed to be known (Table 4). In fact, all similarity indices have higher expected values when both samples are classified, compared to the case when the reference sample is assumed to be known (see Fig. 2 and 3). The index values are often biased downwards when only one of the samples is classified and biased upwards when both samples contain classification errors. This may be caused by the fact that classification errors increase the entropy and evenness of the samples. The higher the evenness in both samples, the more similar they become.

While the aforementioned results (Figures 1, 2, 3 and Tables 3, 4 and 5) are obtained with sample size 500, we also tested sample sizes 200 and 1000 to assess whether error propagation in biological indices varies with sample size (see results in appendix, Tables 8, 9, 10, 11, 12 and 13). Of the diversity, richness, evenness and dominance indices, only S, S_{Chao} and D_{Mg} are affected by sample size. For S and D_{Mg}, the average proportional bias clearly increases with the sample size for all classification methods. For S_{Chao}, the %bias increases for good classifiers. When there are more observations in the sample, the chance of a misclassified observation introducing an extra species is higher. D_{Mg} proportional to sample size which should make it less sensitive to changes in sample size. However, when

Table 6: Average proportional bias for different classifiers over all river types and diversity, richness, evenness and dominance indices (DIV), similarity indices when one sample is classified (SIM1) and similarity indices when both samples are classified (SIM2). Here, $n = 500$. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias		
Classifier	DIV	SIM1	SIM2
GEKELM	0.10 (0.08)	0.04 (0.04)	0.05 (0.04)
KRDA	0.10 (0.08)	0.04 (0.04)	0.05 (0.04)
KSVM	0.12 (0.10)	0.04 (0.05)	0.06 (0.05)
GEELM	0.10 (0.08)	0.05 (0.04)	0.05 (0.04)
RBA	0.11 (0.09)	0.04 (0.05)	0.05 (0.05)
RBFN	0.11 (0.09)	0.05 (0.05)	0.06 (0.05)
LDA	0.13 (0.10)	0.05 (0.05)	0.07 (0.05)
RF	0.13 (0.10)	0.06 (0.06)	0.07 (0.05)
SVM	0.14 (0.12)	0.06 (0.06)	0.08 (0.06)
NB	0.17 (0.12)	0.07 (0.06)	0.11 (0.08)
MLP	0.16 (0.12)	0.22 (0.18)	0.15 (0.12)

calculating the %Bias, the $\log(n)$ terms are cancelled and the %Bias is identical to that of S. Of the similarity indices, the bias increases with sample size for QS, $1-CM$ and J when both samples may contain classification errors. PMA, C_λ and $1-D^2_{Eucl}$ are less sensitive to sample size.

In addition to studying the effect of classification errors on biological indices, we compare the different classification methods. Usually, classifiers are compared on error rate but we are interested in their effect on decision making via the indices. The classifiers which have classification errors under 20 % are very similar with respect to the %bias in biological indices (Table 6). However, note that the third best classifier based on error rate, KSVM, introduces more bias in the indices than some classification methods that have higher error rates than KSVM. This is more clear for diversity, richness, evenness and dominance indices. Even though the differences are small, this does show that classification error should not be the only basis in the selection of classification methods.

Last, we consider the effect of individual river types, i.e. the effect of species distribution combined with the different confusion matrices. When we use automated classification methods, the number of possible taxa is fixed based on the training image data and this sets the dimensions for the confusion matrix. In this setting, taxa distributions with only few taxa are problematic for indices based on presence/absence data (see e.g. Tables 14, 15, 16 and 17 in appendix). When a confusion matrix has many classes, misclassification easily introduces extra taxa into the samples and therefore affects the index values. The problem is even larger if the taxa present in the distribution happen to be ones with a high classification error. On the other hand, taxa distributions with the majority of the taxa present tend to produce smaller %bias in the index values. For indices based on proportions, the most problematic taxa distributions are ones where the most common taxa have high error rates as the highest proportions contribute most in the calculation of these indices.

For example, in our simulation study, proportion based indices for medium-sized woodland streams of northern Finland display higher bias than other river types (see Table 18 in appendix compared to Table 3). This is because almost half of this river type's observations are *Baetis rhodani* which is a taxa identified well only by RBA. Unsurprisingly, RBA is the only classifier with a low bias in proportion based indices for medium-sized woodland streams of northern Finland.

Using Eq. 2, we also study how the standard error of biological indices is affected by classification errors. However, as there is very little difference in the standard errors before and after classification, the results are not shown here.

5 Conclusions

In this paper, we discuss the effect of classification errors on biological indices describing richness, diversity, evenness, dominance and similarity. We study the error propagation into biological indices with simulation experiments based on real data. We train 11 classifiers with benthic macroinvertebrate image data and use these classification results to evaluate how different confusion matrices affect index values calculated from classified macroinvertebrate samples. We study which indices are most sensitive to misclassifications and sample size and how different taxa distributions affect the error propagation.

The most sensitive indices to classification errors are the richness indices based on presence/absence data, i.e. S, S_{Chao} and D_{Mg}. As the calculation of these indices relies on the number of observed species, even one misclassified observation can introduce an extra taxa into the calculation and therefore introduce bias into the index. These indices are even more sensitive to errors when there are fewer taxa in the species distribution than in the confusion matrix since this makes it possible to have false extra taxa. S, S_{Chao} and D_{Mg} are also sensitive to sample size since increasing sample size increases the possibility of misclassifications introducing extra taxa in the sample.

Diversity, evenness, dominance and similarity indices analyzed in this paper are less sensitive to classification errors than richness indices, with proportional bias less than 10 % when using good classifiers. Presence/absence based similarity indices, i.e. QS and J, are less biased than S, S_{Chao} and D_{Mg} because in their calculation extra taxa increase both the numerator and denominator, keeping the ratio roughly the same. Proportion-based indices can also be sensitive to classification errors if the most common taxa in the samples are poorly classified, i.e. identified. However, since biologists have prior knowledge of expected taxa distributions at sampling sites they are likely to choose the classification method accordingly. The classification methods used in this paper can be split into three groups: good classifiers (ce $<$ 20 %), mediocre classifiers (20 % $<$ ce $<$ 50 %) and poor classifiers (ce $>$ 50 %). Although different in error rates, the good classifiers do not really differ when considering the proportional bias they bring into biological indices, allowing to choose the most favourable classifier among them for a given scenario.

We found many of the similarity indices to be sensitive to sample size as well. When both samples being compared are classified, bias caused by misclassifications increases with sample size for QS, $1-CM$ and J. We found that for similarity indices, misclassifications often increase entropy of the samples. Thus, when both samples are classified, their similarity

increases and the similarity index values become over-estimated. Therefore decision makers should carefully consider cases where the necessity of mitigation measures is evaluated based on similarity values. Based on our analyses and simulation experiments, the similarity indices least affected by classification errors, sample size and taxa distributions are $1 - D^2_{Eucl}$, PMA and C_λ. The least biased diversity index is D. We acknowledge that there are other sources of bias, e.g. sampling error, but in this paper we limit our analyses on classification errors and restrict the study of the effect of sample size and taxa distribution to their interaction with classification errors, when the counts follow a multinomial distribution. We also note that the choice of an index ultimately depends on what needs to be measured from a monitored community but we would generally recommend proportion-based indices for diverse communities as these are the most robust against taxa misidentification error propagation, based on our simulation experiments. When shifting the biomonitoring and ecological status assessment process towards automation, the proposed expert system should consider only indices that are robust to automated classification errors.

The results in this paper were obtained using automated classification. A nice property of automated classification given a gold standard training set is the knowledge of confusion matrices. As misclassifications with good classifiers are systematic and predictable, for future work, correction methods will be considered in order to decrease the bias in biological indices due to misclassification. Even though we like to think that human experts rarely make identification errors, it does happen (Culverhouse *et al.*, 2003) and can cause remarkable bias in resulting index values and ecological status evaluations (Haase *et al.*, 2010). Unlike in automated methods, human expert errors rarely include knowledge of the human expert's confusion matrix. Also as human misclassifications may not be as systematic as with automated classifiers it is not sensible to construct a correction method for every single human expert. In contrast, it is highly sensible to construct a correction method for a well performing automated classifier to further boost its performance.

Interesting future research directions include i) the comparison between human experts and automated classifiers and how much human experts introduce bias into biological indices due to prior knowledge affecting identification errors, ii) analysis of the classification errors observed by expert systems when evaluation is conducted in varying conditions, e.g. differences in illumination that might be caused either by product failures or by external factors, iii) analysis of the classification errors observed when the expert system is trained and evaluated using data obtained by different conditions, iv) analysis of the classification errors observed when a test sample from a previously unseen class is processed.

Acknowledgements

We thank the Academy of Finland (projects 288584 (Kiranyaz, Iosifidis), 289364 (Gabbouj), 289076 (Ärje, Kärkkäinen) and 289104 (Meissner)) and the Ellen and Artturi Nyyssönen foundation for the grant of Ärje. The authors would like to thank Marko Vikstedt for the preparation of the monitoring data and Tuomas Turpeinen for the image data. We kindly thank Antti Penttinen for fruitful discussions and support.

References

Abdar, M., Zomorodi-Moghadam, M., Das, R. and Ting, I.-H. (2017) Performance analysis of classification algorithms on early detection of liver disease. *Expert Syst Appl*, **67**, 239–251.

Albatineh, A. N. and Niewiadomska-Bugaj, M. (2011) Correcting jaccard and other similarity indices for chance agreement in cluster analysis. *Adv. Data Anal. Classif.*, **5**, 179–200.

Ali, R., Lee, S. and Chung, T. C. (2017) Accurate multi-criteria decision making methodology for recommending machine learning algorithm. *Expert Syst Appl*, **71**, 259–278. DOI: 10.1016/j.eswa.2016.11.034.

Aroviita, J., Hellsten, S., Jyväsjärvi, J., Järvenpää, L., Järvinen, M., Karjalainen, S., Kauppila, P. and Keto, A. (2012) Guidelines for the ecological and chemical status classification of surface waters for 2012-2013 - updated assessment criteria and their application. *Environ. Adm. Guidel.*, **7**, 144.

Berger, W. H. and Parker, F. L. (1970) Diversity of planktonic foraminifera in deep sea sediments. *Science*, **168**, 1345–1347.

Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., Solimini, A., van de Bund, W., Zampoukas, N. and Hering, D. (2012) Three hundred ways to assess Europe's surface waters: An almost complete overview of biological methods to implement the Water Framework Directive. *Ecol. Indic.*, **18**, 31–41.

Biswas, A. and Biswas, B. (2017) Defining quality metrics for graph clustering evaluation. *Expert Syst Appl*, **71**, 1–17. DOI: 10.1016/j.eswa.2016.11.011.

Blaschko, M., Holness, G., Mattar, M., Lisin, D., Utgoff, P., Hanson, A., Schultz, H., Riseman, E., Sieracki, M., Balch, W. and Tupper, B. (2005) Automatic in situ identification of plankton. *Proceedings of the 7th IEEE Workshops on Application of Computer Vision (WACV/MOTION '05)*, **1**.

Breiman, L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Carayannis, E. G., Grogoroudis, E. and Goletsis, Y. (2016) A multilevel and multistage efficiency evaluation of innovation systems: A multiobjective dea approach. *Expert Syst Appl*, **62**, 63–80. DOI: 10.1016/j.eswa.2016.06.017.

Chao, A. (1984) Non-parametric estimation of the number of classes in a population. *Scand. J. .Stat.*, **11**, 265–270.

Chao, A. (1987) Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, **43**, 783–791.

Chen, X. H., Yamaguchi, Y. and Chen, J. (2010) A new measure of classification error: designed for landscape pattern index. *Int. Arch. Photogramm., Remote Sens. Spat. Inf. Sci.*, **38**, 759–762.

Clifford, H. T. and Stephenson, W. (1975) *An introduction to numerical classification.* London: Academic Press.

Cortes, C. and Vapnik, V. (1995) Support-vector networks. *Mach. Learn.*, **20**, 273–297.

Culverhouse, P., Williams, R., Benfield, M., Flood, P., Sell, A., Mazzocchi, M., Buttino, I. and Sieracki, M. (2006) Automatic image analysis of plankton: future perspectives. *Mar. Ecol.-Prog. Ser.*, **312**.

Culverhouse, P., Williams, R., Reguera, B., Herry, V. and González-Gil, S. (2003) Do experts make mistakes? a comparison of human and machine identification dinoflagellates. *Mar. Ecol.-Prog. Ser.*, **247**, 17–25.

Dickinson, J. L., Shirk, J., Bonter, D., Bonney, R., Crain, R. L., Martin, J., Phillips, T. and Purcell, K. (2012) The current state of citizen science as a tool for ecological research and public engagement. *Front. Ecol. Environ.*, **10**, 291–297.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001) *Pattern Classification (2nd ed.).* Wiley: New York.

Fortier, J.-J. (1992) Best linear corrector of classification estimates of proportions of objects in several unknown classes. *Can. J. Stat.*, **20**, 23–33.

Gardiner, M. M., Allee, L. L., Brown, P. M., Losey, J. E., Roy, H. E. and Smyth, R. R. (2012) Lessons from lady beetles: accuracy of monitoring data from us and uk citizen-science programs. *Front. Ecol. Environ.*, **10**, 471–476.

Haase, P., Pauls, S. U., Schindehütte, K. and Sunderman, A. (2010) First audit of macroinvertebrate samples from an eu water framework directive monitoring program: human error greatly lowers precision of assessment results. *J. N. Am. Benthol. Soc.*, **29**, 1279–1291.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference and Prediction (2nd ed.).* Springer: New York.

Haykin, S. (2009) *Neural Networks and Learning Machines.* Upper Saddle River, NJ: Pearson, third ed. edn.

Healy, J. D. (1981) The effects of misclassification error on the estimation of several proportions. *Bell Syst. Tech. J.*, **60**, 697–705.

Horn, H. S. (1966) Measurement of "overlap" in comparative ecological studies. *Am. Nat.*, **100**, 419–424.

Iosifidis, A., Tefas, A. and Pitas, I. (2014a) Kernel reference discriminant analysis. *Pattern Recogn. Lett.*, **49**, 85–91.

Iosifidis, A., Tefas, A. and Pitas, I. (2014b) On the kernel extreme learning machine classifier. *Pattern Recogn. Lett.*, **54**, 11–17.

Iosifidis, A., Tefas, A. and Pitas, I. (2015) Graph embedded extreme learning machine. *IEEE Transactions on Cybernetics*. D.O.I. 10.1109/TCYB.2015.2401973.

Jaccard, P. (1901) Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *B. Soc. Vaud. Sci. Nat.*, **37**, 547–579.

Janson, S. and Vegelius, J. (1981) Measures of ecological association. *Oecologia*, **49**, 371–376.

Joutsijoki, H., Meissner, K., Gabbouj, M., Kiranyaz, S., Raitoharju, J., Ärje, J., Kärkkäinen, S., Tirronen, V., Turpeinen, T. and Juhola, M. (2014) Evaluating the performance of artificial neural networks for the classification of freshwater benthic macroinvertebrates. *Ecol. Inform.*, **20**, 1–12.

Kiranyaz, S., Ince, T., Pulkkinen, J., Gabbouj, M., Ärje, J., Kärkkäinen, S., Tirronen, V., Juhola, M., Turpeinen, T. and Meissner, K. (2011) Classification and retrieval on macroinvertebrate image databases. *Comput. Biol. Med.*, **41**, 463–472.

Kiranyaz, S., Ince, T., Yildirim, A. and Gabbouj, M. (2009) Evolutionary artificial neural networks by multi-dimensional particle swarm optimization. *Neural Networks*, **22**, 1448–1462.

Lance, G. N. and Williams, W. T. (1967) Mixed-data classificatory programs i. agglomerative systems. *Aust. Comput. J.*, **1**, 15–20.

Lytle, D. A., Martínez-Muñoz, G., Zhang, W., Larios, N., Shapiro, L., Paasch, R., Moldenke, A., Mortensen, E. N., Todorovic, S. and Dietterich, T. G. (2010) Automated processing and identification of benthic invertebrate samples. *J. N. Am. Benthol. Soc.*, **29**, 867–874.

Magurran, A. E. (2004) *Measuring Biological Diversity*. Malden (Ma.): Blackwell.

Magurran, A. E. and McGill, B. J., eds. (2010) *Biological Diversity. Frontiers in Measurement and Assessment*. Oxford University Press.

Margalef, R. (1958) *Temporal succession and spatial heterogeneity in phytoplankton*. Univ. Calif. Press, Berkeley.

Novak, M. A. and Bode, R. W. (1992) Percent model affinity: a new measure of macroinvertebrate community composition. *J. N. Am. Benthol.l Soc.*, **11**, 80–85.

Paninski, L. (2003) Estimation of entropy and mutual information. *Neural Comput.*, **15**, 1191–1253.

Pielou, E. C. (1969) *An introduction to mathematical ecology*. New York: Wiley.

Pielou, E. C. (1975) *Ecological diversity*. New York: Wiley InterScience.

Piltan, M. and Sowlati, T. (2016) Multi-criteria assessment of partnership components. *Expert Syst Appl*, **64**, 605–617. DOI: 10.1016/j.eswa.2016.08.006.

Rasband, W. S. (1997-2010) *ImageJ*. U.S. National Institutes of Health, Bethesda, Maryland, USA. URL `http://rsb.info.nih.gov/ij/`.

Renkonen, O. (1938) Statisch-ökologische Untersuchungen über die terrestrische Käferwelt der finnischen Bruchmoore. *Ann. Zool. Soc. Bot. Fenn. Vanamo*, **6**, 1–231.

Scardi, M., Cataudella, S., Dato, P. D., Fresi, E. and Tancioni, L. (2008) An expert system based on fish assemblages for evaluating the ecological quality of streams and rivers. *Ecol Inform*, **3**, 55–63.

Shannon, C. and Weaver, W. (1963) *The mathematical theory of communication*. University Illinois Press, Urbana.

Shao, G., Liu, D. and Zhao, G. (2001) Relationships of image classification accuracy and variation of landscape statistics. *Can. J. Remote Sens.*, **27**, 33–43.

Simpson, E. H. (1949) Measurement of diversity. *Nature*, **163**, 688.

Smith, B. and Wilson, J. B. (1996) A consumer's guide to evenness measures. *Oikos*, **76**, 70–82.

Smith, E. P. (2002) *Ecological statistics*. John Wiley & Sons.

Smith, W. and Grassel, J. F. (1977) Sampling properties of a family of diversity measures. *Biometrics*, **33**, 283–292.

Sørensen, T. (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content, and its application to analyses of the vegetation on danish commons. *K dan Vidensk Selsk Biol Skr*, **5**, 1–34.

Sugiyama, M. (2007) Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *J. Mach. Learn. Res.*, **8**, 1027–1061.

Tong, Y. L. (1983) Some distribution properties of the sample species-divervsity indices and their applications. *Biometrics*, **39**, 999–1008.

Wickham, J. D., O'Neill, R. V., Riitters, K. H., Wade, T. G. and Jones, K. B. (1997) Sensitivity of selected landscape pattern metrics to land-cover misclassification and differencies in land-cover composition. *Photogramm. Eng. Rem. S.*, **63**, 397–402.

Ärje, J., Choi, K.-P., Divino, F., Meissner, K. and Kärkkäinen, S. (2016) Understanding the statistical properties of the percent model affinity index can improve biomonitoring related decision making. *Stoch. Env. Res. Risk A.* (Published online).

Ärje, J., Kärkkäinen, S., Turpeinen, T. and Meissner, K. (2013) Breaking the curse of dimensionality in quadratic discriminant analysis models with a novel variant of a bayes classifier enhances automated taxa identification of freshwater macroinvertebrates. *Environmetrics*, **24**, 248–259.

Table 7: Taxa used for the classification and simulation experiments. *Baetis muticus* and *Baetis niger* are identified separately in the image data but are combined here into the *Baetis niger group* to have equal taxa lists in both image and monitoring data. Similarly *Protonemura intricata* and *Protonemura meyeri* are combined to *Protonemura spp.*

Taxonomic group	
Ameletus inopinatus	*Habrophlebia spp.*
Arctopsyche ladogensis	*Heptagenia dalecarlica*
Asellus aquaticus	*Hydraena spp.*
Baetis niger group	*Hydropsyche pellucidula*
Baetis rhodani	*Hydropsyche saxonica*
Bithytnia tentaculata	*Hydropsyche siltalai*
Caenis spp.	*Isoperla spp.*
Corixidae	*Leuctra spp.*
Ceratopsyche silfvenii	*Limnius volckmari*
Ceratopogonidae	*Micrasema gelidum*
Cheumatopsyche lepida	*Micrasema setiferum*
Diura spp.	*Nemoura spp.*
Elmis aenea	*Sphaeriidae*
Ephemerella aurivillii	*Protonemura spp.*
Ephemerella ignita	*Rhyacophila nubila*
Ephemerella mucronata	*Taeniopteryx nebulosa*

Appendix

Table 8: Average proportional bias for diversity, richness, evenness and dominance indices for sample size $n = 200$. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
S	0.14 (0.06)	0.15 (0.05)	0.17 (0.06)	0.14 (0.06)	0.16 (0.06)	0.15 (0.07)	0.20 (0.06)	0.19 (0.07)	0.22 (0.05)	0.27 (0.08)	0.22 (0.09)
S_{Chao}	0.18 (0.07)	0.19 (0.06)	0.23 (0.07)	0.17 (0.06)	0.21 (0.07)	0.19 (0.08)	0.24 (0.07)	0.24 (0.08)	0.27 (0.07)	0.30 (0.08)	0.28 (0.09)
D_{Mg}	0.14 (0.06)	0.15 (0.05)	0.17 (0.06)	0.14 (0.06)	0.16 (0.06)	0.15 (0.07)	0.20 (0.06)	0.19 (0.07)	0.22 (0.05)	0.27 (0.08)	0.22 (0.09)
H'	0.07 (0.03)	0.07 (0.03)	0.07 (0.03)	0.07 (0.04)	0.06 (0.03)	0.07 (0.04)	0.09 (0.03)	0.08 (0.04)	0.10 (0.04)	0.14 (0.05)	0.11 (0.08)
J'	0.07 (0.03)	0.07 (0.03)	0.07 (0.03)	0.07 (0.04)	0.06 (0.03)	0.07 (0.04)	0.09 (0.03)	0.08 (0.04)	0.10 (0.04)	0.14 (0.05)	0.11 (0.08)
D	0.03 (0.03)	0.03 (0.02)	0.03 (0.02)	0.03 (0.03)	0.02 (0.02)	0.03 (0.03)	0.04 (0.03)	0.03 (0.03)	0.04 (0.03)	0.05 (0.03)	0.05 (0.05)
$E_{1/D}$	0.04 (0.02)	0.03 (0.02)	0.04 (0.02)	0.04 (0.03)	0.04 (0.02)	0.05 (0.03)	0.03 (0.02)	0.04 (0.02)	0.05 (0.02)	0.05 (0.03)	0.08 (0.07)
d	0.04 (0.04)	0.03 (0.03)	0.04 (0.04)	0.05 (0.04)	0.03 (0.02)	0.05 (0.05)	0.05 (0.04)	0.05 (0.04)	0.05 (0.05)	0.07 (0.05)	0.08 (0.08)

Table 9: Average proportional bias for similarity indices with sample size $n = 200$, when only one of the two samples may contain classification errors. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
QS	0.05 (0.04)	0.05 (0.04)	0.05 (0.04)	0.05 (0.04)	0.05 (0.04)	0.05 (0.04)	0.06 (0.05)	0.06 (0.05)	0.07 (0.05)	0.10 (0.06)	0.16 (0.06)
PMA	0.03 (0.02)	0.02 (0.02)	0.03 (0.02)	0.03 (0.02)	0.02 (0.02)	0.04 (0.03)	0.03 (0.02)	0.04 (0.03)	0.04 (0.02)	0.06 (0.05)	0.32 (0.13)
$1 - CM$	0.05 (0.04)	0.04 (0.04)	0.05 (0.04)	0.05 (0.04)	0.05 (0.04)	0.05 (0.05)	0.06 (0.05)	0.06 (0.06)	0.06 (0.05)	0.09 (0.06)	0.20 (0.09)
$1 - D^2_{Eucl}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	0.01 (0.00)	0.00 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.05 (0.03)
C_λ	0.03 (0.03)	0.02 (0.02)	0.02 (0.03)	0.04 (0.04)	0.02 (0.02)	0.05 (0.04)	0.03 (0.02)	0.04 (0.04)	0.05 (0.04)	0.07 (0.04)	0.45 (0.18)
$Jacc$	0.07 (0.05)	0.07 (0.05)	0.07 (0.06)	0.07 (0.05)	0.07 (0.06)	0.07 (0.06)	0.09 (0.06)	0.09 (0.07)	0.09 (0.07)	0.13 (0.08)	0.20 (0.08)

Table 10: Average proportional bias for similarity indices with sample size $n = 200$, when both samples are classified and may contain classification errors. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
QS	0.04 (0.04)	0.04 (0.04)	0.05 (0.04)	0.04 (0.04)	0.04 (0.04)	0.05 (0.04)	0.06 (0.04)	0.05 (0.04)	0.06 (0.04)	0.08 (0.05)	0.07 (0.06)
PMA	0.04 (0.03)	0.03 (0.02)	0.03 (0.03)	0.04 (0.03)	0.04 (0.02)	0.05 (0.03)	0.04 (0.03)	0.05 (0.04)	0.05 (0.04)	0.09 (0.06)	0.19 (0.10)
$1 - CM$	0.05 (0.04)	0.05 (0.03)	0.06 (0.04)	0.06 (0.04)	0.05 (0.04)	0.06 (0.04)	0.08 (0.04)	0.07 (0.05)	0.08 (0.05)	0.13 (0.05)	0.16 (0.08)
$1 - D^2_{Eucl}$	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.01)	0.03 (0.02)
C_λ	0.04 (0.03)	0.03 (0.03)	0.04 (0.03)	0.05 (0.04)	0.03 (0.03)	0.06 (0.05)	0.04 (0.04)	0.06 (0.05)	0.05 (0.05)	0.08 (0.07)	0.20 (0.14)
J	0.06 (0.05)	0.06 (0.05)	0.06 (0.05)	0.06 (0.06)	0.06 (0.05)	0.06 (0.05)	0.08 (0.06)	0.07 (0.06)	0.08 (0.06)	0.12 (0.07)	0.10 (0.08)

Table 11: Average proportional bias for diversity, richness, evenness and dominance indices with sample size $n = 1000$. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
S	0.20 (0.07)	0.20 (0.06)	0.24 (0.07)	0.19 (0.06)	0.22 (0.07)	0.21 (0.07)	0.24 (0.07)	0.24 (0.07)	0.28 (0.08)	0.29 (0.08)	0.28 (0.09)
S_{Chao}	0.21 (0.08)	0.21 (0.08)	0.25 (0.09)	0.20 (0.07)	0.22 (0.08)	0.23 (0.08)	0.24 (0.08)	0.25 (0.08)	0.28 (0.09)	0.27 (0.09)	0.28 (0.09)
D_{Mg}	0.20 (0.07)	0.20 (0.06)	0.24 (0.07)	0.19 (0.06)	0.22 (0.07)	0.21 (0.07)	0.24 (0.07)	0.24 (0.07)	0.28 (0.08)	0.29 (0.08)	0.28 (0.09)
H'	0.07 (0.03)	0.07 (0.03)	0.08 (0.03)	0.07 (0.04)	0.07 (0.03)	0.07 (0.04)	0.10 (0.04)	0.09 (0.05)	0.10 (0.04)	0.14 (0.05)	0.11 (0.08)
J'	0.07 (0.03)	0.07 (0.03)	0.08 (0.03)	0.07 (0.04)	0.07 (0.03)	0.07 (0.04)	0.10 (0.04)	0.09 (0.05)	0.10 (0.04)	0.14 (0.05)	0.11 (0.08)
D	0.03 (0.03)	0.03 (0.02)	0.03 (0.02)	0.03 (0.03)	0.02 (0.02)	0.03 (0.03)	0.04 (0.03)	0.03 (0.03)	0.04 (0.03)	0.05 (0.03)	0.05 (0.05)
$E_{1/D}$	0.03 (0.02)	0.03 (0.02)	0.04 (0.03)	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)	0.03 (0.02)	0.04 (0.02)	0.04 (0.02)	0.06 (0.03)	0.08 (0.07)
d	0.05 (0.04)	0.03 (0.03)	0.04 (0.04)	0.05 (0.04)	0.03 (0.02)	0.05 (0.05)	0.05 (0.05)	0.05 (0.04)	0.05 (0.05)	0.07 (0.05)	0.08 (0.08)

Table 12: Average proportional bias for similarity indices with sample size $n = 1000$, when only one of the two samples may contain classification errors. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
QS	0.06 (0.04)	0.06 (0.04)	0.06 (0.05)	0.06 (0.04)	0.06 (0.04)	0.06 (0.04)	0.07 (0.05)	0.07 (0.05)	0.07 (0.05)	0.08 (0.05)	0.09 (0.05)
PMA	0.03 (0.02)	0.02 (0.02)	0.03 (0.02)	0.03 (0.02)	0.02 (0.02)	0.04 (0.03)	0.03 (0.02)	0.04 (0.03)	0.04 (0.02)	0.07 (0.05)	0.33 (0.13)
$1 - CM$	0.06 (0.04)	0.05 (0.04)	0.06 (0.04)	0.06 (0.04)	0.05 (0.04)	0.06 (0.05)	0.06 (0.05)	0.07 (0.05)	0.07 (0.05)	0.08 (0.06)	0.20 (0.09)
$1 - D^2_{Eucl}$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.01 (0.00)	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)	0.05 (0.03)
C_λ	0.03 (0.03)	0.02 (0.02)	0.02 (0.03)	0.04 (0.04)	0.02 (0.02)	0.05 (0.05)	0.03 (0.02)	0.04 (0.04)	0.05 (0.04)	0.07 (0.04)	0.46 (0.19)
J	0.08 (0.06)	0.08 (0.06)	0.09 (0.07)	0.08 (0.05)	0.09 (0.06)	0.09 (0.06)	0.1. (0.07)	0.10 (0.07)	0.11 (0.07)	0.11 (0.07)	0.12 (0.07)

Table 13: Average proportional bias for similarity indices with sample size $n = 1000$, when both samples are classified and may contain classification errors. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
QS	0.06 (0.03)	0.07 (0.03)	0.08 (0.03)	0.06 (0.03)	0.07 (0.03)	0.06 (0.03)	0.08 (0.03)	0.08 (0.04)	0.10 (0.04)	0.11 (0.04)	0.10 (0.05)
PMA	0.05 (0.03)	0.04 (0.02)	0.05 (0.03)	0.05 (0.03)	0.05 (0.03)	0.06 (0.03)	0.06 (0.03)	0.07 (0.04)	0.07 (0.04)	0.12 (0.06)	0.24 (0.10)
$1 - CM$	0.11 (0.03)	0.11 (0.03)	0.13 (0.03)	0.11 (0.03)	0.12 (0.03)	0.12 (0.04)	0.15 (0.03)	0.15 (0.04)	0.17 (0.04)	0.23 (0.05)	0.29 (0.08)
$1 - D^2_{Eucl}$	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.01 (0.01)	0.02 (0.01)	0.03 (0.02)
C_λ	0.05 (0.04)	0.03 (0.03)	0.04 (0.03)	0.05 (0.04)	0.04 (0.03)	0.06 (0.05)	0.05 (0.04)	0.06 (0.05)	0.06 (0.04)	0.10 (0.07)	0.23 (0.14)
J	0.10 (0.05)	0.11 (0.05)	0.13 (0.05)	0.10 (0.05)	0.11 (0.05)	0.11 (0.05)	0.14 (0.05)	0.13 (0.06)	0.16 (0.06)	0.18 (0.07)	0.17 (0.07)

Table 14: Proportional bias for richness indices for large or extra large woodland reference streams of southern Finland with sample size $n = 500$. For this river type, $c = 22$.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
S	0.28	0.27	0.33	0.28	0.30	0.31	0.33	0.32	0.34	0.42	0.42
S_{Chao}	0.31	0.30	0.36	0.29	0.31	0.32	0.35	0.33	0.37	0.41	0.44
D_{Mg}	0.28	0.27	0.33	0.28	0.30	0.31	0.33	0.32	0.34	0.41	0.42

Table 15: Proportional bias for richness indices for small peatland reference streams of southern Finland with sample size $n = 500$. For this river type, $c = 19$.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
S	0.29	0.29	0.34	0.27	0.34	0.30	0.36	0.36	0.39	0.46	0.41
S_{Chao}	0.36	0.37	0.43	0.34	0.41	0.40	0.41	0.42	0.46	0.48	0.46
D_{Mg}	0.29	0.29	0.34	0.27	0.34	0.30	0.36	0.39	0.46	0.41	

Table 16: Proportional bias for richness indices for medium-sized peatland non-reference streams of northern Finland with sample size $n = 500$. For this river type, $c = 30$.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
S	0.07	0.09	0.12	0.07	0.10	0.07	0.13	0.11	0.17	0.16	0.17
S_{Chao}	0.10	0.10	0.13	0.10	0.11	0.10	0.12	0.12	0.16	0.14	0.16
D_{Mg}	0.07	0.09	0.12	0.07	0.10	0.07	0.13	0.11	0.17	0.16	0.17

Table 17: Proportional bias for richness indices for large or extra large peatland non-reference streams of northern Finland with sample size $n = 500$. For this river type, $c = 29$.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
S	0.06	0.08	0.08	0.05	0.09	0.05	0.10	0.08	0.12	0.13	0.09
S_{Chao}	0.07	0.08	0.09	0.07	0.08	0.09	0.09	0.10	0.12	0.11	0.10
D_{Mg}	0.06	0.08	0.08	0.05	0.09	0.05	0.10	0.08	0.12	0.13	0.09

Table 18: Proportional bias for proportion-based indices for medium-sized woodland non-reference streams in northern Finland with sample size $n = 500$. Standard deviation of the proportional bias is presented in parenthesis.

	%Bias										
Index	GEKELM	KRDA	KSVM	GEELM	RBA	RBFN	LDA	RF	SVM	NB	MLP
H'	0.13	0.11	0.14	0.14	0.06	0.15	0.15	0.15	0.19	0.20	0.27
J'	0.13	0.11	0.14	0.14	0.06	0.15	0.15	0.15	0.19	0.20	0.27
D	0.11	0.09	0.11	0.12	0.04	0.13	0.12	0.12	0.14	0.14	0.18
$E_{1/D}$	0.07	0.02	0.03	0.08	0.04	0.09	0.04	0.05	0.05	0.07	0.14
d	0.18	0.12	0.15	0.19	0.04	0.23	0.17	0.18	0.21	0.19	0.27