UNIVERSITY OF JYVÄSKYLÄ
DEPARTMENT OF MATHEMATICS
AND STATISTICS

REPORT 156

UNIVERSITÄT JYVÄSKYLÄ
INSTITUT FÜR MATHEMATIK
UND STATISTIK

BERICHT 156

IMPROVING STATISTICAL CLASSIFICATION METHODS AND ECOLOGICAL STATUS ASSESSMENT FOR RIVER MACROINVERTEBRATES

JOHANNA ÄRJE

To be presented, with permission of the Faculty of Mathematics and Science
of the University of Jyväskylä, for public criticism in Auditorium S212
on September 17th, 2016, at 12 o'clock noon.

JYVÄSKYLÄ
2016

Editor: Pekka Koskela
Department of Mathematics and Statistics
P.O. Box 35 (MaD)
FI–40014 University of Jyväskylä
Finland

ISBN 978-951-39-6706-2 (print)
ISBN 978-951-39-6707-9 (pdf)
ISSN 1457-8905

University Printing House
Jyväskylä 2016

Abstract

Aquatic ecosystems are facing a growing number of human-induced stressors and the need to implement more biomonitoring to assess the ecological status of water bodies is eminent. This dissertation aims at providing tools to reduce the costs and improve the accuracy of freshwater benthic macroinvertebrate biomonitoring. To improve the cost-efficiency, we consider automated classification and develop a novel classifier suitable for complex macroinvertebrate image data. To enhance the accuracy of macroinvertebrate biomonitoring, we study the statistical properties of the Percent Model Affinity index crucial to current Finnish biomonitoring and the factors affecting these statistics. Finally, we perform a simulation study to analyze how different biological indices are affected by misclassifications in automated identification of macroinvertebrates.

Acknowledgements

I would like to thank my two supervisors, Dr. Salme Kärkkäinen and Dr. Kristian Meissner for their relentless efforts on helping me graduate. They have been my advisors and support system from way back when I started my master's thesis. Thank you for helping me grow as a researcher. And thank you, Kristian, for making sure we never lost sight of the biologist's point of view.

A special thanks goes to the Finnish Environment Institute for making this research possible by providing me with interesting research questions as well as all of my data. I truly hope that this research benefits you in your efforts to monitor, study and preserve the environment. For the image data, I owe my gratitudes also to Dr. Tuomas Turpeinen and Dr. Ville Tirronen.

This work was funded by several entities. I am grateful to the COMAS graduate school for helping me start this study and the University of Jyväskylä for providing me with not only the facilities but also food on my plate when I needed it. I would like to thank the Maj and Tor Nessling foundation for allowing me to not stress about grant applications for a few years. I owe my gratitude to the Ellen and Artturi Nyyssönen foundation and the Academy of Finland for funding the final stage of the project.

I have been lucky enough to collaborate with Associate Professor Fabio Divino and Associate Professor Kwok-Pui Choi. The third article would have never come true without you. I also wish to thank our collaborators, Professor Moncef Gabbouj, Associate Professor Türker Ince, Dr. Alexandros Iosifidis and Professor Serkan Kiranyaz, who are co-authors in the fourth article. I am grateful to Professor Seppo Pynnönen and Professor Bikas K. Sinha for taking the time to review my thesis, and to Tuula Blåfield for proof-reading the compilation part. I want to express my deepest gratitude to the faculty of Statistics at the University of Jyväskylä and the administrative staff for making this such a friendly and great environment to study and work.

The past six years have been exciting, inspirational and challenging - sometimes driving me to edge of my sanity. I would like to thank my lovely co-PhD students for their vital support and understanding. I have been lucky enough to get to work with dear friends and to make new ones.

Thank you mom, dad and little sister for always staying interested although I realize statistics can sound like a language from another planet to people not studying the subject. Thank you, too, big brother, who wrestled with your own dissertation and the same challenges, for all our talks. You have always made me believe I can do this. I have been quite surrounded by the academia with my family and in-laws. Thank you dear mother-in-law and father-in-law for all your amazing support helping me carry on through the hard parts. The most heartfelt thanks go to my dear husband. I hope you know there would be no point doing all this or

anything else for the matter, if you wouldn't be here to share it. Last but not least, I would like to thank a gorgeous group of friends, my dear Excessive Knitters Anonymous for a place to think, talk and laugh about something completely different.

Jyväskylä, July 2016

Johanna Ärje

List of original publications

This thesis consists of an introductory part and the publications listed below.

I Ärje, J., Kärkkäinen, S., Meissner, K. & Turpeinen, T. (2010). Statistical classification and proportion estimation – an application to a macroinvertebrate image database. *Proceedings of the 2010 IEEE Workshop on Machine Learning for Signal Processing (MLSP 2010)*, Kittilä, Finland, pages 373–378.

II Ärje, J., Kärkkäinen, S., Turpeinen, T. & 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(4), 248–259.

III Ärje, J., Choi, K.-P., Divino, F., Meissner, K. & Kärkkäinen, S. (2016). Understanding the statistical properties of the percent model affinity index can improve biomonitoring related decision making. *Stochastic Environmental Research and Risk Assessment*. In press.

IV Ärje, J., Kärkkäinen, S., Meissner, K., Iosifidis, A., Ince, T., Gabbouj, M. & Kiranyaz, S. The effect of automated taxa identification errors on biological indices. Submitted.

The author of this dissertation is the main contributor of the joint papers I, II and IV. The theory of papers I and II was developed together with the second author. Except for some classification results in papers II and IV, the author of this dissertation has performed all the analyses and simulations in the joint papers I–IV. The co-authors have contributed in writing minor parts of the papers I, II and IV. The theoretical results in paper III were developed jointly by the first, second, third and fifth author. All the authors had an active role in writing paper III.

Contents

Chapter 1

Introduction

The problems addressed in this thesis originate from the need to improve the accuracy and cost-effectiveness of aquatic biomonitoring. This thesis deals with statistical classification, the statistical properties of Percent Model Affinity index (*PMA*, Novak and Bode, 1992) and how classification errors propagate into different biological indices. The proposed methods and results have been investigated in extensive simulation experiments and were validated with real data collected by the Finnish Environment Institute.

In aquatic biomonitoring, changes in the communities of biological indicator groups are measured to evaluate the ecological status of surface waters. These biological indicator groups include fish, periphyton (i.e. a mixture of algae, bacteria and other similar organisms living on the surface of aquatic plants), phytoplankton and benthic macroinvertebrates (Fig. 1.1). The ecological status assessment of surface waters is required by the European Union's Water Framework Directive (WFD, 2000) by all member states. In this thesis, we concentrate on freshwater benthic macroinvertebrates.

Figure 1.1: Freshwater benthic macroinvertebrates used in aquatic biomonitoring. From the top left: *Baetis muticus*, *Ceratopsyche silfvenii*, *Isoperla sp.*, *Protonemura intricata*, *Rhyacophila nubila* and *Taeniopteryx nebulosa*. These images have been scaled and do not reflect the size differences between the taxonomic groups.

Benthic macroinvertebrates are bottom dwelling animals without backbones. They are a diverse group of species that react quickly and strongly to changes in their environment (Rosenberg and Resh, 1993). Different taxonomic groups have varying sensitivities to different stressors and their community composition can reflect subtle human-induced changes

over long-term time scales (Rosenberg and Resh, 1993). The monitoring process for macroinvertebrates is presented in Figure 1.2. First, macroinvertebrates are sampled and identified to taxonomical groups by a human expert. Then, the observed taxa frequencies are used to calculate several biological indices, and the index values – together with other monitoring data – are used to evaluate the ecological status of the monitored waterbody. Further, decision makers use these ecological status assessments to determine whether or not monitored waterbodies are in need of mitigation measures.

Figure 1.2: The process of biomonitoring

Aquatic ecosystems are facing a growing number of anthropogenic pressures, e.g. eutrophication and global warming. The growing global need to implement more biomonitoring is apparent but due to the cost-intensity of manual taxonomic identification of samples, this need cannot currently be met. The apparent mismatch between monitoring demands and funding thus calls for more efficient sample processing. In order to lower the cost of the identification step of the biomonitoring process, the automated identification of macroinvertebrates has been studied in Tirronen et al. (2009); Lytle et al. (2010); Kiranyaz et al. (2010a,b, 2011); Joutsijoki et al. (2014); Joutsijoki and Juhola (2012) as well as in Articles I and II of this thesis.

To improve the accuracy of the ecological status assessment of surface waters, one must understand the variety of errors and biases that can be introduced into the process chain at each step (Fig. 1.2). First, the conclusions based on samples are dependent on how well the samples represent the community of interest. Sampling variation is also a source of uncertainty in biomonitoring. Second, as some of the sampled individuals may be misidentified, the identification step introduces bias and variation into the process via classification error. Third, the statistical properties of the biological indices affect the precision of the ecological status assessment. When choosing an index, one should know what the expected value, bias and variance of the index are and what factors affect them. In this thesis, we concentrate on the statistical properties of biological indices especially in the case of automated identification. For simplicity, our investigations are based on frequencies modelled by a multinomial distribution. The effects of other models and design are not considered.

The aim of this thesis is to provide tools to reduce the costs, bias and variation of the biomonitoring of freshwater benthic macroinvertebrates. Articles I and II explore the automated classification of benthic macroinvertebrates in order to reduce the costs of expensive and time-consuming manual identification. Article I presents a novel classification method, Random Bayes Array (RBA, named random Bayes forest in this paper). In Article I, RBA is used to classify a pilot study image data set of eight macroinvertebrate taxa and its performance is compared with several other popular classifiers. Article II provides a more thorough description of RBA and justifies its basic idea. In Article II, RBA and other popular classification methods are tested on much larger macroinvertebrate image data of 35 different taxonomical identities. Compared with Article I, Article II also explores new rules for making the final class decision in RBA and utilizes the varying importance of explanatory variables in building RBA. Articles III and IV study the different sources of bias and variation in the

monitoring process. Article III presents the statistical properties of the *PMA* index crucial to current Finnish aquatic biomonitoring and explores what factors increase or decrease its bias and variation. Finally, Article IV presents a simulation study that investigates how different types of classification errors propagate into different biological indices causing bias in the index values. The bias in taxa proportions due to classification errors is also studied in Article I.

The automated identification, i.e. classification of macroinvertebrates, is a complex task because of the large number of taxonomic identities and features needed to separate them. The problem of high-dimensional data is common in genomics and computational biology. It often results in high variance and overfitting of classifiers (Hastie et al., 2009; Domingos, 2012). In an attempt to evade the curse of dimensionality, one can perform feature selection, regularization or make assumptions on the independence of the features (see e.g. Duda et al., 2001, 2012; Hastie et al., 2009). Hastie et al. (2009) recommend quadratic discriminant analysis (QDA) as a standard approach to pattern recognition problems, because it provides good classification records (Michie et al., 1994). When the number of features greatly exceeds the number of observations per class, the use of QDA is not supported, but linear methods, such as linear discriminant analysis (LDA) or Naïve Bayes (NB), might be appropriate (Duda et al., 2001). However, the strict assumptions of equal covariance matrices in LDA and the independence of features in NB are unrealistic with complex data. One option is to use regularized discriminant analysis (RDA, Friedman, 1989), but in this work we will present a novel approach – an ensemble of random QDA classifiers, called Random Bayes Array (RBA, Articles I & II). Nonlinear methods, such as QDA, can benefit from introducing randomness into the model (Hastie et al., 2009). In RBA, selecting a random subset of the features reduces multicollinearity and enables the use of QDA by constraining the dimension of the covariance matrix. The ensemble approach reduces the high variance caused by high-dimensional data (Breiman, 1996b) and makes the classifier more robust to outliers and deviations from normality (see Article II and Dietterich, 2000).

Understanding the statistical properties of biological indices is a key component in improving the accuracy of the biomonitoring process. The *PMA* index is a popular measure of the similarity of two probability profiles used in many fields of applications, e.g. biomonitoring (Alahuhta et al., 2009; Kauppila et al., 2012), ecology (Renkonen, 1938), sawmill industry (Bergstrand, 1989) and sociology (Jahn et al., 1947; Duncan and Duncan, 1955). In biomonitoring, *PMA* is used to measure the similarity between a monitored community and a reference community. The statistical properties of this measure, such as bias, variance and confidence intervals, depend on the model used for the counts in the samples, e.g. multinomial, log-series or negative binomial distribution. For simplicity, we consider the multinomial case. Koskela et al. (2007) presented the expected value and variance of the index, known as apportionment index (AI) in the sawmill industry, using a normal approximation. Ransom (2000) approximated the asymptotic distribution of $1 - PMA$, known as dissimilarity index (D) in sociology, using the normal distribution and the delta method. However, both approaches require assumptions that are not valid in the biomonitoring context. In ecology, the measure is known as percentage similarity (PS, Renkonen, 1938), and Smith (1982) presented the expected value and variance of the index in the case when the reference profile is known. Smith and Zaret (1982) studied the expected value of the index when both profiles are unknown and estimated from samples. Smith (1982); Smith and Zaret (1982); Ricklefs and Lau (1980) presented also limited simulation studies on the index. As a detailed understanding of

the *PMA* index is needed, compared with previous works, we present exact formulas for the expected value and variance of the *PMA* beyond the works of Smith (1982) and Smith and Zaret (1982), calculated independently (Article III). Further, we study the asymptotics of the index when the reference profile is known. We also present an extensive simulation study relevant to the context of biomonitoring of macroinvertebrates. In the simulation study, we analyze the effects of sample size, the number of taxa, evenness and the true value of the index on the bias and variance of *PMA*.

In biomonitoring, reliable taxa identification is an important premise for index calculation. It is crucial to understand not just the statistical properties of different biological indices, but also how classification errors propagate into indices calculated from classified samples – especially when shifting from manual to automated identification. The effect of automated identification errors on indices has received minor attention in remote sensing. Wickham et al. (1997); Shao et al. (2001) and Chen et al. (2010) studied the effect of classification errors on landscape pattern indices, including Shannon's diversity (Shannon, 1948) and Simpson's diversity (Simpson, 1949), which are also often used in biomonitoring. However, these remote sensing studies focused mostly on indices specific to landscape patterns and concentrated more on the variation rather than bias caused by misclassifications. In this thesis, we present a simulation study exploring the effect of errors due to automated identification on a variety of biological richness, diversity, dominance, evenness and similarity indices (including the *PMA*) using several different classification methods with varying error rates. Further, we consider the reasons why certain types of indices are more sensitive to classification errors than others.

Chapter 2 gives an introduction to classification and problems related to high-dimensional data. RBA is presented as a possible solution to those issues with macroinvertebrate image data. In Chapter 3, the statistical properties of the *PMA* index are presented in two scenarios. Chapter 4 deals with error propagation and Chapter 5 summarizes the conclusions of this thesis.

Chapter 2

Automated identification

Automated identification, i.e. pattern recognition or classification, is a learning task. In supervised learning, a set of variables, called inputs or features, is used to predict one or more outputs. When the output is a qualitative, categorical variable, the prediction task is called classification. It is an exercise in function approximation with the objective to find a model or a classification rule that minimizes the expected prediction error – the expected loss (Hastie et al., 2009). The classifier is built based on data where the class labels are known. Often, the features are calculated from image data, as in our case.

There are thousands of different classification methods using different models for the data. To give a brief description of the different types of classification methods, Domingos (2012) divided classifiers by their representation into instances, hyperplanes, decision trees, sets of rules, neural networks and graphical models. Instance-based methods include such classifiers as K-Nearest-Neighbors (KNN, Hastie et al., 2009) and Support Vector Machines (SVM, Cortes and Vapnik, 1995). SVM can also be viewed as a hyperplane-based method, as it uses a hyperplane to separate the classes. Other hyperplane-based classifiers include e.g. Naïve Bayes (NB, John and Langley, 1995), Quadratic Discriminant Analysis (QDA, Duda et al., 2001; Hastie et al., 2009) and logistic regression (Hastie et al., 2009). Decision trees are nonmetric classifiers particularly useful with nominal features (Quinlan, 1986; Duda et al., 2001) but can be used for real-valued features as well. Decision trees can also be used to extract classification rules (Duda et al., 2001). Neural networks use nonlinear functions to obtain decision regions (Ripley, 1996), and graphical models include e.g. Bayesian networks (Nielsen and Jensen, 2009). This is just one coarse way of grouping the ever-growing number of classification methods available. In this thesis, we consider QDA models and their modifications.

2.1 Bayes decision rule

Let us consider a finite set of classes $\{\omega_1, \ldots, \omega_c\}$ and let $\mathbf{z} = (z_1, \ldots, z_r)$ be an r-dimensional random feature vector in the Euclidian space $\mathbb{R}^r$ with the class-conditional probability density function $f(\mathbf{z}|\omega_h)$ given the true class ω_h. Let π_h be the prior probability that $\mathbf{z}$ originates from class ω_h such that $\sum_{h=1}^{c} \pi_h = 1$. The posterior probability is given by

$$P(\omega_h|\mathbf{z}) = \frac{f(\mathbf{z}|\omega_h)\pi_h}{f(\mathbf{z})},$$

where $f(\mathbf{z}) = \sum_{h=1}^{c} f(\mathbf{z}|\omega_h)\pi_h$.

Let $\hat{h}(\mathbf{z})$ be a classification rule predicting the class from which $\mathbf{z}$ originates. Following the presentations in Duda et al. (2001, chapt. 2.2) and Hastie et al. (2009, chapt. 2.4), let $L\big(\omega_h, \hat{h}(\mathbf{z})\big)$ be a loss function indicating the loss for predicting the class $\hat{h}(\mathbf{z})$ when the observation belongs to class ω_h. The most common choice for a loss function is 0–1 loss:

$$L(\omega_h, \omega_{h'}) = \begin{cases} 0, & \omega_{h'} = \omega_h \\ 1, & \omega_{h'} \neq \omega_h \end{cases} .$$

Now, the conditional expected loss, i.e. risk, for prediction $\omega_{h'}$, is given by

$$R(\omega_{h'}|\mathbf{z}) = \sum_{h=1}^{c} L(\omega_h, \omega_{h'}) P(\omega_h|\mathbf{z})$$

and the overall risk associated with the classification rule $\hat{h}(\mathbf{z})$ is

$$R = \mathbb{E}_{\mathbf{z}} \left[\sum_{h=1}^{c} L\big(\omega_h, \hat{h}(\mathbf{z})\big) P(\omega_h|\mathbf{z}) \right]. \tag{2.1}$$

To minimize the overall risk, it suffices to minimize it pointwise by choosing

$$\hat{h}(\mathbf{z}) = \text{argmin}_{\omega_{h'}} \left(\sum_{h=1}^{c} L\big(\omega_h, \omega_{h'}\big) P(\omega_h|\mathbf{z}) \right).$$

For the 0–1 loss, minimizing Equation (2.1) simplifies to choosing the class with the highest posterior probability because now, for a fixed $\mathbf{z}$,

$$\hat{h}(\mathbf{z}) = \text{argmin}_{\omega_{h'}} \big(1 - P(\omega_{h'}|\mathbf{z})\big).$$

This is known as the Bayes decision rule and it achieves optimal performance by minimizing the expected prediction error (Duda et al., 2001).

The Bayes decision rule can also be represented in terms of discriminant functions $g_h(\mathbf{z})$ by setting $g_h(\mathbf{z}) = P(\omega_h|\mathbf{z})$ (e.g. Duda et al., 2001, Chapt. 2.4). The discriminant functions may be transformed using monotonically increasing functions without changing the classification. Thus, choosing the class based on the highest discriminant function

$$g_h(\mathbf{z}) = \ln f(\mathbf{z}|\omega_h) + \ln \pi_h$$

is equivalent to basing the classification on the highest posterior probability and following the optimal Bayes decision rule (Duda et al., 2001). Of course, the problem is that the choice of a prior is not always straightforward and we seldom know the true conditional distribution $f(\mathbf{z}|h)$.

If the features follow a multivariate normal distribution, i.e. $\mathbf{z}|\omega_h \sim N(\boldsymbol{\mu}_h, \boldsymbol{\Sigma}_h)$, the discriminant functions are quadratic:

$$g_h(\mathbf{z}) \quad = \quad -\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu}_h)' \boldsymbol{\Sigma}_h^{-1} (\mathbf{z} - \boldsymbol{\mu}_h) - \frac{1}{2} \ln |\boldsymbol{\Sigma}_h| + \ln \pi_h \tag{2.2}$$

and the subsequent classifier is known as Quadratic Discriminant Analysis (QDA, Hastie et al., 2009). If the covariance matrices are assumed equal, $\mathbf{\Sigma}_h = \mathbf{\Sigma}\ \forall h$, the quadratic terms in Equation (2.2) disappear and the discriminant functions become linear (Duda et al., 2001). The Bayes classifier is then called Linear Discriminant Analysis (LDA, Hastie et al., 2009). In an even more simplified scenario where the features are assumed independent, $\mathbf{\Sigma}_h = diag(\sigma_{h1}, \ldots, \sigma_{hr})$, the classifier is known as Naïve Bayes (NB, John and Langley, 1995). The mean vector and covariance matrix in Equation (2.2) are estimated from data (see e.g. Hastie et al., 2009), which can hinder the optimal performance of the Bayes classifier if the data is not truly Gaussian. Actually, NB, LDA and QDA perform well on a large array of classification tasks (Langley et al., 1992; Michie et al., 1994) even though the normal assumption is rarely met. Friedman (1997) and Hastie et al. (2009) suggest this is due to the fact that the data can often only support simple decision boundaries, and because the Gaussian model produces stable estimates.

2.2 Evaluating classifier performance

The performance of a classification method can be evaluated with the expected prediction error, also known as the classification error (Hastie et al., 2009, Chapt. 7.2)

$$Err = \mathbb{E}_T[\mathbb{E}_{\mathbf{z},h|T}\big(L(\omega_h, \hat{h}(\mathbf{z}))|T\big)],$$

where T is a training set used to fit the classifier and obtain the predictor $\hat{h}(\mathbf{z})$. When compared to Equation (2.1), this expectation averages also over the training set T with fixed sample size n. With 0–1 loss function, the classification error is the expected proportion of misclassified observations in a new data.

The classification error can be approximated in many ways. The most common way to do this is to split the data into separate training and test sets (see e.g. Hastie et al., 2009). The training set is used to build the classifier and the portion of misclassified observations in the test set serves as an estimate of the classification error.

Another option is to use cross-validation. In K-fold cross-validation, the data is split into K equal parts. For the kth part, the other $K-1$ parts are combined to form the training set the classifier is trained with, and that classifier is used to predict the classes for the observations in the kth data split. This is repeated for all K data splits, and the final classification error is a combination of the K estimates of the prediction error (Hastie et al., 2009, Chapt. 7.10). Let $\hat{h}^{-k}(\mathbf{z})$ denote the classifier trained with the kth data split removed. Now, the estimate for the classification error is the proportion of misclassified observations in the entire data set:

$$\widehat{Err}_{CV}(\hat{h}) = \frac{1}{n}\sum_{i=1}^{n} L(\omega_{h_i}, \hat{h}^{-k(i)}(\mathbf{z}_i)),$$

where ω_{h_i} is the true class of the observation i and $\hat{h}^{-k(i)}$ denotes the classifier where the observation i belonged to the kth data split that was not used for training (Hastie et al., 2009, Chapt. 7.10).

A third alternative is a bootstrap estimate of the classification error. A bootstrap sample is a random sample drawn with replacement from the original data and having the same sample size as the original data. When we take a bootstrap sample of the data,

$$
\begin{aligned}
P(\text{observation } i \in \text{bootstrap sample}) &= 1 - (1 - 1/n)^n \\
&\approx 1 - e^{-1} = 0.632. \qquad (2.3)
\end{aligned}
$$

Thus, about a third of the data is left out of the bootstrap sample and can be used as a separate test data (Hastie et al., 2009, Chapt. 7.11). Taking many bootstrap samples of the data and only keeping track of predictions from bootstrap samples not containing that observation mimicks cross-validation and produces an estimate for the classification error. Let $\hat{h}^b(\mathbf{z})$ denote the classifier trained on a bootstrap sample b, $b = 1, \ldots, B$. Now, the leave-one-out bootsrap estimate of the classification error is

$$
\widehat{Err}^{(1)} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|C^{-i}|} \sum_{b \in C^{-i}} L(\omega_{h_i}, \hat{h}^b(\mathbf{z}_i)), \qquad (2.4)
$$

where C^{-i} is the set of indices of the bootstrap samples that do not contain the observation i, and $|C^{-i}|$ is the number of such samples (Hastie et al., 2009, chapt. 7.11). The leave-one-out bootstrap estimate of the classification error has been criticized for being biased upwards, and corrections for this bias have been proposed in Efron and Tibshirani (1997). However, we present it here as it is closely related to another bootstrap estimate of the classification error discussed in Section 2.4.

As the performance of a classifier may depend on several parameters, finding optimal parameter values introduces a new layer to the problem of classifier evaluation. To avoid underestimating the classification error, parameter optimization, i.e. model selection, should be done either with a separate validation data or within the training data (e.g. Hastie et al., 2009, Chapt. 7). If there is enough data, it can be split into three parts, one for training, one for validation and one for testing. The classifier is built on the training data, the optimal parameter values are chosen based on the classification error of the validation data, and the final classification error is estimated from the test data. If the data is limited, it can be split into two parts, one for training and one for testing. Then the optimal parameter values for the classifier can be obtained based on the cross-validation error or the bootstrap error of the training data. Again, the final classification error is estimated from the test data.

2.3 Bias–variance decomposition

With a quantitative response variable and squared error loss, the bias–variance decomposition of the prediction error is easily understood due to its additive nature. In classification, the decomposition is not as simple. In fact, even the definitions of bias and variance vary in the classification framework, and there have been several different attempts to understand the relation of bias and variance to classification error in the case of 0–1 loss and two classes. Kong and Dietterich (1995); Kohavi and Wolpert (1996) and Breiman (1996a) derived additive bias-variance decompositions for the classification error, while Tibshirani (1996) and Friedman (1997) gave formulations where the interaction of bias and variance is non-linear and multiplicative. Friedman (1997) defined variance in the classification framework as the stability of a classifier, i.e. how much the predictions of a classifier might vary due to different

training sets. For the definition of bias, Friedman (1997) discussed the estimation bias similar to the case of the squared error loss, and the boundary bias referring to the bias in the estimation of the decision boundaries. He stated that in all of these decompositions of Kong and Dietterich (1995); Kohavi and Wolpert (1996); Breiman (1996a); Tibshirani (1996) and Friedman (1997), variance usually dominates bias, i.e. low variance is often more important for accurate classification than low estimation bias. However, Friedman (1997) noted that the effect of variance depends on the sign of the boundary bias. Domingos (2000) presented unified definitions for bias and variance that can be used with other loss functions and generalized the bias–variance decomposition of 0–1 loss to multi-class classification. He found that low variance becomes increasingly important as the number of classes increases.

The bias–variance decomposition is closely related to the problem of overfitting. If the classifier is sensitive to changes in the training data, it can result in poor generalization. Overfitting a classifier to the training data reduces bias but increases variance (Duda et al., 2001), which is the dominating factor in the classification error. With high dimensional data, avoiding overfitting becomes even harder because the training data covers only a fraction of the input space (Domingos, 2012).

2.3.1 Curse of dimensionality

Classification becomes increasingly more difficult as the dimension of the feature space grows. The problem of high-dimensional data was coined by Bellman (1961) as the curse of dimensionality. Classification methods that work well in lower dimensions cannot be used in higher dimensions because there is not enough information in the data to estimate the models accurately (Duda et al., 2001; Hastie et al., 2009). The limited amount of data can result in singular covariance matrices, overfitting and high variance. Some possible solutions to this are dimension reduction, i.e. reducing the number of features, assuming that the features are independent, or regularization. For example, QDA models are too complex to fit if the number of features exceeds the number of observations per class as the class conditional covariance matrices become singular (Friedman, 1989; Hastie et al., 2009). Friedman (1989) stated that even if the covariance matrices could be estimated, the estimates would be extremely unstable giving rise to high variance. He therefore proposed Regularized Discriminant Analysis (RDA), which shrinks the class-conditional covariance matrices of QDA towards the common covariance matrix of LDA. Though, with a complex data, also RDA can suffer from computational accuracy problems (Article II).

One common way to improve the classification accuracy by reducing variance is to build ensemble classifiers. Two popular methods for building ensembles are bagging and boosting. In bagging, i.e. bootstrap aggregating, the ensemble is built by taking bootstrap samples of the training data and fitting a classifier on each of these samples (Breiman, 1996b). Typically, the classifiers in the ensemble are of the same type, e.g. decision trees (Quinlan, 1986), and only the parameter values vary due to the different training samples (Duda et al., 2001). The final class decision can be obtained by a majority vote or by averaging the predicted class probabilities over the ensemble (Breiman, 1996b). Bagging improves classification accuracy because it greatly reduces variance while only slightly increasing bias (Domingos, 2012). However, in order to improve classification accuracy, bagging requires an unstable base classifier, i.e. a classifier with high variance (Breiman, 2001). For example, NB is a very stable (although often biased) classifier and cannot be improved with bagging (Bauer and Kohavi,

1999). The idea in bagging is to approximate the underlying distribution of the data that produced the training set (Breiman, 1996b). This makes bagged ensembles quite robust to noise, i.e. incorrect labels of training observations (Dietterich, 2000). Another benefit with bagging is that it gives a built-in estimate of the classification error (Eq. 2.5) without the need for a separate test set (Breiman, 2001).

In boosting, the ensemble is a sequence of classifiers. The next classifier is trained on observations that were found the most informative based on the previous classifier (Duda et al., 2001, Chapt. 9.5). The most popular boosting algorithm is called Adaboost (Freund and Schapire, 1997). Adaboost, or adaptive boosting, gives weights to each observation in the training data. These weights are adjusted after each classifier is built. The weights are increased for observations the previous classifier misclassified and decreased for those that were classified correctly (Duda et al., 2001, Chapt. 9.5). The final prediction is a weighted vote over the ensemble and the weights for the votes depend on the classification error of each classifier (Hastie et al., 2009, Chapt. 10.1). For Adaboost, the base classifier need not be as unstable as for bagging because Adaboost can make larger changes in the training data (Dietterich, 2000). However, unlike bagging, Adaboost is sensitive to noise in the training data (Dietterich, 2000).

Variance can also be decreased by introducing randomness into the classifier since this encourages diversity which can further improve the accuracy of ensembles (Dietterich, 2000). Breiman (2001) combined bagging and randomization in Random Forest (RF) which is an ensemble of random decision trees. Breiman noted that the classification error of a forest depends on the strength of the trees as well as on the correlation between them. Randomness reduces the correlation between the base classifiers and therefore improves the accuracy.

2.4 Contributions

We propose a novel classifier called Random Bayes Array (RBA, Articles I and II) by applying both bagging and randomness to quadratic discriminant analysis, mimicking the RF approach of Breiman (2001). With high-dimensional data, QDA becomes a very unstable classifier and as such can benefit from these variance reduction techniques. The randomness is introduced through randomly selecting a subset of the features for each QDA in the ensemble. This not only reduces the variance of the ensemble but also enables the use of QDA with high dimensional feature space since it reduces the chance of obtaining ill-conditioned covariance matrices (Article II). In addition to the majority vote and averaging posterior probabilities over the ensemble, we explore weighted voting. In weighted voting, the vote of each base classifier in the ensemble is weighted with its corresponding posterior probability, giving higher weight for more confident votes (Article II).

As RF and RBA are ensembles of random classifiers trained on bootstrap samples of the original training data, one third of the original training observations are left out of each bootstrap sample (see Eq. (2.3)). Breiman (2001) named these out-of-bag or oob observations which can be used to construct an internal estimator for the classification error. Let $\hat{h}'(\mathbf{z})$ be the predictor pooled over those classifiers in the ensemble where $\mathbf{z}$ is out-of-bag. Now, the oob estimate for the classification error is

$$\widehat{Err}_{oob} = \frac{1}{n}\sum_{i=1}^{n} L(\omega_{h_i}, \hat{h}'(\mathbf{z}_i)). \tag{2.5}$$

One standard approach to evading the problems associated with high-dimensional data is feature selection, i.e. selecting a subset of the original feature space. Similarly as in RF, the out-of-bag observations in RBA can be used to evaluate feature importance. Breiman (2001) proposed to explore the importance of a feature z_j, $j = 1, \ldots, r$, by permuting the values of the feature. If a feature z_j is an integral part of successful classification, permuting its values in the oob-observations should increase the out-of-bag error in Equation (2.5). The importance of the feature z_j is given by

$$\begin{aligned} imp(z_j) &= \frac{1}{B}\sum_{b=1}^{B}\Bigg[\frac{\#\{\text{Correct classifications of } b\text{'s oob-observations}\}}{\#\{b\text{'s oob-observations}\}} \\ &\quad - \frac{\#\{\text{Correct classifications when } z_j \text{ values of } b\text{'s oob-observations permuted}\}}{\#\{b\text{'s oob-observations}\}}\Bigg], \end{aligned} \tag{2.6}$$

where b denotes the different bootstrap samples. Feature importance is a highly useful measure since it takes into account each feature individually as well as multivariate interactions with other features (Strobl and Zeileis, 2008).

Breiman and Cutler (2008) suggested to calculate standardized importance scores that, based on the central limit theorem, are assumed asymptotically standard normal under the null hypothesis of zero feature importance. Thus, they could be used for feature selection. However, Strobl and Zeileis (2008) demonstrated in simulation experiments that the power of the test does not increase with the relevance of the feature but with the number of random decision trees in RF – a parameter chosen by the researcher. Therefore, the test should not be used to identify the most important features.

Instead of selecting a subset of features based on the ranking of feature importance or the standardized importance score test, in Article II we propose to utilize the feature importances as weights when drawing the random subset of features for each QDA in the ensemble. This way, all the features are included and no information is lost, yet the most important features will be used more often. The weights are obtained from the raw feature importance as follows:

$$w(z_j) = \frac{imp(z_j)}{\sum_{j'=1}^{r} imp(z_{j'})}.$$

The performance of RBA was tested and compared to other classification methods with macroinvertebrate image data in Articles I, II and IV. In article I, the data comprised of 15 features and 1350 observations belonging to 8 classes. With this data, RBA matched the performance of QDA and RF. All three classifiers produced classification errors of 7–8 %. Article II used a much more complex data of 64 features and 6814 observations belonging to 35 classes. In Article II, RBA yielded the lowest classification error out of all tested classifiers, comparing favourably to such classifiers as LDA, RF, SVM and neural networks. We achieved the lowest classification error (18.8 %) when basing the predictions on the averaged posterior probabilities and using feature importance as weights for the features. The three pooling methods, i.e. using majority vote, average posterior probabilities or posterior weighted voting, yielded similar accuracies but using feature importance weights produced a clear improvement in the classification accuracy. We also compared RBA to other Bayesian classifiers (QDA, LDA and NB) and found it to outperform them in small training sample scenarios – even when the data was multivariate normal (Article II). With real image data deviating from the Gaussian assumption, RBA gave higher classification accuracy than QDA even with larger

training samples (Article II). In Article IV, the data of 35 classes was modified slightly to include 32 classes (see Article IV for details) and classified with a wider range of classification methods. RBA produced a similar classification error to that in Article II, ranking fifth among the classifiers. The proportion of data used for training was higher in Article IV than Article II, and the benefit of the ensemble approach in RBA is precisely that it works well even with limited training data. Other classifiers may outperform RBA when there is more training data available but RBA's performance endures cuts in data size.

Chapter 3

Percent Model Affinity Index

3.1 On biological indices

In biomonitoring, ecological status assessments of surface waters are based on scores of indices calculated from samples of biological indicator groups, such as benthic macroinvertebrates. Different indices use different aspects of taxa abundancies identified from samples and aim to assess the diversity of a monitored community or to determine whether two communities are similar to each other.

The biological diversity of a community is a combination of its species richness and the evenness of those species (Krebs, 1999). Consequently, diversity indices can be divided into groups based on which aspects of diversity they cover the most. Diversity can also be viewed as the average rarity within a community (Patil and Taillie, 1982). Some popular and widely used diversity, richness, evenness and dominance indices are presented in Article IV.

Similarity indices are an important tool in biomonitoring to measure similarity over time and space. They are often used to compare a monitored community to an ideal reference community considered to be either in a natural state or largely undisturbed by human action. If a monitored community is very similar to the reference community, the monitored waterbody is estimated to have good ecological status. If the similarity between the monitored community and the reference community is low, the monitored waterbody is in poor ecological condition, indicating a need for mitigation measures. Besides measuring the similarity between two communities, similarity indices can also be used to detect changes in community composition over time. This is done by calculating the similarity of two samples taken from the same community at different times. Article IV presents six well-known similarity indices, including the Percent Model Affinity (PMA) index.

The accuracy of the ecological status assessment depends strongly on the statistical properties, i.e. bias and variation, of the chosen indices. As for many of the diversity indices, it is difficult to give reliable confidence intervals for most similarity measures (Ricklefs and Lau, 1980): The statistical properties of most similarity indices have only been evaluated with simulation procedures, such as jackknife and bootstrap experiments. The statistical properties of an index arise from the underlying distribution assumed for the samples. Typical choices include multinomial distribution (Smith, 1982; Smith and Zaret, 1982), Poisson distribution (Farley and Johnson, 1985), log-series distribution (Wolda, 1981) and negative binomial distribution (Venrick, 1983; El-Shaarawi, 2006). Here, we limit the study on biological indices to the case of multinomial distribution. The statistical properties of the PMA index, assuming

multinomial distribution for the samples, are presented in the following sections.

3.2 Statistical properties of the Percent Model Affinity index

The Percent Model Affinity (PMA) index was introduced to the aquatic biomonitoring field by Novak and Bode (1992). In their article, Novak and Bode proposed PMA as a new index of stream water quality that quantitatively compares the community composition of a monitored community with an expected, model community. In PMA, the proportional taxa distributions in the monitored community and the model community are compared using Percentage Similarity (PS), originally proposed by Renkonen (1938). This makes the PMA index an application of PS. Mathematically, PS is defined as follows: Let us consider a finite set of classes (e.g. taxonomic groups) $\{\omega_1, \ldots, \omega_c\}$ and let $\mathbf{p} = (p_1, ..., p_c)$ and $\mathbf{q} = (q_1, ..., q_c)$ be two probability mass distributions over Ω, such that $\sum_{h=1}^{c} p_h = 1$ and $\sum_{h=1}^{c} q_h = 1$. Then,

$$PS = 1 - \frac{1}{2} \sum_{h=1}^{c} |p_h - q_h|. \tag{3.1}$$

In the case of perfect similarity, when $\mathbf{p} = \mathbf{q}$, the PS is equal to 1. In the case of perfect dissimilarity, i.e. $p_h q_h = 0$, $h = 1, \ldots, c$, the PS is equal to 0. In all other cases, $0 < PS < 1$. For the PMA index, $\mathbf{p}$ represents the taxa distribution of the monitored community and $\mathbf{q}$ represents the taxa distribution of the model community.

Generally, the probability profiles $\mathbf{p}$ and $\mathbf{q}$ are unknown, and therefore it is necessary to consider an estimator that can be derived from empirical data. Following the presentation of Novak and Bode (1992), the taxa distribution of the monitored community is estimated from a sample and the model community is obtained as the average taxa distribution from reference (sensu Stoddard et al., 2006) stream samples. For the monitored community, let $\mathbf{X} = (X_1, \ldots, X_c)$ be a random sample drawn from a multinomial distribution $Multinom(n, \mathbf{p})$. Then, the maximum likelihood estimator for p_h is $\hat{p}_h = X_h/n$ for each $h = 1, \ldots, c$. For the reference community, let $\mathbf{Y} = (Y_1, \ldots, Y_c) \sim Multinom(m, \mathbf{q})$. Respectively, the maximum likelihood estimator for q_h is $\hat{q}_h = Y_h/m$. This leads to three possible estimators for the PS: If the model community is assumed known,

$$\widehat{PS} = 1 - \frac{1}{2} \sum_{h=1}^{c} |\hat{p}_h - q_h|, \tag{3.2}$$

or if both probability profiles are estimated from one sample,

$$\widehat{PS} = 1 - \frac{1}{2} \sum_{h=1}^{c} |\hat{p}_h - \hat{q}_h|. \tag{3.3}$$

The third estimator, where $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ are obtained from multiple samples, is discussed in Article III.

Percentage similarity is a popular measure of similarity between two probability profiles and has applications in many fields, such as the sawmill industry (Bergstrand, 1989), ecology (Renkonen, 1938) and sociology (Duncan and Duncan, 1955). The PMA index has gained popularity and is widely used in the biomonitoring of aquatic ecosystems (see e.g. Passy and Bode, 2004; Alahuhta et al., 2009; Kauppila et al., 2012). Yet, in the field of aquatic

biomonitoring, the statistical properties of the estimators in Equations (3.2) and (3.3) have received very little attention even though these properties may have a significant impact on the ecological status assessment of waterbodies. In the other fields of study, some theoretical results exist on the statistical properties of percentage similarity.

3.2.1 Known reference profile

In the sawmill industry, PS is used to measure the fit between the demand distribution and the observed distribution of logs (Bergstrand, 1989). In this field of application, the index is referred to as Apportionment Index and it is a measure of the successfulness of the harvesting operation. Using multinomial distribution for the samples, Koskela et al. (2007) presented the expected value and variance of the estimated index in Equation (3.2) using a large-sample normal approximation. This approach assumes that the observed frequencies of log size classes follow a known demand distribution, i.e. $\mathbf{X} \sim Multinom(n, \mathbf{q})$, fixing the true value of the index to be 1. The expected value was found to approach 1 with increasing the sample size, making the index asymptotically unbiased. Koskela et al. (2007) noted the variance to be a function of the sample size as well, decreasing with increases in the sample size. Besides sample size, they found the size and shape of the demand distribution $\mathbf{q}$ to affect the expected value of the estimator $\widehat{PS}$ in Equation (3.2). Koskela et al. (2007) also derived a lower tolerance limit for the index that can be used to determine when deviations from the demand distribution are statistically significant.

Percentage similarity is also used in the field of ecology to measure niche breadth by comparing resource use with resource availability (Feinsinger et al., 1981). Smith (1982) derived exact formulas for the expected value and variance of the estimator $\widehat{PS}$ in Equation (3.2) when $\mathbf{X} \sim Multinom(n, \mathbf{p})$ and $\mathbf{q}$ is fixed. Smith further derived an approximate formula for the variance using the delta method (Seber, 1973). However, these results are not reliable as the transformation function (Eq. 3.2) in the delta method is not differentiable.

3.2.2 Estimated reference profile

In sociology, PS is utilized to calculate Dissimilarity index ($D = 1 - PS$, Jahn et al., 1947; Duncan and Duncan, 1955) which is a common tool to measure segregation. Having both $\mathbf{p}$ and $\mathbf{q}$ estimated from single samples following the multinomial distributions $Multinom(n, \mathbf{p})$ and $Multinom(m, \mathbf{q})$, Ransom (2000) used a normal approximation for large samples to derive the asymptotic distribution of D. The variance for D was calculated using the delta method (Seber, 1973). As with Smith (1982), the results obtained with the delta method are not reliable as the transformation function in Equation (3.3) is not differentiable.

Smith and Zaret (1982) extended the results of Smith (1982) to the more general framework where both probability profiles in PS are estimated from single samples. They studied the bias of the estimator $\widehat{PS}$ in Equation (3.3) using exact formulas from Goldstein and Wolf (1977) in order to calculate the expected value of the estimator. Smith and Zaret (1982) found the bias of $\widehat{PS}$ to decrease with increasing sample sizes and to increase with increasing the number of classes and evenness of the probability profiles $\mathbf{p}$ and $\mathbf{q}$. They also claimed that the true value of the index has no effect on the bias.

Ricklefs and Lau (1980) studied the bias and variation of the estimator $\widehat{PS}$ in Equation (3.3) with simulation experiments. In contrast to the claims of Smith and Zaret (1982),

simulating samples from two multinomial distributions and estimating $\mathbf{p}$ and $\mathbf{q}$ from the samples, Ricklefs and Lau (1980) found the bias to be a problem when the true value of PS approaches 1. They noticed both the bias and variance of the estimated index to decrease with increases in sample size and to increase with increasing the number of classes. In addition, Ricklefs and Lau (1980) argued that the bias and variance are not affected by changes in evenness.

3.2.3 Contributions

Although some theoretical results exist on the statistical properties of percentage similarity, not all of the results are reliable or valid in the context of aquatic biomonitoring. First, for the case of a known reference profile, the results of Koskela et al. (2007) assume that the monitored sample follows a known reference distribution. In biomonitoring, this is not a reasonable assumption for samples from streams impacted by human action. For the case of an estimated reference profile, the assumption in Ransom (2000) that all taxa proportions in the monitored and the reference community must be unequal is not reasonable for samples from streams that are in reference condition. There is a need for a general framework for all possible taxa profiles $\mathbf{p}$ and $\mathbf{q}$ and all three possible estimators of PS.

In Article III, we derive and present the exact formulas for the expected value and variance of the estimators in Equations (3.2) and (3.3), calculated beyond and independently of the results of Smith (1982) and Smith and Zaret (1982). In addition, we explain how the results for the latter estimator can be extended to multiple samples. To our knowledge, the exact variance for Equation (3.3) has not been studied previously. As another novel contribution, we provide asymptotic results for the expected value of the estimator in Equation (3.2). We find the estimator to be asymptotically unbiased with the bias approaching zero faster if the probability profiles $\mathbf{p}$ and $\mathbf{q}$ are unequal (Article III).

Using simulated taxa profiles $\mathbf{p}$ and $\mathbf{q}$, we study the effects of sample size, number of taxa, evenness of the taxa profiles and the true value of the index on the bias and variance of the estimators in Equation (3.2) and (3.3). Our analyses support the results of Ricklefs and Lau (1980) and Smith and Zaret (1982) that increasing sample size decreases the bias and variance and increasing the number of taxa increases the bias (Article III). We concur with the findings of Smith and Zaret (1982) that increasing the evenness of the taxa profiles $\mathbf{p}$ and $\mathbf{q}$ increases the bias of the estimator in Equation (3.3). This is also true for the variance (Article III). In addition, we show that the true value of the PS index indeed does affect the bias and variance of the estimated index, as suggested by Ricklefs and Lau (1980). There is a steep increase in the bias as the true value of the index approaches 1. The results are similar for both estimators of the PS index (Eqs. (3.2) and (3.3)). In comparison with the previous analytical and simulation studies, we use parameter values more relevant to the field of aquatic biomonitoring where e.g. the number of taxonomic groups is very large (Article III).

Chapter 4

Error propagation

Measuring data quality has become increasingly important for bioassessment and should be done based on the ultimate uses of the data (Cao et al., 2003). One major source of uncertainty in the biomonitoring process chain are identification errors. As biological indices are calculated from samples that are identified to taxa, identification errors may affect the statistical properties of the indices. In addition to the sampling distribution of biological indices, it is crucial to understand how identification errors propagate into the indices and into the ecological status evaluation of waterbodies. This error propagation depends not only on the total accuracy of the identification but also on the probabilities of different misclassifications.

The important issue of error propagation has received some attention in the field of biomonitoring. The European Union funded STAR project (Furse et al., 2006) aimed at identifying different sources of uncertainty and their effects on biological indices and the ecological status assessment of waterbodies. In the project, 6 countries were audited for identifying macroinvertebrates. In the audit, Haase et al. (2006) studied the identification differences between the primary analysts and the auditors on a group of indices (e.g. Shannon's diversity Shannon and Weaver, 1963) and found statistically significant differences in the index values using Wilcoxon's test (Wilcoxon, 1945). Stribling et al. (2008) reported on quality control of biomonitoring of macroinvertebrates in the USA, auditing 8 laboratories performing taxa identification. In this study, the identifications of the laboratories and the auditors differed on average 21 % leading to different ecological status classes in 22.2 % of the audited samples. In a similar auditing of the biomonitoring of macroinvertebrates in Germany, Haase et al. (2010) found a 33.8 % difference in identifications between the primary analysts and the auditors, leading to differences in index values. The final ecological status class was determined as the worst-case-scenario out of three indices and was different for 8/50 samples (Haase et al., 2010).

In biomonitoring, the decreasing funding drives the search for alternative identification methods with lower costs, such as automated classification (Articles I & II, Blaschko et al., 2005; Culverhouse et al., 2006; Lytle et al., 2010; Kiranyaz et al., 2011; Joutsijoki et al., 2014) and citizen-science monitoring (Dickinson et al., 2012). While being cost-effective, these approaches may introduce additional bias and variation to indices calculated from identified samples. In fact, Gardiner et al. (2012) studied how identification errors of lady beetles propagated into Menhinick's index of species richness (Magurran, 2004) and Simpson's diversity index (Simpson, 1949) in citizen-science monitoring programs. Both indices were found biased upwards due to identification errors. The effect of automated classification on indices has

mainly been studied in the field of remote sensing, concentrating mostly on additional variation due to misclassifications. Wickham et al. (1997) simulated automatically classified maps using information on spatial autocorrelation and the confusion matrix - a matrix containing the probabilities of different classification decisions. They calculated bias due to classification errors for four landscape pattern indices and approximated the confidence intervals of these indices using t-distribution. Shao et al. (2001) studied the effects of 23 different classifications of a thematic map on 18 landscape pattern indices calculated from the map. They noted that classifiers with similar total accuracy resulted in varying index values, indicating the effect of the shape of the confusion matrix. Using the same data, Shao and Wu (2004) found that the mean index values did not differ very much between classifiers of different total accuracy but that variation in the index values was greatly affected by the amount of classification errors. They also discovered certain landscape pattern indices to be more sensitive to misclassifications than others.

The probabilities of different classification decisions are stored in a confusion matrix with the rows representing the predicted classes and the columns representing the true classes of the observations. Therefore, if a gold standard training set is used, then the confusion matrix can be utilized to correct the bias due to classification errors in the predicted proportions or frequencies (see e.g. Card, 1982; Hay, 1988; Fortier, 1992; Buonaccorsi, 2010). Studying the effect of misclassifications on proportion estimation, Healy (1981) gave a formula for the mean square error (MSE) of the predicted proportions based on the confusion matrix. Healy (1981) also advised how to build a classifier that minimizes the MSE of the estimated proportions due to misclassifications. In an effort to approximate the uncertainty caused by classification errors, Hess and Bay (1997) used a confusion matrix correction and bootstrapping to construct confidence intervals for several landscape pattern indices. This study included Simpson's diversity and Shannon's diversity (Shannon and Weaver, 1963) also used in biomonitoring. Of course, for any confusion matrix correction to work, we must have a good estimate for the confusion matrix, which is not always possible.

4.1 Contributions

In Article IV, we study the effects of automated taxa identification errors on a wide range of biological indices describing richness, diversity, dominance, evenness and similarity (Table 4.1). In that article, we study how different classification methods introduce bias and variation into biological indices. We also look at the reasons why certain indices are more sensitive to misclassifications than others.

Using a multinomial model, we simulate macroinvertebrate samples from 12 stream types, estimating the taxa probabilities from monitoring data collected by the Finnish Environment Institute. For each stream type, we simulate both reference and non-reference samples, having all together 24 different taxa profiles for the multinomial model. To simulate classified samples, we use multinomial distribution, where the taxa probabilities are mixed with a confusion matrix. The confusion matrices for 11 classifiers are estimated from image data described in Articles II and IV.

Table 4.1: Biological indices and their ranges (Article IV).

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	$C_\lambda = \frac{2\sum_{h=1}^{c} X_h Y_h}{(D_x + D_y) nm}$	0	1
14) Jaccard similarity coefficient	$J = \frac{S_{xy}}{S_x + S_y - S_{xy}}$	0	1

To evaluate the bias in index values caused by classification errors, we use proportional bias

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

where I is a general notation for an index with correct classification and $\tilde{I}$ a general notation for the same index with possible classification errors. Chen et al. (2010) used similar proportional bias to represent bias in landscape pattern indices in remote sensing. There are other possibilities for scaling the bias (see e.g. Shao and Wu, 2004) but they usually depend on the class probabilities. We scale the bias with the range of each index to make it comparable between indices with varying ranges, different classifiers and varying taxa distributions (Article IV). The %bias serves a measure of the biological significance of the bias. Similarly, the effect of misclassifications on the standard deviation of biological indices can be evaluated with

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

In contrast to the findings of Shao and Wu (2004) with landscape pattern indices, we find that classification errors cause bias in biological indices but leave the variation mostly unaffected (Article IV). The results from the simulation study suggest that richness indices based on the number of observed taxa are the most vulnerable to classification errors because even a single misclassification can cause overestimation (Article IV). Interestingly, these indices are also sensitive to the dimensions of the training data used to train the classifiers: If the training data has a larger number of taxa than the sample from which we estimate the number of observed species, the confusion matrix tends to spread out the observations so that species richness will be biased upwards. Other presence/absence-based indices, where the number of the observed taxa affects both the denominator and the numerator of the index, are less affected by classification errors. Indices based on the observed proportions of taxa are quite robust to misclassifications except if the most common taxa are poorly classified (Article IV). We find Simpson's diversity to be the least biased diversity index and Euclidian similarity, *PMA* index and Morisita-Horn index to be the least biased similarity indices. Similarity indices measure the similarity between two conditions – in our study the reference and non-reference communities of a fixed stream type. In our simulations, we explore a scenario where the reference sample is known, i.e. classified correctly, and a scenario where both the reference and the non-reference sample can contain misclassifications. We find the similarity indices to be biased downwards if the reference sample is known and biased upwards if also it contains classification errors (Article IV). We suspect this is due to the fact that our chosen classifiers tend to increase the entropy of the samples. As the entropy increases, the samples for reference and non-reference condition become increasingly more similar, thus overestimating the similarity indices (Article IV).

As many of the biological indices are based on estimated taxa abundances or proportions, it might be worthwhile to explore correction methods based on the confusion matrix. In Article I, we study the bias in proportion estimation due to classification errors and apply a correction method based on the inverse of the confusion matrix (Hay, 1988). The correction yields promising results with the RBA and a C4.5 decision tree classifier (Quinlan, 1993) – although in Article I the data consists of only 8 taxa making the estimation of the confusion

matrix a considerably easier task than in the case of larger image data, as in Articles II and IV. The effect of the confusion matrix correction on biological indices is not considered here.

Chapter 5

Discussion

The aim of this thesis was to provide tools to improve the cost-efficiency and accuracy of freshwater benthic macroinvertebrate biomonitoring. To achieve this objective, we studied existing and developed new automated classification methods suitable for challenging macroinvertebrate image data, and strove to understand the different sources of errors and bias that introduce uncertainty to biological indices calculated from macroinvertebrate samples.

Shifting from manual to automated identification can greatly reduce the costs of macroinvertebrate biomonitoring. For example, automated classification of a sample of 3400 individuals can be over 300 times faster compared with manual expert identification, and even more if we take into account the time it takes to train an expert or a classifier. The shift to automated identification requires classification methods with high accuracy. The results for RBA (and other classifiers) are very promising, although the image data used in this thesis contained only 35 taxa and the true number of different taxonomic groups for a sample can be much higher. According to our results, RBA performed well with complex data where the number of classes and features is high and the observations per class are limited. As the imaging systems for macroinvertebrates develop further, other classifiers that benefit from having a large training data may outperform RBA. However, due to bagging and random feature selection, RBA remains a good option for classifying challenging data that suffers from the curse of dimensionality. Furthermore, we only used equal priors for the classes in RBA in this thesis. As with other Bayesian classifiers, we would expect RBA's performance to benefit from more informative priors.

The accuracy of ecological status assessment of water bodies can be improved by utilizing knowledge on the different sources of uncertainty in the biomonitoring process chain. To achieve this, we derived the mean and variance of the estimated *PMA* index used in Finnish biomonitoring, and explored the effects of sample size, the number of taxa, evenness and the true value of the index on these statistics. Our results increase the understanding of e.g. whether the *PMA* scores for two different streams are comparable or not. Namely, the most disconcerting result for the properties of the *PMA* index was the amount of bias as the true value of the index approaches its maximum. The *PMA* index is strongly biased downwards for samples from streams in reference condition. In practice, the sample size needs to be much larger for streams in reference condition to achieve as low bias as for streams in non-reference condition. In this thesis, we limited the study on biological indices to the case of multinomial distribution. In nature, however, the variation in macroinvertebrate samples is much greater, and to understand the behavior of the *PMA* index in those circumstances, we need to develop

a more realistic model which takes overdispersion into account.

Besides the sampling distributions of biological indices, another source of uncertainty arises from identification errors. In this dissertation, we studied how automated classification errors affect a variety of common biological indices. We found presence/absence-based indices to be very sensitive to misclassifications in the sense of bias but not standard deviation, and proportion-based indices to be the most robust. Naturally, the indices have to be chosen such that they measure the subject of interest from a community. However, if an index is highly biased, it is not a good measurement of the phenomenon altogether. Choosing another distribution instead of the multinomial model for the samples could result in more variation. However, in Article IV we were only interested in bias and extra variation caused by classification errors.

For future research, besides different models for the counts, it would be interesting to consider how different correction methods based on the confusion matrix can decrease the bias in biological indices. Studies comparing the performance of automated classifiers and human experts from the identification of samples to the calculation of biological indices and, finally, the ecological status assessment of water bodies would also be very advantageous as biomonitoring is slowly moving towards automation.

Summary of original publications

Article I introduces the RBA classifier with a small preliminary macroinvertebrate image data set. RBA – called Random Bayes Forest in this paper – is compared with several other popular classification methods and found to work well, although its potential lies with more problematic data. The article studies the effect of classification errors on sample proportions estimated from predicted classes and briefly discusses the confusion matrix correction.

Article II gives a thorough presentation of the RBA classifier. Its origins in random forests and quadratic discriminant analysis are discussed. RBA's benefits with complex data suffering from the curse of dimensionality are considered and several ways of using the produced posterior probabilities are examined, including the possibility of assessing feature importance and using it as weights for better classification results. RBA is compared with a variety of classification methods with a large macroinvertebrate image data and found to yield the lowest error rate.

In biomonitoring, the identified samples are used to calculate biological indices that are then used to assess the ecological status of waterbodies. One popular measure is the *PMA* index which compares a monitored taxa profile with a reference profile. In article III, the statistical properties of the *PMA* index are studied in two different scenarios, 1) when the reference profile is assumed to be known and 2) when the reference profile is estimated. Theoretical formulas for expected value and variance are given in both scenarios and asymptotics are considered when the reference profile is assumed to be known. The effect of sample size, true value of the index, evenness and number of classes on the bias and variance of the *PMA* index are studied.

Article IV studies the effect of classification errors on biological indices, such as the *PMA* index, calculated from classified samples. Error propagation is a problem with both manual and automated identification but in automated identification the errors are more systematic and the confusion matrix of a classifier can be an indicator of how the classification errors propagate into indices calculated from the identified samples. A simulation study is performed to assess the amount of bias resulting in different indices using different classifiers.

Bibliography

Alahuhta, J., Vuori, K.-M., Hellsten, S., Järvinen, M., Olin, M., Rask, M., and Palomäki, A. (2009). Defining the ecological status of small forest lakes using multiple biological quality elements and palaeolimnological analysis. *Fundamental and Applied Limnology / Archiv fr Hydrobiologie*, 175(3):203–216.

Bauer, E. and Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine Learning*, 36(1–2):105–139.

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton University Press.

Bergstrand, K.-G. (1989). Fördelningsaptering med näroptimalmetoden reviderad version [bucking to order with a close-to-optimal method revised version]. *Forskningsstiftelsen Skogsarbeten*, 1989-12-11. 11 p. (In Swedish).

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. (1996a). Arcing classifiers. Technical report, Statistics Department, University of California, Berkley.

Breiman, L. (1996b). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L. and Cutler, A. (2008). *Random forests - Classification manual*. http://www.math.usu.edu/∼adele/forests/.

Buonaccorsi, J. P. (2010). *Measurement Error: Models, Methods, and Applications*. CRC Press.

Cao, Y., Hawkins, C. P., and Vinson, M. R. (2003). Measuring and controlling data quality in biological assemblage surveys with special reference to stream benthic macroinvertebrates. *Freshwater Biology*, 48:1898–1911.

Card, D. H. (1982). Using known map category marginal frequencies to improve estimates of thematic map accuracy. *Photogrammetric Engineering & Remote Sensing*, 48(3):431–439.

Chen, X. H., Yamaguchi, Y., and Chen, J. (2010). A new measure of classification error: designed for landscape pattern index. *International Archives of Photogrammetry, Remote Sensing and Spatial Information Science*, 38(8):759–762.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 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. *Marine Ecology Progress Series*, 312.

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. *Frontiers in Ecology and the Environment*, 10:291–297.

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, 40(2):139–158.

Domingos, P. (2000). A unified bias-variance decomposition. *Proceedings of the 17th International Conference on Machine Learning*, pages 231–238.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

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

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern Classification*. John Wiley & Sons.

Duncan, O. D. and Duncan, B. (1955). A methodological analysis of segregation indexes. *American Sociological Review*, 20(2):210–217.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

El-Shaarawi, A. H. (2006). *Encyclopedia of Environmetrics*, chapter Negative Binomial Distribution. John Wiley & Sons.

Farley, R. and Johnson, R. (1985). On the statistical significance of the index of dissimilarity. *Proceedings of the Social Statistics Section*, pages 415–420. Washington DC: American Statistical Association.

Feinsinger, P., Spears, E. E., and Poole, R. W. (1981). A simple measure of niche breadth. *Ecology*, 62:27–32.

Fortier, J.-J. (1992). Best linear corrector of classification estimates of proportions of objects in several unknown classes. *The Canadian Journal of Statistics*, 20(1):23–33.

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139.

Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175.

Friedman, J. H. (1997). On bias, variance, 0/1–loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77.

Furse, M., Hering, D., Moog, O., Verdonschot, P., Johnson, R. K., Brabec, K., Gritzalis, K., Buffagni, A., Pinto, P., Friberg, N., Murray-Bligh, J., Kokes, J., Alber, R., Usseglio-Polatera, P., Haase, P., Sweeting, R., Bis, B., Szoszkiewicz, K., Soszka, H., Springe, G., Sporka, F., and Krno, I. (2006). The star project: context, objectives and approaches. *Hydrobiologia*, 566:3–29.

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. *Frontiers in Ecology and the Environment*, 10(9):471–476.

Goldstein, M. and Wolf, E. (1977). On the problem of bias in multinomial classification. *Biometrics*, 33(2):325–331.

Haase, P., Murray-Bligh, J., Lohse, S., Pauls, S., Sundermann, A., Gunn, R., and Clarke, R. (2006). Assessing the impact of errors in sorting and identifying macroinvertebrate samples. *Hydrobiologia*, 566:505–521.

Haase, P., Pauls, S. U., Schindehtte, 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. *Journal of the North American Benthological Society*, 29(4):1279–1291.

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

Hay, A. M. (1988). The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9(8):1359–1398.

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

Hess, G. R. and Bay, J. M. (1997). Generating confidence intervals for composition-based landscape indexes. *Landscape Ecology*, 12:309–320.

Jahn, J., Schmidt, C. F., and Schrag, C. (1947). The measurement of ecological segregation. *American Sociological Review*, 12(3):293–303.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann Publishers: San Mateo.

Joutsijoki, H. and Juhola, M. (2012). *Machine Learning and Data Mining in Pattern Recognition*, chapter DAGSVM vs. DAGKNN: an experimental case study with benthic macroinvertebrate dataset, pages 439–453. Springer.

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. *Ecological Informatics*, 20:1–12.

Kauppila, T., Kanninen, A., Viitasalo, M., Räsänen, J., Meissner, K., and Mattila, J. (2012). Comparing long term sediment records to current biological quality element data - implications for bioassessment and management of eutrophic lake. *Limnologica - Ecology and Management of Inland Waters*, 42(1):19–30.

Kiranyaz, S., Gabbouj, M., Pulkkinen, J., Ince, T., and Meissner, K. (2010a). Classification and retrieval on macroinvertebrate image databases using evolutionary RBF neural networks. Proceedings of the International Workshop on Advanced Image Technology (IWAIT).

Kiranyaz, S., Gabbouj, M., Pulkkinen, J., Ince, T., and Meissner, K. (2010b). Network of evolutionary binary classifiers for classification and retrieval in macroinvertebrate databases. *Proceeding of IEEE ICIP 2010*, pages 2257–2260.

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. *Computers in Biology and Medicine*, 41(7):463–472.

Kohavi, R. and Wolpert, D. H. (1996). Bias plus variance decomposition for zero-one loss functions. In Saitta, L., editor, *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 275–283. Morgan Kaufmann.

Kong, E. B. and Dietterich, T. G. (1995). Error-correcting output coding corrects bias and variance. In Prieditis, A. and Russell, S., editors, *Machine Learning: Proceedings of the Twelgth International Conference*, pages 313–321. Morgan Kaufmann.

Koskela, L., Sinha, B. K., and Nummi, T. (2007). Some aspects of the sampling distribution of the apportionment index and related inference. *Silva Fennica*, 41(4):699–715.

Krebs, C. J. (1999). *Ecological Methodology.* Benjamin/Cummings, 2nd edition.

Langley, P., Iba, W., and Thomas, K. (1992). An analysis of Bayesian classifiers. *Proceedings of the Tenth National Conference on Artificial Intelligence.* pp. 223-228. AAAI Press: Stanford.

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. *Journal of the North American Benthological Society*, 29(3):867–874.

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

Michie, D., Spiegelhalter, D., and Taylor, C. (1994). *Machine Learning, Neural and Statistical Classification.* Ellis Horwood Series in Artificial Intelligence. Ellis Horwood.

Nielsen, T. D. and Jensen, F. V. (2009). *Bayesian Networks and Decision Graphs*. Springer Science & Business Media.

Novak, M. A. and Bode, R. W. (1992). Percent model affinity: a new measure of macroinvertebrate community composition. *Journal of the North American Benthological Society*, 11(1):80–85.

Passy, S. I. and Bode, R. W. (2004). Diatom model affinity (dma), a new index for water quality assessment. *Hydrobiologia*, 524:241–251.

Patil, G. P. and Taillie, C. (1982). Diversity as a concept and its measurement. *Journal of the American Statistical Association*, 77(379):548–561.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1:81–106.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Ransom, M. R. (2000). Sampling distributions of segregation indexes. *Sociological Methods & Research*, 28(4):454–475.

Renkonen, O. (1938). Statisch-kologische Untersuchungen ber die terrestrische Kferwelt der finnischen Bruchmoore. *Ann. Zool. Soc. Bot. Fenn. Vanamo*, 6:1–231.

Ricklefs, R. E. and Lau, M. (1980). Bias and dispersion of overlap indices: results of some Monte Carlo simulations. *Ecology*, 61(5):1019–1024.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks*. Cambridge, Cambridge University Press.

Rosenberg, D. M. and Resh, V. H., editors (1993). *Freshwater Biomonitoring and Benthic Macroinvertebrates*. Chapman & Hall.

Seber, G. A. F. (1973). *The Estimation of Animal Abundance and Related Parameters*. C. Griffin, London, England.

Shannon, C. and Weaver, W. (1963). *The Mathematical Theory of Communication*. University Illinois Press, Urbana.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Shao, G., Liu, D., and Zhao, G. (2001). Relationships of image classification accuracy and variation of landscape statistics. *Canadian Journal of Remote Sensing*, 27(1):33–43.

Shao, G. and Wu, W. (2004). *Remote Sensing and GIS Accuracy Assessment*, chapter The effects of classification accuracy on landscape indices, pages 209–220. CRC Press.

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

Smith, E. P. (1982). Niche breadth, resource availability and inference. *Ecology*, 63(6):1675–1681.

Smith, E. P. and Zaret, T. M. (1982). Bias in estimating niche overlap. *Ecology*, 63(5):1248–1253.

Stoddard, J. L., Larsen, D. P., Hawkins, C. P., Johnson, R. K., and Norris, R. H. (2006). Setting expectations for the ecological condition of streams: the concept of reference condition. *Ecological Applications*, 16(4):1267–1276.

Stribling, J. B., Pavlik, K. L., Holdsworth, S. M., and Leppo, E. W. (2008). Data quality, performance, and uncertainty in taxonomic identification for biological assessments. *Journal of North American Benthological Society*, 27(4):906–919.

Strobl, C. and Zeileis, A. (2008). Danger: High power! – exploring the statistical properties of a test for random forest variable importance. *In Proceedings of the 18th International Conference on Computational Statistics*, Porto, Portugal.

Tibshirani, R. (1996). Bias, variance and prediction error for classification rules. Technical report, Department of Statistics, University of Toronto.

Tirronen, V., Caponio, A., Haanpää, T., and Meissner, K. (2009). Multiple order gradient feature for macroinvertebrate identification using support vector machines. *Lecture Notes in Computer Science*, pages 489–497.

Venrick, E. (1983). Percent similarity: the prediction of bias. *Fishery Bulletin*, 81(2):375–387.

WFD (2000). Directive 2000/60/EC of the European Parliament and the Council of 23, October 2000. A framework for community action in the field of water policy. *Off. J. Eur. Commun.*, L327:72.

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. *Photogrammetric Engineering & Remote Sensing*, 63(4):397–402.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1:80–83.

Wolda, H. (1981). Similarity indices, sample size and diversity. *Oecologia (Berl)*, 50:296–302.

I

Ärje, J., Kärkkäinen, S., Meissner, K. & Turpeinen, T. (2010) Statistical classification and proportion estimation – an application to a macroinvertebrate image database. *Proceedings of the 2010 IEEE Workshop on Machine Learning for Signal Processing (MLSP2010)*, Kittilä, Finland, page 373-378.

STATISTICAL CLASSIFICATION AND PROPORTION ESTIMATION – AN APPLICATION TO A MACROINVERTEBRATE IMAGE DATABASE

Johanna Ärje, Salme Kärkkäinen

Department of Mathematics and Statistics
University of Jyväskylä
johanna.arje@jyu.fi
salme.karkkainen@jyu.fi

Kristian Meissner, Tuomas Turpeinen***

*Finnish Environment Institute
Kristian.Meissner@ymparisto.fi
**Department of Physics
University of Jyväskylä
tuomas.turpeinen@phys.jyu.fi

ABSTRACT

We apply and compare a random Bayes forest classifier and three traditional classification methods to a dataset of complex benthic macroinvertebrate images of known taxonomical identity. Since in biomonitoring changes in benthic macroinvertebrate taxa proportions correspond to changes in water quality, their correct estimation is pivotal. As classification errors are passed on to the allocated proportions, we explore a correction method known as a confusion matrix correction. Classification methods were compared using the misclassification error and the χ^2 distance measures of the true proportions to the allocated and to the corrected proportions. Using low misclassification error and smallest χ^2 distance measures as performance criteria the classical Bayes classifier performed best followed closely by the random Bayes forest.

1. INTRODUCTION

Macroinvertebrate samples are commonly used in biomonitoring to study human induced changes on aquatic ecosystems (e.g. [1], [2]). Traditionally species in samples are classified by human experts. In this work, we compare the efficiency of the image-based classification methods and explore the effect of a confusion matrix correction to species proportions after initial classification (e.g. [3]).

Benthic macroinvertebrates living on the bottom of waterbodies are quick to react to changes in water quality. For biomonitoring, it is crucial to obtain highly accurate estimates of macroinvertebrate species abundances and proportions since observed changes in them are good indicators of ecosystem changes. The traditional human-made classification is, however, both time-consuming and expensive.

We would like to warmly thank Phil Culverhouse, Harri Högmander, Serkan Kiranyaz, Simon Oliver, Antti Penttinen, Mats Rudemo and Sara Taskinen for valuable suggestions and discussions. The financial support of STATCORE (a project of the University Alliance Finland) is appreciated.

In our work, classification, i.e. automated identification, is based on the features such as area extracted from greyscale images of individuals (Fig. 1). We present random Bayes forest classifier, which is the implementation of the random forest approach on the Bayes classifier. A similar type of approach, i.e. random naive Bayes, can be found in [4]. We briefly describe traditional Bayes, decision tree and random forest (e.g. [5]). Besides forming the basis of our novel classifier, classifiers of this type have been used for automated identification [6, 7]. Since class conditional classification errors cause bias to the allocated proportions (the proportions resulting from the automated classification), we further explore the correction of the allocated proportions using the inverse of the confusion matrix [3].

With a predefined macroinvertebrate image database, we can apply and compare the classification methods using the misclassification error, the χ^2 distance measures between true and allocated proportions and the distance measures between true and corrected proportions, respectively [8].

2. DATA

Due to the high amount of manual expert labor needed to identify true species identities, the number of species used here to get an initial estimate of the viability of automated recognition for benthic macroinvertebrate was deliberately limited. Although the taxa included into this digitized data set are representative of taxa typically found in rivers, they form a small subset of the 30-75 taxa typically encountered at individual river sites. We used eight common macroinvertebrate taxa (*Baetis rhodani, Diura nanseni, Heptagenia sulphurea, Hydropsyche pellucidulla, Hydropsyche siltalai, Isoperla sp., Rhyacophila nubila* and *Taeniopteryx nebulosa*) for automated identification, see examples in Fig. 1.

Specimens included were first keyed traditionally by taxonomic experts and then scanned onto a computer in single species batches. The segmentation for each batch-scan pic-

Fig. 1. Example of segmented macroinvertebrate taxa images used for feature extraction and classification. Taxa depicted from upper left to bottom right are: *Baetis rhodani, Diura nanseni, Heptagenia sulphurea, Hydropsyche pellucidulla, Hydropsyche siltalai, Isoperla sp., Rhyacophila nubila* and *Taeniopteryx nebulosa*.

ture file was performed in multiple steps. First an estimation of the image background was obtained by median filtering the image with kernel size larger than the radius of the largest object in the image. The background image was used to normalize the background value of each pixel in the original image to the average of the background level. A global threshold was used to generate a mask that separated the specimens from the background in the normalized image. This procedure enables the detection of individual specimens as connected (we used 8-connectivity) areas in the mask. We used an automatic technique for choosing the threshold value first developed by Ridler & Calvard [9]. In some cases, when the specimens were touching or overlapping, the segmentation was solved manually. We acknowledge that there are more automatic ways to handle touching or overlapping specimens. However, due to the low number of overlapping cases requiring our attention, we did not make use of e.g. sophisticated blob detectors (e.g., LoG, DoG) to detect the amount of specimens in the image and then finding the minimum cost path (e.g, seam carving) between the objects for this dataset.

After segmentating all the greyscale batch-scan picture files for the eight species, 1350 images of indivuals were subsequently obtained. From those images, used features were extracted as follows: First, using ImageJ the indivual was automatically separated from the background of the image and subsequently the features were extracted. Second, we reduced the initial number of features through the use of canonical discriminant analysis on the data [5]. Finally, the following set of simple a) geometry and b) intensity -based features were obtained for classification:

a) area, perimeter, height and width of smallest rectangle, major and minor axes of smallest ellipse, circularity, Feret's diameter

b) mean, median, mode, standard deviation, kurtosis, skewness and sum of grey values,

see ImageJ [10] for more details.

3. CLASSIFIERS

3.1. Bayes classifier

Let $\{\omega_1, \ldots, \omega_k\}$ be the finite set of k classes and let $P(\omega_i)$ be the prior probability of class ω_i such that $\sum_{i=1}^{k} P(\omega_i) = 1$. Let $\mathbf{x}$ be a p-dimensional feature vector in the Euclidian space $\mathbb{R}^p$ with the class-conditional probability density function $p(\mathbf{x}|\omega_i)$ given the true class ω_i. Now the posterior probability is

$$P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{p(\mathbf{x})},$$

where $p(\mathbf{x}) = \sum_{i=1}^{k} p(\mathbf{x}|\omega_i)P(\omega_i)$. The Bayes decision rule classifies each individual to the class corresponding to the largest posterior probability, thus minimizing the probability of classification error [11].

The Bayes decision rule can also be represented in the form of a discriminant function $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$, instead of the posterior probability. If the features are distributed normally, $\mathbf{x}|\omega_i \sim N_p(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$,

$$\begin{aligned} g_i(\mathbf{x}) &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{p}{2}\ln 2\pi \\ &\quad -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \end{aligned} \tag{1}$$

[5]. The Bayes classifier is known as quadratic discriminant analysis (QDA), if the covariances for each class are arbitrary. If all classes have a common covariance matrix $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ for all i, the quadratic terms disappear, and we get linear discriminant functions (LDA). In this work, we use QDA and equal priors.

3.2. Decision tree

A decision tree consists of nodes, branches and leaves. First the feature that best splits the data into the classes is chosen as a root node. Observations are then split according to the values of the root node into subsets, which form the first branches of the decision tree. The procedure is repeated for all branches until the resulting subsets are pure - that is, all the observations in the subset belong to the same class. At the end of each pure branch a leaf is grown and labeled according to the class of the individuals on that branch.

Let $\{\omega_1, \ldots, \omega_k\}$ be values of a discrete random variable Y representing the species class, and let X be a discrete random feature with s classes. The choice of a feature

for a node of the decision tree is made based on entropy. The entropy of Y is

$$H(Y) = -\sum_{i=1}^{k} p_i \log p_i, \tag{2}$$

where $p_i = \mathbb{P}(Y = \omega_i) = P(\omega_i) \geq 0$ for all $i = 1, \ldots, k$. The conditional entropy of Y conditioned on feature X is

$$H(Y|X) = \sum_{j=1}^{s} p_{\cdot j} H(Y|X = x_j), \tag{3}$$

where $p_{\cdot j} = \sum_{i=1}^{k} p_{ij}$ and within the class x_j

$$\begin{aligned} H(Y|X = x_j) \quad = \quad & -\sum_{i=1}^{k} P(Y = \omega_i|X = x_j) \\ & \times \log(P(Y = \omega_i|X = x_j)) \end{aligned}$$

equals the entropy of the distribution of taxa into the classes. Low entropy values indicate that the feature X discriminates the classes well, consequently that information can be used in selecting the feature for a node. The Id3 algorithm uses Information gain, $G(Y, X) = H(Y) - H(Y|X)$, and chooses the feature with the largest $G(Y, X)$ (smallest $H(Y|X)$) for the node variable [12]. Although decision trees were originally developed for discrete features, some of them can also deal with continous ones, an example is the C4.5 decision tree, which uses Gain ratio:

$$\text{Gain ratio}(X) = \frac{G(Y, X)}{H(X)}$$

to determine the node features [13]. In this work, the C4.5 decision tree is used.

3.3. Random Forests

Random forest (RF) uses bagging (random inputs) and random feature selection on decision trees resulting in a forest of random decision trees [14, 5]. Each tree is grown on a different bootstrap sample drawn randomly with replacement from the training data. At each node m features are randomly drawn from the original M features and from these the best split is chosen. For the classification of a new observation each tree casts a vote and the class for each observation is assigned based on the majority vote.

A typical estimate for the misclassification error is the test error, the ratio of misclassifications of all observations in the test data [5]. Since random forest uses bagging there is no need for a separate test set. When a tree is grown on a bootstrap sample, about one-third of the observations are left out. These cases form the out-of-bag (oob) data that can subsequently be used in estimating the test error. Each observation is classified using the majority vote from all the trees in which it was "out-of-bag". The oob error e_{oob} gives a running unbiased estimate for the test error when the trees are added.

3.4. Random Bayes forest

Random Bayes forest (RBF) is the application of the random forests approach on a Bayes classifier. It produces a collection of random Bayes classifiers, each built on a different bootstrap sample of the training data and using m randomly drawn features of the original M features. Bagging enhances the robustness of the base classifier whereas the random feature selection improves its accuracy [4].

Each random Bayes classifier computes posterior probabilities or discriminant functions for all classes and classifies each observation to the class corresponding to the largest quadratic discriminant function. The classifier then votes for this class. Similarly to random forests, the final classification is made based on the majority vote of all individual random Bayes classifiers. As with random forests in random Bayes forests there is no need for a separate test set because the oob error (e_{oob}) provides an estimate for the test error.

4. CORRECTION OF ALLOCATED PROPORTIONS

Let us consider objects from k different classes with their true proportions $\mathbf{p} = (p_1, \ldots, p_k)$ such that $\sum p_i = 1$. The classification of objects is based on the observed features $\mathbf{x} = (x_1, \ldots, x_p)$ of each object. Using only the features of objects, we are not able to discriminate the classes perfectly. Misclassification errors are then included in the allocated proportions $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_k)$, referred to in the literature also as raw estimates [3]. Our focus is on the estimation of true proportions $\mathbf{p}$ using allocated proportions $\hat{\mathbf{p}}$ and the confusion matrix (misclassification error matrix).

Assuming that the confusion matrix $\mathbf{A}$ of a specified classification procedure is known, as proposed by Fortier [3], the element a_{ij} of $\mathbf{A}$ is the probability of classifying an object in the class i when originating from the class j. Further, $\sum_i a_{ij} = 1$ and $a_{ij} \geq 0, i, j = 1, \ldots, k$. Following Fortier's model [3], let us consider the classification result in matrix form where the element n_{ij} is the number of objects classified in the class i when originating from the class j. Allocated proportions are obtained as follows:

$$\hat{p}_i = \frac{N_i}{N},$$

where the sample size is denoted by N and the number of objects classified in the class i,

$$N_i = \sum_j n_{ij}, \tag{4}$$

is the sum of the elements of the ith row.

When considering the real sample, only N_i's can be observed but not the elements of the matrix, $\{n_{ij}\}$. The elements of the jth column, however, follow a multinomial distribution

$$(n_{1j}, \ldots, n_{kj}) \sim multin(p_j N, a_{1j}, \ldots, a_{kj}), \quad (5)$$

where $p_j N = \sum_i n_{ij}$ is the sum of the elements of the jth column. Since the classification in the given class is independent from the classification of other classes, the vector $(N_1, \ldots, N_k)$ is the sum of the independent random vectors following the multinomial distribution. Consequently, the expectations for $i = 1, \ldots, k$ are

$$E(N_i) = \sum_j E(n_{ij}) = \sum_j a_{ij} p_j N,$$

see formulas (4) and (5). Then $E(\hat{p}_i) = \sum_j a_{ij} p_j$. In the matrix form, the latter equals

$$E(\hat{\mathbf{p}}) = \mathbf{A}\mathbf{p}. \quad (6)$$

The elimination of the systematic bias in formula (6) is achieved by using the estimator $\mathbf{A}^{-1}\hat{\mathbf{p}}$, when $\mathbf{A}$ is nonsingular.

5. COMPARISON OF CLASSIFIERS

The four classifiers were compared using the misclassification error. In the case of Bayes classifier and decision tree the data was randomly split into the training and test subsets. Given the training subset τ we estimated the test error Err_τ from the test data [5, p. 220]. Further, we calculated the mean test error $\overline{\mathrm{Err}}_\tau$ over the 100 training and test set splits. For both the RF and RBF classifier, the estimate for the misclassification error was obtained using the out-of-bag data as the test data and calculating the mean oob error $\overline{e_{oob}}$ over the 100 forests.

In addition, we compared the results of four classifiers using the allocated proportions $\hat{\mathbf{p}}$ and the corrected proportions $\mathbf{A}^{-1}\hat{\mathbf{p}}$ after the confusion matrix correction. Since the confusion matrix $\mathbf{A}$ is unknown, the data was divided into three parts (training data, validation data and test data) to be able to utilize the confusion matrix correction. The classification rule was obtained from the training data. Using this classification rule, the confusion matrix was estimated from the validation data. From the test data we calculated the true $\mathbf{p}$, and both the raw $\hat{\mathbf{p}}$, and the corrected estimates $\mathbf{A}^{-1}\hat{\mathbf{p}}$ of the true proportions. We estimated the difference between the true and the estimated proportions using the χ^2 distance measure [8]

$$\chi^2 = \sum_{i=1}^{k} \frac{(h_i - p_i)^2}{p_i}, \quad (7)$$

where p_i is the true proportion of the class i and h_i is either the raw estimate or the corrected estimate of p_i. Note that the χ^2 distance measures of (7) for both the raw and corrected proportion estimates are affected by chance variation. In addition, corrected proportion estimates are also affected by the estimation of the confusion matrix itself.

Using a bootstrap method frequently used in the estimation of variance (e.g. [5]), we assessed the variation and distribution of the distance measure for each classifier. Given the classification rule, the bootstrap was used for both the validation and test data as follows. We randomized B bootstrap samples from the validation and test data. From B pairs of bootstrap samples, we calculated the raw $\hat{\mathbf{p}}_\mathbf{1}, \ldots, \hat{\mathbf{p}}_\mathbf{B}$, and the corrected proportion estimates $\mathbf{A}^{-1}\hat{\mathbf{p}}_\mathbf{1}, \ldots, \mathbf{A}^{-1}\hat{\mathbf{p}}_\mathbf{B}$. Since the true proportions $\mathbf{p_l}$ of lth bootstrap sample of the test data were known, we were able to calculate the lth distance measure

$$\chi^2_l = \sum_{i=1}^{k} \frac{(h_{li} - p_{li})^2}{p_{li}}, \quad (8)$$

where p_{li} is the true proportion of the class i and h_{li} is either the raw or the corrected estimate of p_{li}. Using (8), we obtained the distribution of the distance measure for both raw and corrected proportion estimates for each classifier.

With the forest classifiers (RF and RBF), we could estimate the confusion matrix from the oob data. Consequently, we picked one third of our data as the test data and the rest of the data was used both for training and validation.

6. RESULTS OF COMPARISON

6.1. Bayes classifier

For the comparison of the misclassification errors, we drew 100 samples which each used 650 observations as training data and the remaining 700 as test data for the Bayes classifier. For every sample we calculated the test error and averaged it over all the samples, $\overline{\mathrm{Err}}_\tau = 0.0736$.

For the comparison of the proportion estimates we split the data into three subsets. We built the Bayes classifier with the first subset (i.e. the training data) and randomized 100 bootstrap samples both from the second and third subset; i.e. the validation data and test data, respectively. Raw proportion estimates $\hat{\mathbf{p}}_\mathbf{1}, \ldots, \hat{\mathbf{p}}_\mathbf{B}$, the corrected proportion estimates $\mathbf{A}^{-1}\hat{\mathbf{p}}_\mathbf{1}, \ldots, \mathbf{A}^{-1}\hat{\mathbf{p}}_\mathbf{B}$ and their distance measures χ^2 were calculated using the aforementioned subsets. Thus, we were able to obtain the empiric bootstrap distibution of the distance measure for both estimates. For the Bayes classifier, raw estimates of proportions clearly outperformed the corrected ones (Fig. 2, Tables 1-2).

Table 1. Summary statistics of the empiric distribution of χ^2 distance measure between true and allocated proportions for each classifier. See text for abbreviations.

	Mean	Median	Range	Std.
Bayes	0.0005	0.0003	0.0000 – 0.0049	0.0007
C4.5	0.0084	0.0082	0.0007 – 0.0238	0.0045
RF	0.0035	0.0027	0.0000 – 0.0141	0.0030
RBF	0.0015	0.0009	0.0000 – 0.0074	0.0017

Table 2. Summary statistics of the empiric distribution χ^2 distance measure between true and corrected proportions for each classifier. See text for abbreviations.

	Mean	Median	Range	Std.
Bayes	0.0025	0.0015	0.0000 – 0.0153	0.0030
C4.5	0.0040	0.0022	0.0000 – 0.0225	0.0046
RF	0.0043	0.0025	0.0000 – 0.0228	0.0048
RBF	0.0009	0.0004	0.0000 – 0.0068	0.0013

6.2. Decision tree

For the C4.5 decision tree we drew 100 samples which each used 650 observations as training data and the remaining 700 observations as test data. Averaging over all the samples, we obtained $\overline{\text{Err}}_\tau = 0.1731$.

The distributions of the distance measures for the raw estimates and the corrected ones were calculated using analogous procedures as with the Bayes classifier. In the case of the C4.5 decision tree we observed more small distances for the corrected estimates than for the raw estimates. The post-classification correction improved the accuracy of proportion estimates for the C4.5 decision tree (Fig. 2, Tables 1-2).

6.3. Random forest

Since random forest does not require a separate test data, we were able to use all observations to build the classifier. We simulated 100 random forests with 70 trees each. Individual trees were built using five randomly selected features at each node. Averaging the out-of-bag error over all simulations we obtained $\overline{e_{oob}} = 0.0762$.

We randomly picked one third of the observations to be our test data for the comparison of the raw and corrected estimates of proportions. From the remaining data we simulated 100 random forests and calculated the confusion matrices from the out-of-bag observations for each forest. Thus, with each forest this data were used both for the training and validation data. We randomized 100 bootstrap samples from the test data to compute the estimates for raw and corrected proportions and their distance measures. Allocated and corrected proportion estimates produced quite similar distributions

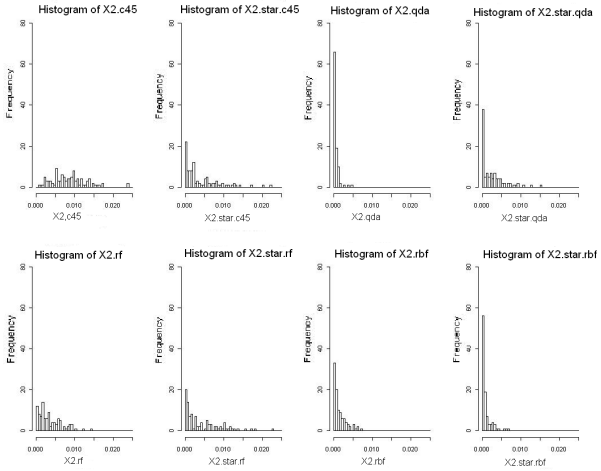

Fig. 2. The empiric bootstrap distributions of the χ^2 distance measure between true and allocated proportions (upper row) and between true and corrected proportions (bottom row) for each classifier. Depicted, from left to right: the Bayes classifier, the C4.5 decision tree, the random forest (RF) and the random Bayes forest (RBF).

of distance measures in the case of random forests (Fig. 2, Tables 1-2).

6.4. Random Bayes forest

As random Bayes forest operates similarly to random forest, it does not require a separate test data either. We simulated 100 random Bayes forests with 90 Bayes classifiers built from 8 randomly drawn features. We averaged the out-of-bag error over all simulations and obtained $\overline{e_{oob}} = 0.0769$.

For comparisons between raw and corrected proportion estimates we randomly picked one third of our observations to be the test data and simulated 100 random Bayes forests with the remaining data. Confusion matrices were computed from the out-of-bag-observations of each forest. We then randomized 100 bootstrap samples from the test data and calculated the raw and the corrected proportion estimates and their empiric distance measure distributions. As for the C4.5 decision trees, the corrected proportion estimates outperformed the raw ones for random Bayes forests (Fig. 2, Tables 1-2).

Of the classifiers examined in this work the decision tree performed poorest, whereas the Bayes classifier produced the lowest misclassification error with RF and RBF achieving almost similar results. Obtaining such a good result for the Bayes classifier is no surprise, for detailed examples and reasons see [5, p. 111]. Further, Hastie *et al.* point out that

as a highly nonlinear classification method the decision tree stands to benefit the most from random inputs and random feature selection [5, p. 589]. The macroinvertebrate image feature data used here mirror these predictions (Table 1, Fig. 2, upper row).

When the correction for allocated proportions is used, the result of the decision tree improves the most, followed closely by random Bayes forest. However, the Bayes classifier performs worse after correction (Tables 1-2, Fig. 2). According to Fortier [3], allocated proportions can beat corrected ones in cases where the data for classification is small. Note also that the confusion matrix of the Bayes classifier was estimated from the data set of same size. In both previous cases, the data was quite small with respect to the number of species.

7. DISCUSSION

In this work, we applied and compared three classical classification methods to a random Bayes forest. We further explored the effect of a confusion matrix correction to allocated proportions. Our data were features extracted from complex benthic macroinvertebrate images.

The Bayes classifier yielded the best results. With it we obtained the smallest misclassification error and the best estimates for the proportions of species when compared using the χ^2 distance measures. Random forest and random Bayes forest achieved almost as good results as the Bayes classifier. Especially when examining corrected proportion estimates, the random Bayes forest produced results similar to the raw proportion estimates of the Bayes classifier. We believe that through fine tuning of the parameters and with larger simulations random Bayes forest may outperform the Bayes classifier.

We feel that it may be worthwhile exploring the use of posterior probabilities in the voting process more profoundly, rather than just determining the most likely class for each random Bayes classifier in the forest. Posterior probalities could also be used in identifying outliers, e.g. observations that do not actually contain a macroinvertebrate. In this test study on a limited selection of macroinvertebrate taxa we used equal prior probabilities for the Bayes classifier and the random Bayes forest. However, in biomonitoring of complex benthic river communities experts can make very accurate predictions on probabilities of species occurences based on the geological and physico-chemical characteristics of the target river. Exploring the use and validating these expert derived priors in future studies will prove a very interesting challenge.

8. REFERENCES

[1] J. F. Wright, D. Moss, P. D. Armitage, and M. T. Furse, "A preliminary classification of running-water sites in great britain based on macro-invertebrate species and the prediction of community type using environmental data," *Freshwater Biology*, vol. 14, pp. 221–256, 1984.

[2] J. R. Karr and E. W. Chu, "Sustaining living rivers," *Hydrobiologia*, vol. 422/423, pp. 1–14, 2000.

[3] J.-J. Fortier, "Best linear corrector of classification estimates of proportions of objects in several unknown classes," *The Canadian Journal of Statistics*, vol. 20, pp. 23–33, 1992.

[4] A. Prinzie and D. Van den Poel, "Random multiclass classification: Generalizing random forests to random MNL and random NB," 2007, LNCS 4653, Springer-Verlag.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2nd edition, 2009.

[6] M. Blaschko et al., "Automatic in situ identification of plankton," 2005, Proceedings of the Seventh IEEE Workshops on Application of Computer Vision (WACV/MOTION'05), vol. 1, pp. 79–86, IEEE Computer Society.

[7] P. Culverhouse et al., "Automatic classification of field-collected dinoflagellated by articial neural network," *Marine Ecology Progress Series*, vol. 139, pp. 281–287, 1996.

[8] W. J. Krzanowski and F. H. C. Marriot, *Multivariate Analysis, Part 1, Distributions, Ordination and Inference*, Edward Arnold, 1994.

[9] T. W. Ridler and S. Calvard, "Picture thresholding using an iterative selection method," 1978, IEEE Trans. System, Man and Cybernetics, SMC-8.

[10] S. W. Rasband, *ImageJ*, U. S. National Insitutes of Health, Bethesda, Maryland, USA, 1997-2010, http://rsb.info.nih.gov/ij/.

[11] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, Wiley, 2nd edition, 2001.

[12] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 1986.

[13] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

[14] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

II

Ärje, J., Kärkkäinen, S., Turpeinen, T. & 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(4), 248–259.

Research Article Environmetrics

Received: 1 June 2012, Revised: 27 February 2013, Accepted: 27 February 2013, Published online in Wiley Online Library: 30 April 2013

(wileyonlinelibrary.com) DOI: 10.1002/env.2208

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

J. Ärje[a]*, S. Kärkkäinen[a], T. Turpeinen[b] and K. Meissner[c]

Macroinvertebrate samples are commonly used in biomonitoring to study changes on aquatic ecosystems. Traditionally, specimens are identified manually to taxa by human experts being time-consuming and cost intensive. Using the image data of 35 taxa and 64 features, we propose a novel variant of the quadratic discriminant analysis for breaking the curse of dimensionality in quadratic discriminant analysis models. Our variant, called a random Bayes array (RBA), uses bagging and random feature selection similar to random forest. We explore several variations of RBA. We consider three classification (i.e taxa identification) decisions: majority vote, averaged posterior probabilities, and a novel approach; a score of weighted votes. Besides modifying the voting, we propose to weight features according to their importance instead of eliminating the least important features. We compared the performance of RBA with traditional Bayesian and several other popular classification methods and assessed how the methods behave in relation to each other and the different macroinvertebrate species. Further, we investigate how severely misclassifications affect the performance of different methods when set into a biomonitoring context. We found that the lowest and least severe classification error (i.e. most accurate taxa identification) was achieved with RBA by using averaged posterior probabilities and weighted features.

Keywords: biomonitoring; classification; correspondence analysis; random Bayes array

1. INTRODUCTION

Aquatic ecosystems worldwide face a large variety of anthropogenic pressures. Among these, climate change, pollution, and land use are causing biodiversity loss and deterioration in water quality (Millennium Ecosystem Assessment, 2005). In biomonitoring, benthic macroinvertebrates are often used to detect these human-induced changes in aquatic ecosystems. Macroinvertebrate taxa have varying sensitivities to certain pressures, for example, pollution, and are not only quick to react to changes in water quality but also reflect subtle long-term water quality changes in their community composition.

Despite the obvious advantages of using macroinvertebrate biomonitoring as an indicator of water quality change, currently the manual identification of macroinvertebrates is laborious. The traditional human-made classification is time-consuming, expensive, and has recently been found to be unexpectedly error prone (Haase *et al.*, 2010). Recent advances have demonstrated that both in terms of speed and accuracy, automated classification matches and even outperforms manual taxa identification (Ärje *et al.*, 2010; Kiranyaz *et al.*, 2011; Lytle *et al.*, 2010; Joutsijoki and Juhola, 2012). Compared with previous works, this data is more extensive, comprising 6814 individuals from 35 taxa with 64 features extracted from an image of each individual. The data includes classes with few observations compared with the number of features (Table A1, second left column) and images with overlapping and damaged individuals (Figure 2).

Hastie *et al.* (2009) recommend to have linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), because these provide good classification records (Michie *et al.*, 1994). The reason for the good records is likely to be because the data often supports the linear or quadratic boundaries with stable, Gaussian model-based estimates. When the class-specific feature space is high dimensional

* *Correspondence to: J. Ärje, Department of Mathematics and Statistics, P.O.Box 35 (MaD), FI-40014, University of Jyväskylä, Finland. E-mail: johanna.arje@jyu.fi*

a *Department of Mathematics and Statistics, University of Jyväskylä, P.O. Box 35 (MaD), FI-40014, Finland*

b *Department of Mathematical Information Technology, University of Jyväskylä, P.O. Box 35 (Agora), FI-40014, Finland*

c *Finnish Environment Institute SYKE, Jyväskylä Office, Survontie 9, 40500, Jyväskylä, Finland*

compared with the number of observations, the use of QDA is not supported, but linear methods LDA and even naïve Bayes (NB) with independent features may be appropriate to efficiently estimate the high-dimensional covariance matrices and their inverses.

Because the conditional independence assumption of NB or the common covariance matrix assumption of LDA rarely hold with real data, we propose a novel modification of the QDA model designed for small sample sizes and a large number of features, to reduce the dimension of the feature space. Prinzie and van den Poel (2007) proposed a random NB (RNB) classifier for this purpose. The RNB is based on the random forest (RF) of Breiman (2001) and requires the selection of feature subsets to better satisfy the conditional independence assumption (Langley and Sage, 1994). According to Hastie *et al.* (2009), nonlinear classification methods (such as QDA, not NB and LDA) can be improved by introducing randomness to the model. For example, the use of RNB in Prinzie and van den Poel (2007) only slightly enhances the prediction accuracy of NB. To build our novel modification, we extend the RNB classifier to dependent Gaussian features, that is, to QDA, which is a nonlinear classifier and can benefit from random shaking. Our classification method, random Bayes array (RBA), is a collection of random QDA classifiers trained with bootstrap samples of a training set and random feature selection of the original features. The performance of this novel RBA approach (Ärje *et al.*, 2010), based on dependent Gaussian features has, to our knowledge, not been investigated earlier. Using an ensemble approach, we reduce the high variance of the QDA model in cases where a large number of parameters needs to be estimated from small samples – as for our data set. Random feature selection may reduce the chance of obtaining ill-conditioned covariance matrices when some of the features are highly correlated, as in our data. Finally, utilizing random subset selection of the features enables the use of QDA with smaller sample sizes.

In this article, we explore several variations of RBA. We first consider three alternative ways of making the final classification decision: majority vote, averaged posterior probabilities (Breiman, 1996), and a method of our own where we use the score of weighted votes. Further, we propose a method to improve the random selection of the features by using variable importance (Breiman, 2001) as weights.

By using the macroinvertebrate image database, we compare RBA with QDA and other Bayesian classifiers with real and simulated data to show that bagging and random feature selection can improve the performance of the QDA classifier on complex data. We compare RBA to several popular methods for automated classification: multilayer perceptron (MLP; Ripley, 1996), k-nearest neighbors (KNN; Hastie *et al.*, 2009), radial basis function (RBF; Buhmann, 2003), support vector machine (SVN; Cortes and Vapnik, 1995), and RF (Breiman, 2001). Comparisons of the different methods are based on their error rates and confusion matrices. Further, we assess what type of images typically lead to misclassifications. By using correspondence analysis (e.g. Greenacre, 2007), we investigate how the classifiers behave in relation to each other and to the different macroinvertebrate species.

2. DATA

Macroinvertebrate specimens from 33 lotic taxa (Table A1) were obtained through research projects of the Finnish Environment Institute and through the national freshwater biomonitoring program. Three taxonomic experts ensured the identity of the taxonomic specimens. Further, two lentic gastropod taxa (*Bithynia tentaculata* and *Myxas glutinosa*) were collected through a project of the department of Biological and Environmental Sciences of the University of Jyväskylä. Specimens were batch imaged by taxa by 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 by using a calibration target. Individual specimens were segmented, and each specimen was saved as a single posture image (for examples, see Figure 2).

In total, we obtained 6814 images from freshwater macroinvertebrates belonging to 35 different taxa. From each image, we extracted a total of 64 geometrical and statistical features. The features are a basic set of features provided by ImageJ (Rasband, 1997–2010). The following basic features were selected by using ImageJ's built in measurement and analysis functions: pixel value (grayscale and RGB) statistics (mean, SD, mode, minimum, maximum, center of mass coordinates, integrated density, median, kurtosis, and skewness) and geometric features (area, centroid, perimeter, bounding rectangle's coordinates, width and height, fitted ellipse's angle, major and minor axis, Feret's diameter, coordinates, angle and minimum caliber diameter, roundness, circularity, and solidity). A detailed description of these features can be found in the ImageJ manual (Rasband, 1997).

This data exhibits several special characteristics. First, imaged specimens are semirigid and fragile, which creates a large amount of variation in the actual shape between specimens belonging to the same taxa. Specimens may also overlap during the digitizing, exhibit rotations and vary in intensity levels (Figure 2). In addition, part of the taxa groups are small compared with the number of features extracted (i.e. 64). The number of individuals in a taxa class varies from 63 individuals to 633 individuals, which given a 50/50 division of the image data into test and training sets leads to smaller class sizes than feature numbers for many taxa (Table A1).

3. RANDOM BAYES ARRAYS

Our data is characterized by a large feature number to class number ratio. Traditionally, QDA models are considered too complex to fit to small data sets because of their large number of parameters (Friedman, 1989), which often leads to high variance (i.e. "overfitting"). Therefore, simpler models such as NB or LDA have been employed. However, those simpler models impose very strong assumptions that are unlikely to be true. To alleviate the problems associated with QDA and LDA, Friedman (1989) proposed regularized discriminant analysis (RDA). With our complex data, RDA suffers from computational accuracy problems when inversing the covariance matrices, similarly to QDA. Instead of regularization, our novel approach applies an ensemble approach on the basis of RF (Breiman, 2001) to QDA models. Further, we compare different methods for making the final classification decision and introduce a method for weighting the features with their variable importance instead of dropping out the least important features.

3.1. A single Bayes classifier

More formally, let Ω be a finite set of classes $c = 1, \ldots, C$, and let $p(c)$ be the prior probability of class c such that $\sum_{c=1}^{C} p(c) = 1$. Let $\mathbf{x} = (x_1, \ldots, x_p)$ be a p-dimensional feature vector in the Euclidean space $\mathbb{R}^p$ with the class-conditional probability density function $f(\mathbf{x}|c)$ given the true class c. Then the posterior probability is given by

$$p(c|\mathbf{x}) = \frac{f(\mathbf{x}|c)\,p(c)}{f(\mathbf{x})},$$

where $f(\mathbf{x}) = \sum_{c=1}^{C} f(\mathbf{x}|c)\,p(c)$. The Bayes decision rule classifies each individual to the class corresponding to the largest posterior probability, thus minimizing the probability of classification error; see for example, Duda *et al.* (2001).

If the features given the class c are distributed normally, $\mathbf{x}|c \sim N_p(\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)$, the Bayes decision rule can also be represented in the form of a discriminant function,

$$\begin{aligned} g_c(\mathbf{x}) &= \ln f(\mathbf{x}|c) + \ln p(c) \\ &= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)'\boldsymbol{\Sigma}_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c) - \frac{p}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_c| + \ln p(c) \end{aligned} \tag{1}$$

where $\boldsymbol{\mu}_c$ is the mean vector, and $\boldsymbol{\Sigma}_c$ is the covariance matrix; they are estimated from the training data; see for example, Hastie *et al.* (2009). The Bayes classifier is known as QDA if the covariance matrix of each class is arbitrary. However, if all classes have a common covariance matrix $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$, the quadratic terms in equation 1 disappear, and we get LDA. The model where the features are further conditionally independent in each class, that is, $\boldsymbol{\Sigma}_c = \mathrm{diag}(\sigma_{c1}, \ldots, \sigma_{cp})$, is an NB classifier (John and Langley, 1995). Friedman (1989) proposed a compromise between LDA and QDA: RDA with two regularization parameters (λ, γ). With special choices of parameter values, we obtain LDA $(\lambda = 0, \gamma = 0)$, QDA $(\lambda = 1, \gamma = 0)$, and NB with equal variances within group $(\lambda = 0, \gamma = 1)$.

3.2. A basic random Bayes array

As a possible solution to overfitting experienced with basic QDA on data sets like ours, we extend the RNB classifier (Prinzie and van den Poel, 2007) to dependent Gaussian features. We propose a collection of B random Bayes classifiers (QDAs), each built on a different bootstrap sample that we subsequently refer to as a RBA. For a single random Bayes classifier, N observations are randomly sampled with replacement from the original training data of N observations. For each bootstrap sample, we further sample randomly m features from the original p ones to be used in building the classification model.

In RBA, each random Bayes classifier b, $b = 1, \ldots, B$, produces the posterior probability $\hat{p}_b(c|\mathbf{x})$ for each class c. The predicted class $\hat{c}(\mathbf{x})$ of a new individual $\mathbf{x}$ can be obtained from the posterior probabilities. First, each random Bayes classifier votes for the class with the highest posterior probability, that is, a classifier b votes the class $\hat{c}_b(\mathbf{x}) = c$, if $\hat{p}_b(c|\mathbf{x}) = \max_{c'} \hat{p}_b(c'|\mathbf{x})$. Then, the final class $\hat{c}(\mathbf{x})$ is determined by majority vote $\{\hat{c}_b(\mathbf{x})\}$. We have applied this voting rule earlier in Ärje *et al.* (2010). Here it is denoted by $\mathrm{RBA}_{\mathrm{basic}}$.

3.2.1. *Modifications of voting*

For cases where the classification method provides a class probability estimate, Breiman (1996) suggested an alternative strategy to voting, namely the averaging of probabilities. We use this approach for the RBA and average the class posterior probabilities over its B classifiers as follows:

$$\hat{p}_{\mathrm{ave}}(c|\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} \hat{p}_b(c|\mathbf{x}) \tag{2}$$

The final class is the class with the highest averaged posterior; $\hat{c}(\mathbf{x}) = c$, if $\hat{p}_{\mathrm{ave}}(c|\mathbf{x}) = \max_{c'} \hat{p}_{\mathrm{ave}}(c'|\mathbf{x})$. In the later text, this voting method is denoted by $\mathrm{RBA}_{\mathrm{ave}}$.

In addition, we explore another method for the classification decision, which we call $\mathrm{RBA}_{\mathrm{weight}}$. Instead of the number of votes, each class c, $c = 1, \ldots, C$, obtains a score of votes weighted by the posterior probabilities

$$\mathrm{score}(c|x) = \sum_{b=1}^{B} 1_{\{\hat{p}_b(c|\mathbf{x}) = \max_{c'} \hat{p}_b(c'|\mathbf{x})\}} \hat{p}_b(c|\mathbf{x}),$$

where the indicator has value 1 if the class c was voted by a single classifier b and the vote is weighted by its corresponding posterior probability. The decision on the final class $\hat{c}(\mathbf{x}) = c$, if $\mathrm{score}(c|\mathbf{x}) = \max_{c'} \mathrm{score}(c'|\mathbf{x})$, is based on the class with the highest score.

3.2.2. *Using the importance of features as weights in sampling*

As with random forests (Breiman, 2001), we can explore the importance of a feature x_j, $j = 1, \ldots, p$, by permuting the values of the feature x_j. If a feature x_j is important for successful classification, its permutation in a test set should decrease the number of correctly classified objects. Because approximately one third of the original training observations (i.e. out-of-bag (oob) cases) are left out of each bootstrap

sample, they can be used as a test set in the calculation of feature importance. As an importance measure for the feature x_j, $j = 1, \ldots, p$, Breiman (2001) used the *variable importance*:

$$\text{imp}(x_j) = \frac{1}{B}\sum_{b=1}^{B}\left[\frac{\#\{\text{correct classifications of } b'\text{s oob cases}\}}{\#\{b'\text{s oob cases}\}} - \frac{\#\{\text{correct classifications when } x_j \text{ values of } b'\text{s oob cases permuted}\}}{\#\{b'\text{s oob cases}\}}\right] \tag{3}$$

Breiman (2001) stated that the variable importance of a permuted feature is decreased in the presence of highly correlated features. However, in simulation studies, Archer and Kimes (2008) discovered that the true predictors of highly correlated features were ranked among those features with the highest variable importance values.

Variable importance can be utilized in the sampling of m features for each classifier in the ensemble. Using the *weights* for the features x_j, $j = 1, \ldots, p$,

$$w(x_j) = \frac{\text{imp}(x_j)}{\sum_{j'=1}^{p}\text{imp}(x_{j'})},$$

the most important features are sampled more often than others. These weights can be used with the different classification decisions denoted by $\text{RBA}^*_{\text{basic}}$, $\text{RBA}^*_{\text{ave}}$, and $\text{RBA}^*_{\text{weight}}$.

In addition to the weights, we also considered the use of z-score (Breiman and Cutler, 2008), which is the standardized value of the variable importance

$$\widetilde{\text{imp}}(x_j) = \frac{\text{imp}(x_j)}{\hat{\sigma}/\sqrt{B}},$$

where $\hat{\sigma}$ is the estimated standard deviation of the difference term in the definition (3). On the basis of the central limit theorem, the mean $\widetilde{\text{imp}}(x_j)$ is assumed to be asymptotically standard normal under the null hypothesis of zero variable importance and therefore can be used for feature selection. However, using simulation studies, Strobl and Zeileis (2008) showed that the power of the test-based z-score does not increase with sample size but with the number of trees chosen by the researcher. Thus, rather than basing variable selection on z-scores, we propose the use of weights, as this includes all features and thus prevents information loss.

3.3. Out-of-bag error

The RBA produces an internal estimate for the classification error, mimicking cross validation. Let $L(c, \hat{c}(\mathbf{x}))$ be the 0-1 loss function indicating if the class $\hat{c}(\mathbf{x})$ predicted by the RBA matches the correct class c. As with RF, any versions of RBA produce an estimate for the prediction error by using the oob observations

$$\widehat{\text{Err}}_{\text{oob}} = \frac{1}{N}\sum_{i=1}^{N} L(c(i), \hat{c}'(\mathbf{x}_i)),$$

where $c(i)$ is the correct class of the observation $\mathbf{x}_i$, and $\hat{c}'(\mathbf{x}_i)$ is the RBA predictor constructed using the bootstrap samples in which $\mathbf{x}_i$ does not appear (Breiman, 2001). Because the oob error can be used for validation, a separate validation set is not required.

3.4. Benefits of introducing randomness to quadratic discriminant analysis

There are a number of possibilities to improve classification results by introducing randomness to the classification model. For instance, bagging can enhance the accuracy of a QDA classifier by reducing its variance. In addition, there are three possible advantages for the use of random feature selection. First, as random feature selection encourages diversity it can improve the predictive accuracy of ensembles (Dietterich, 2000). Second, when two features are highly correlated, random feature selection can also reduce the chance of obtaining ill-conditioned covariance matrices. Finally, random feature selection can reduce the impact of noisy or non-Gaussian features on the fit of the model. We explore the benefits of random feature selection and bagging by using simulated data and a part of our real data, where the classes are sufficiently large for QDA and compare the performance of RBA, QDA, and other Bayesian classifiers in detail in Section 6. In Section 7, RBA, NB, and LDA are compared with common classifiers by using all of the real data, which suffers from small class sizes.

4. OTHER CLASSIFICATION METHODS TESTED

We also tested the performance of a number of alternative classification methods such as the MLP (Ripley, 1996; Duda *et al*., 2001), the KNN (Hastie *et al.*, 2009), RF (Breiman, 2001), the SVM (Boser, 1992; Cortes and Vapnik, 1995) by using an implementation described by Chang and Lin (2011), and the RBF (Buhmann, 2003). Although RBFs are not classifiers as such, they can be built into a classification method. Here, we have implemented a naïve classifier on the basis of RBFs by fitting an RBF for each class in the training data and classifying the test observations into the class producing the highest RBF score.

5. COMPARISON METHODS FOR CLASSIFIERS

Classifiers were compared using their error rates, that is, the proportions of misclassifications in a test set. The data were split randomly into the training and test subsets. After fitting the classifier to the training set, we estimated its test error, e_{test} from the test data (Hastie *et al.*, 2009, p. 220). This was repeated 100 times to obtain the standard deviation of the error rate.

In addition to the overall error rate, we also rank the classifiers by using class-specific error rates obtained from the confusion matrix. As a measure, we use the sum of the weighted distances between the elements of the confusion matrix of each classifier and the confusion matrix of a perfect classification, which is a diagonal matrix. The distances are weighted with an expert defined loss matrix. More mathematically, let $L(c, \hat{c})$ be the loss function, where the correct class is c and the predicted class $\hat{c}$. The weighted distance between matrices is then

$$D(\mathbf{R}, \mathbf{T}) = \sum_{i=1}^{C} \sum_{j=1}^{C} L(c_i, \hat{c}_j)|r_{ij} - t_{ij}|,$$

where $\mathbf{R}$ is the confusion matrix of a classifier with elements r_{ij}, $\mathbf{T}$ is the confusion matrix of a perfect classification with elements t_{ij}. Without the loss functions, ranking the classifiers on the basis of these distances would be the same as ranking them on the basis of their error rates.

With the help of correspondence analysis (CA; e.g. in Greenacre, 2007), we explore how misclassification differs between methods in relation to each other and the different classes. To do so, we use the class-specific error rate type of information obtained from the diagonal of the confusion matrix of each classifier.

6. COMPARISON OF RANDOM BAYES ARRAY WITH OTHER BAYESIAN CLASSIFIERS

6.1. Real data

The use of the classical QDA model requires that the class sizes are larger than the number of features. Therefore we are able to include only 18 of the original 35 classes in the comparison of classic QDA and RBA based on QDA models (see Table A1). We restrict the comparison with RBA_{ave}, because it provides the lowest error rate with this smaller data of 18 classes. We also compare the performance of RBA_{ave} with other Bayesian classifiers: NB, RNB, and LDA.

The reduced data consisting of 5078 observations and 18 classes was split into training and test sets 100 times. Each classifier used the same training data. NB, LDA, and QDA used all 64 features and RNB used 300 single NB classifiers, each with 15 randomly selected features. RBA was built by randomly selecting $m = 24$ features for each of the 500 single QDA classifiers. For both RNB and RBA, the number of single classifiers in the ensemble and the number of randomly selected features, were determined using the oob error. The final classification decision for RBA_{ave} was based on the averaged posterior; see Equation (2).

All of the classifiers were built using R (R Development Core Team, 2011). The estimation of QDA and RBA was performed using the qda-function (Venables and Ripley, 2002) on the basis of a spherical transformation of the data. NB and RNB were estimated using the naiveBayes-function and LDA using the lda-function (Venables and Ripley, 2002).

The lowest classification error is produced by RBA_{ave} (Table 1). We propose that this is because of the complexity of our data and the fact that some of the features may suffer from non-normalities. The difference in the error rates is clearer with a 50/50 (i.e. training/test) split, demonstrating that the use of bagging and random feature selection improves results even more with a smaller training set. A bagged QDA classifier by using all 64 features and a 80/20 split produces mean(e_{test}) = 0.1653, SD(e_{test}) = 0.0135. The performance of the QDA classifier is clearly not enhanced by bagging alone, but random feature selection is needed to improve the classification results, as the minimum classification error is reached with 24 features (Figure 1). Similarly to Prinzie and van den Poel (2007), the RNB also did not substantially improve NB's error rate with our data. Because of the poor performance of RNB with the real data, we did not use it in the performance tests of Section 7 and the CA of Section 7.4.

We injected outliers into the data by randomly switching the class labels of 5% of the training observations, the error rate for QDA rises to mean(e_{test}) = 0.148, SD(e_{test}) = 0.016 in the 80/20 split, whereas for RBA_{ave} mean(e_{test}) = 0.126 and SD(e_{test}) = 0.012. With 5% of outliers, in the 50/50 split into training and test data, the error rate for QDA is mean(e_{test}) = 0.200, SD(e_{test}) = 0.015 and for RBA_{ave}

Table 1. Means and standard deviations of classification errors using 80/20 and 50/50 division for a training and test data in 100 splits of the real data

	80/20		50/50	
	$\overline{e_{test}}$	$SD(e_{test})$	$\overline{e_{test}}$	$SD(e_{test})$
NB	0.440	0.014	0.439	0.009
RNB	0.440	0.015	0.437	0.010
LDA	0.206	0.011	0.215	0.008
QDA	0.122	0.010	0.180	0.013
RBA_{ave}	0.118	0.009	0.127	0.006

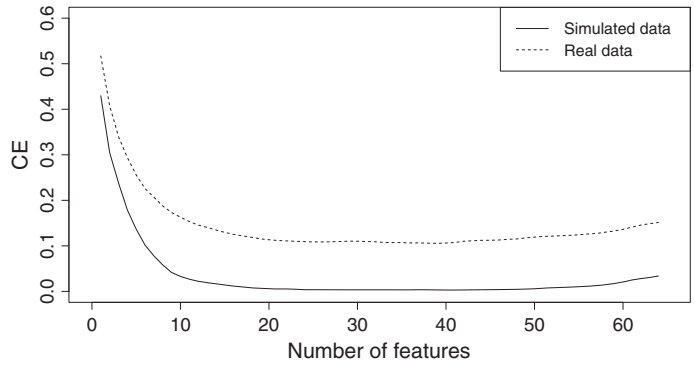

Figure 1. The mean classification error (CE) is plotted against the number m of features used in RBA by using 80/20 division for a training and test data in 100 random splits at each value of m

mean(e_{test}) = 0.133, SD(e_{test}) = 0.009. Bagging and random feature selection clearly reduced the effect of outliers on error rates obtained with RBA.

6.2. Simulation example

To further investigate the performance of RBA compared with QDA, we simulated a setup of the 18 largest classes of the real data. First, from each class of the real data, we estimated the (1×64) mean vector $\hat{\boldsymbol{\mu}}_c,\ c = 1,\ldots,18$, and the (64×64) covariance matrix $\hat{\boldsymbol{\Sigma}}c,\ c = 1,\ldots,18$. Apart from the aforementioned estimates, we used the same class sizes $n_c,\ c = 1,\ldots,18$ as in the real data in the simulation as well. For each class c, we simulated n_c observations from the multivariate normal distribution of 64 features with the mean vector $\hat{\boldsymbol{\mu}}_c$ and covariance matrix $\hat{\boldsymbol{\Sigma}}_c$. We obtained 5078 simulated observations equaling the number of observations in the real data of 18 classes. As the data now satisfied the multivariate normal distribution required by QDA, it should be the optimal classifier.

The simulated data was split into a training and test set 100 times, and the classifiers were built as with the real data. With an 80/20 split into training and test data, a single QDA classifier produced a slightly lower error rate (Table 2) than RBA, which is to be expected. However, with a 50/50 split (Table 2), RBA actually achieves a lower classification error than a single QDA classifier, demonstrating the benefits of bagging combined with random feature selection in the case of smaller training data.

As with the real data, we investigated how the number of features affects the error rate of RBA, and whether bagging alone would enhance the QDA classifier (Figure 1). Again, the error rate for QDA with bagging ($m = 64$), mean(e_{test}) = 0.034, SD(e_{test}) = 0.001 is higher than for RBA with 24 randomly selected features, mean(e_{test}) = 0.004, SD(e_{test}) = 0.001.

7. PERFORMANCE COMPARISONS OF THE CLASSIFIERS

7.1. Setup

For the performance comparison, we used all 6814 observations of the 35 taxa. The data set was randomly split into half training and test data. The following classification methods were compared: MLP, KNN, NB, RNB, RBF, SVM, RF, LDA, and RBA with all of its modifications.

Table 2. Means and standard deviations of classification errors by using 80/20 and 50/50 division for a training and test data in 100 splits of simulated multivariate normal data

	80/20		50/50	
	$\overline{e_{\text{test}}}$	SD(e_{test})	$\overline{e_{\text{test}}}$	SD(e_{test})
NB	0.367	0.014	0.370	0.008
RNB	0.366	0.014	0.367	0.009
LDA	0.184	0.012	0.192	0.006
QDA	0.002	0.001	0.036	0.012
RBA_{ave}	0.004	0.002	0.010	0.003

NB, naïve Bayes; RNB, random naïve Bayes; LDA, linear discriminant analysis; QDA, quadratic discriminant analysis; RBA, random Bayes array.

Table 3. Classification errors using 50/50 split for the training and test data

Classifier	$\overline{e_{\text{test}}}$	SD(e_{test})
NB	0.498	0.007
KNN	0.374	0.007
MLP	0.285	0.021
LDA	0.276	0.007
RF	0.258	0.008
RBF	0.223	0.007
SVM	0.200	0.007
$\text{RBA}_{\text{basic}}$	0.195	0.007
$\text{RBA}_{\text{weight}}$	0.195	0.006
RBA_{ave}	0.193	0.007
$\text{RBA}^{*}_{\text{basic}}$	0.191	0.007
$\text{RBA}^{*}_{\text{weight}}$	0.190	0.007
$\text{RBA}^{*}_{\text{ave}}$	0.188	0.007

NB, naïve Bayes; KNN, k-nearest neighbor; MLP, multilayer perceptron; LDA, linear discriminant analysis; RF, random forest; RBF, radial basis function; SVM, support vector machine; RBA, random Bayes array.

The MLP was built holding out 20% of the training data for validation. The best model was selected on the basis of the classification error of the validation set and had one hidden layer with 49 hidden units. The activation function used was sigmoid, and the model was built using batch learning.

The KNN classifier was built using 10 inverse distance weighted nearest neighbors. The optimal number of nearest neighbors was determined with leave-one-out cross-validation.

The NB and LDA models were formed using all of the 64 features, which were assumed Gaussian.

The RBF classifier was built using polyharmonic splines of rank one. The chosen basis function is nonparametric and smooth requiring no validation or optimizing. The features of observation $\mathbf{x}_i$ were scaled to $[-10, 10]$ and the class was recoded as a vector $[-1, \ldots, -1, 1, -1, \ldots, -1]$, where the correct class had value 1, and the other classes had value -1. The classifier was built fitting RBF for each class in the training data and the test observations were classified into the class producing the highest RBF score.

The SVM construction followed Hsu *et al.* (2003) and was built using cost $= 2048$ and RBF (Gaussian) kernel with $\gamma = 3.125e-2$. The parameter values were determined using 10-fold cross validation within the training data.

The RF was built using 700 trees with 20 features randomly selected at each node. The parameter values were obtained using oob error as an estimate for the classification error.

All the different versions of RBA were built using 300 QDA classifiers with 11 random features each. The number of features was limited, because some of the taxonomical groups had too few observations. For the same reason the QDA classifier by using all 64 features cannot be implemented on our 35 class data. The variable importance used as weight (Figure A1) for the features was calculated from the oob observations of the training data.

The RF (Breiman *et al.*, 2007), the LDA (Venables and Ripley, 2002), the NB, the SVM (Chang and Lin, 2011), and the RBA classifiers were built using R (R Development Core Team, 2011), and the MLP and the KNN were computed with Weka (Hall *et al.*, 2009). The RBF is a personal code written with MATLAB (The MathWorks Inc. 2009).

7.2. Error rates and the effect of costs

The classifiers were compared according to their error rates. The data were split randomly into the training and test set, both including 3407 individuals. Each classifier was built using the training data, and the error rate was estimated from the test data. This procedure was repeated 100 times to obtain the mean and standard deviation of the error rates. From Table 3, we see that the lowest classification errors are achieved with the RBAs. SVM and RBF also work quite well. The most ineffective classification methods for our macroinvertebrate data are the simple and popular methods of KNN and NB. As the modifications of RBA do not differ much from each other, we will use only one of them, $\text{RBA}^{*}_{\text{ave}}$ to evaluate CA results.

From Table 3, we see that it is advantageous to use the variable importance as a weight for features. To investigate the benefits of our method, we performed a comparison in which the data was randomly split into equal sized training and test sets. Variable importance for each feature was obtained with RBA by using 300 random Bayes classifiers each with 11 randomly selected features and utilizing the oob observations of the training data. Next, $\text{RBA}_{\text{basic}}$, RBA_{ave}, and $\text{RBA}_{\text{weight}}$ were built on the training data, using the following: (i) variable importance weights; (ii) z-score weights; and (iii) z-scores to eliminate half of the features. For each scenario, the error rate was estimated from the test data. This was repeated 100 times to obtain the mean and standard deviation of the error rates. From Table 4, we see that the lowest classification errors for all versions of RBA are achieved by using the variable importance weights.

Table 4. Classification errors by using variable importance scores

	Classifier	$\overline{e_{\text{test}}}$	SD(e_{test})
$\text{RBA}_{\text{basic}}$	1) Variable importance as weights.	0.191	0.007
	2) Z-score as weights.	0.194	0.006
	3) Thirty-two features eliminated.	0.207	0.017
RBA_{ave}	1) variable importance as weights.	0.188	0.007
	2) Z-score as weights.	0.191	0.007
	3) Thirty-two features eliminated.	0.204	0.016
$\text{RBA}_{\text{weight}}$	1) Variable importance as weights.	0.190	0.007
	2) Z-score as weights.	0.192	0.006
	3) Thirty-two features eliminated.	0.205	0.016

Table 5. Loss matrix of classification errors

		Predicted taxa		
		Index	Non-index	Non-typical
Correct taxa	Index	0.5	1	1
	Non-index	1	0.3	0.8
	Non-typical	1	1	0.3

We devised a loss matrix of misclassifications on the basis of an expert opinion approximation of error severity to biomonitoring (Table 5). In the loss matrix, index taxa are taxa used to calculate indices from samples and non-typical taxa are taxa not expected to be found in a sample. Utilizing this matrix, we explored the loss-weighted distances from the confusion matrices of the classifiers to the confusion matrix of a perfect classification for small peatland and woodland streams. Accounting for approximate losses of different types of classification errors did not change the ranking of the top four methods, (RBA, SVM, RBF, RF).

7.3. Examples of observed misclassification results

Our theoretical investigation in Section 6 of the effect of outliers on classification was based on the fact that this type of biological data is likely to include a significant proportion of suboptimal images. To illustrate this, we chose images of the taxa *Elmis aenea, Hydropsyche siltalai*, and *Micrasema setiferum* as examples (Figure 2). The three images on the left side were classified correctly with all classifiers used in the current work. In contrast, the three images on the right side were misclassified.

Although some misclassified cases are intuitively obvious, others are less obvious as they are based on subtle differences in size or color. The similar looking *H. siltalai* seem to be misclassified because of their size. The individuals on the right side of the middle row are much smaller than those on the left side. This type of misclassification is due to the limited size range covered by the training data and may be overcome by better future representation of taxa size ranges. For the *M. setiferum*, the difference seems to be in the shape of the individuals. The misclassified *M. setiferum* are wider or narrower, lower or less circular than those classified correctly. We believe that improvements in classification could be achieved if multiposture images and a suite of more tailored features of each specimens were available and used for classification.

7.4. Correspondence analysis on classification of macroinvertebrates

To gain further insights into the behavior of individual classifiers and their relationships to one another and identified taxa, we performed CA. CA was run on the table of correctly classified individuals (Table A1) with taxa in rows and classification methods in columns. Note that specific individuals can be associated with more than one classification method or none at all. Further, the grand total of the table is not the total number of observations but the number of "correctly classified individual to classification method" pairs (see the Hatco data in Hair *et al.*, 1998, for a similar approach). The total inertia of the table was 0.021 and the first, and second principal axis explained 75.6 % of the total inertia. In the resulting CA biplot classification methods are plotted close to taxa they classify well (Figure 3). Given the number of classes in our data set, plotting classes in vertex coordinates would make it difficult if not impossible to visualize differences between the classification methods. Therefore, both classes and classifiers were plotted in principal coordinates.

The first principal axis separates the NB classifier from the rest, whereas the second principal axis separates LDA, SVM, and RBA_{ave} into a group of their own. This is to be expected, because NB is the weakest classifier, and SVM and $\text{RBA}^{*}_{\text{ave}}$ produce the best results. The Bayesian classifiers LDA and $\text{RBA}^{*}_{\text{ave}}$ group close because of similar behavior. The second principal axis also separates MLP and RBF from

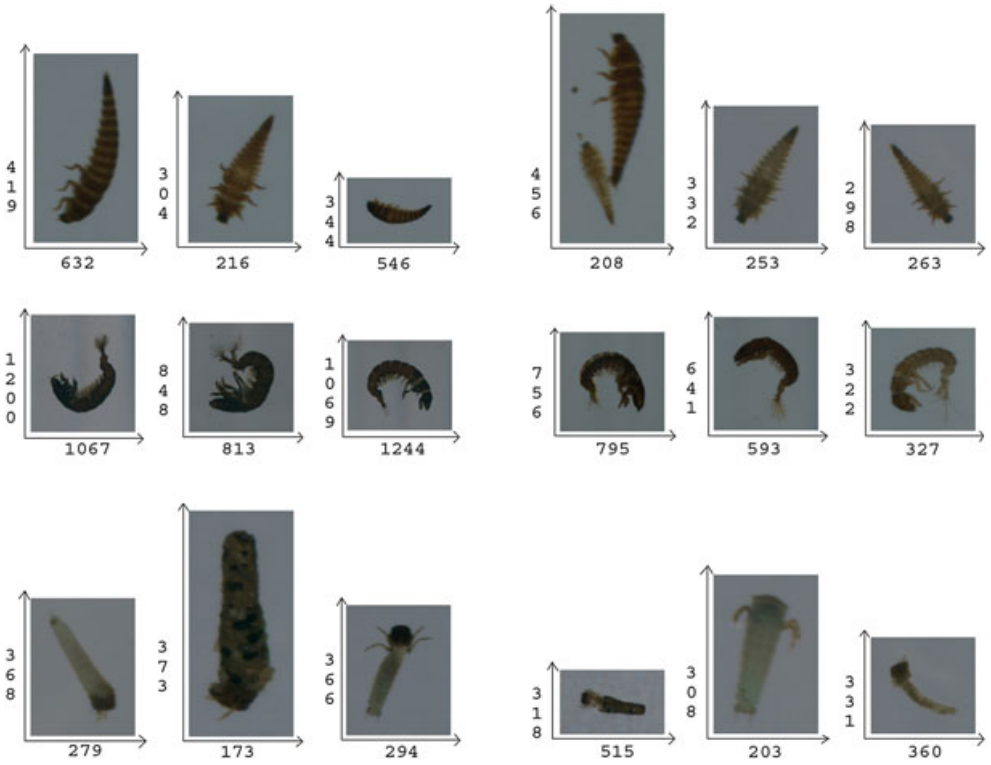

Figure 2. The three images on the left side are individuals that are always classified correctly. The three images on the right side are individuals that are misclassified. Taxonomical groups from the top: *Elmis aenea, Hydropsyche siltalai*, and *Micrasema setiferum*. The images are all scaled to be 180 pixels wide. The actual pixel values are shown next to each image

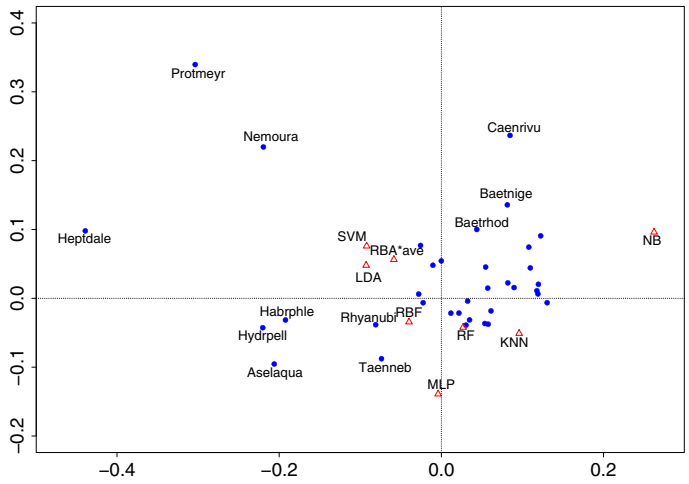

Figure 3. Correspondence analysis biplot with row and column variables in principal coordinates. For the ease of interpretation, only some of the classes are labeled

KNN and RF. The most challenging taxa, *Protonemura meyeri*, *Nemoura sp.* and *Heptagenia dalecarlica* clearly differ from other taxa and group especially far away from MLP, KNN and NB.

Protonemura meyeri and *Nemoura sp.* lie closest to LDA, SVM and RBA^*_{ave}, because these methods classify these taxa the best (Figure 3, Table A1). LDA and SVM are closer to *H. dalecarlica* than $\text{RBA}^*_{\text{ave}}$, because $\text{RBA}^*_{\text{ave}}$ produces a higher error rate for *H. dalecarlica* than the former. *Asellus aquaticus*, *Hydropsyche pellucidula*, and *Habrophlebia sp.* lie on the opposite side of the plot from NB and KNN, indicating especially poor classification results. Average classification error is not well represented by CA, because RF and KNN group close even though RF produces a much lower classification error. The same applies for $\text{RBA}^*_{\text{ave}}$ and LDA, which are in close proximity on the biplot despite lower error rates for $\text{RBA}^*_{\text{ave}}$.

To see which classifiers produced predicted taxa proportions closest to the actual ones, we performed CA on a table of predicted individuals where we included a column with the correct number of individuals in each taxa. With CA on this table, RBA and SVM plotted

close to the correct classification, with SVM being the closest. Although SVM made more misclassifications than RBA, its predicted class proportions were the closest to the correct ones. To investigate the influence of a rarefied data set, we also explored CA in a situation when all the classes in the test data have the same number of observations. However, the results did not change much, so the graphs are not shown.

8. CONCLUSIONS

We tested a variety of classifiers in automated identification of freshwater macroinvertebrate taxa. We developed a novel classification method, RBA, and demonstrated its benefits compared with the traditional QDA and other Bayesian classifiers by using real and simulated data. RBA performance in comparison with several popular methods was assessed using our complex benthic macroinvertebrate image data set with 64 features, 35 classes, and 6814 observations. The input data was deliberately kept realistic to ensure conservative estimates of automated identification results. The image data included routinely encountered outliers in normal macroinvertebrate samples, that is, overlapping or severely damaged individuals. Also the classification task was far from standard, because not all of the easily extractable features used were relevant to the classification task. Based on our results with this challenging data, we are convinced that RBA can overcome many of these challenges with the help of bagging and random feature selection.

Quadratic discriminant analysis is known to be the optimal classifier if we have sufficiently large multivariate normal data. This is rarely the case in practice, especially with biological data. We showed with simulation that even with a multivariate normal data set, RBA produces a lower error rate than QDA when the size of the training data decreases. The same effect was observed with noisy and partially non-Gaussian real data of 18 classes.

Classification error rates for a larger data of 35 classes showed that RBA produces a lower error rate than the other classifiers tested. The best results were obtained with a particular variant, $\text{RBA}^{*}_{\text{ave}}$. In cases of data with many irrelevant features, it is best to calculate the variable importance and use it to weight the features as this produces lower classification errors without the need to eliminate features.

If our automated identification results are put into a biomonitoring context, it should be acknowledged that a low error rate alone does not necessarily equate with reliable information. Rather, we need to understand the error structure of the chosen automated classification method to avoid introducing costly artifactual errors that propagate further into decision making. To assess this risk, we explored the effects of different approximate losses associated with our classifiers by weighting the distances between the produced confusion matrices and the optimal confusion matrix with a loss matrix. However, weighting automated misidentification with the severity of this error (i.e. loss) did not change the ranking of the top four classifiers. Our novel RBA method still performed best, followed by SVM, RBF, and RF.

The CA-based visualization of classifier methods in relation to taxa identification results confirmed NB as the worst and RBA, SVM, and LDA as the best performing methods.

In conclusion, our novel variant of QDA, RBA, provides tools to utilize the benefits of QDA with less strict assumptions in situations when the data suffer from the curse of dimensionality. RBA provided very promising results on this complex single posture macroinvertebrate image data set. We anticipate further improvements in overall performance when multiple posture data sets become available.

Acknowledgements

All image data was created at the Finnish Environment Institute. We kindly thank Timo Ruokonen for providing the gastropod specimens, Ville Tirronen for his help with the SVM, and Tomi Haanpää for the RBF classification results. We would also like to warmly thank Thomas Dietterich, two anonymous referees, and the associate Editor for useful comments made on an earlier draft of this article, as well as Joe F. Hair and Antti Penttinen for valuable discussions and suggestions. Financial support from the Maj and Tor Nessling Foundation and COMAS graduate school (to JÄ) and the University Alliance Finland project STATCORE (to JÄ and SK) is appreciated.

REFERENCES

Archer KJ, Kimes RV. 2008. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis* **52**: 2249–2260.

Boser B, Guyon I, Vapnik V. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*. ACM Press: New York; 144–152.

Breiman L. 1996. Bagging predictors. *Machine Learning* **24**: 123–140.

Breiman L. 2001. Random forests. *Machine Learning* **45**: 5–32.

Breiman L, Cutler A. 2008. Random forests - classification manual. http://www.math.usu.edu/~adele/forests/.

Breiman L, Cutler A, Liaw A, Wiener M. 2007. Breiman and Cutler's random forests for classification and regression. R package version 4.6-7 http://CRAN.R-project.org/.

Buhmann MD. 2003. Radial Basis Functions: Theory and Implementations (Ciarlet PG, Iserles A, Kohn RV, Wright MH, eds), Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press: West Nyack, New York.

Chang CC, Lin CJ. 2011. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* **2**(3): 27:1–27.

Cortes C, Vapnik V. 1995. Support-vector networks. *Machine Learning* **20**: 273–297.

Dietterich TG. 2000. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning* **40**(2): 139–158.

Duda RO, Hart PE, Stork DG. 2001. Pattern Classification, (2nd edn). Wiley: New York.

Friedman JH. 1989. Regularized discriminant analysis. *Journal of the American Statistical Association* **84**(405): 165–175.

Greenacre M. 2007. Correspondence Analysis in Practice, (2nd edn). Chapman & Hall: London.

Haase P, Pauls SU, Schindehütte K, Sunderman A. 2010. First audit of macroinvertebrate samples from an EU Water Framework Directive monitoring program: human error greatly lowers precision of assessment results. *Journal of the North American Benthological Society* **29**(4): 1279–1291.

Hair JF, Tatham RL, Anderson RE, Black W. 1998. Multivariate Data Analysis, (5th edn). Prentice Hall: Upper Saddle River.

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. 2009. The Weka data mining software: an update. *SIGKDD Explorations* **11**(1): 10–18.

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

Hsu CW, Chang CC, Lin CJ. 2003. A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/ [accessed 2003].

John GH, Langley P. 1995. Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers: San Mateo; 338–345.
Joutsijoki H, Juhola M. 2012. Dagsvm vs. dagknn: an experimental case study with benthic macroinvertebrate dataset. In *Proceedings of the 8th International Conference on Machine Learning and Data Mining (MLDM 2012)*, Vol. 7376, Springer Lecture Notes in Artificial Intelligence. Springer-Verglag: Berlin, Heidelberg; 439–453.
Kiranyaz S, Ince T, Pulkkinen J, Gabbouj M, Ärje J, Kärkkäinen S, Tirronen V, Juhola M, Turpeinen T, Meissner K. 2011. Classification and retrieval on macroinvertebrate image databases. *Computers in Biology and Medicine* **41**(7): 463–472.
Langley P, Sage S. 1994. Induction of selective Bayesian classifiers. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann: Seattle; 399–406.
Lytle DA, Martínez-Muñoz G, Zhang W, Larios N, Shapiro L, Paasch R, Moldenke A, Mortensen EN, Todorovic S, Dietterich TG. 2010. Automated processing and identification of benthic invertebrate samples. *Journal of the North American Benthological Society* **29**(3): 867–874.
MEA. 2005. Millennium Ecosystem Assessment.
Michie D, Spiegelhalter D, Taylor C (eds). 1994. Machine Learning, Neural and Statistical Classification, Ellis Horwood Series in Artificial Intelligence. Ellis Horwood: New York.
Prinzie A, van den Poel D. 2007. Random multiclass classification: generalizing random forests to random MNL and random NB. In *Volume 4653 of Lecture Notes in Computer Science (LNCS)*. Springer Verlag: Berlin; 349–358.
R Development Core Team. 2011. R: a language and environment for statistical computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org.
Rasband WS. 1997. Imagej manual. http://rsbweb.nih.gov/ij/docs/menus/analyze.html.
Rasband WS. 1997–2010. Imagej, U.S. National Institutes of Health, Bethesda, Maryland, USA. http://rsb.info.nih.gov/ij/.
Ripley BD. 1996. Pattern Recognition and Neural Networks. Cambridge University Press: Cambridge.
Ärje J, Kärkkäinen S, Meissner K, Turpeinen T. 2010. Statistical classification methods and proportion estimation – an application to a macroinvertebrate image database, *Proceedings of the 2010 IEEE Workshop on Machine Learning for Signal Processing (MLSP)*: Kittilä, Finland.
Strobl C, Zeileis A. 2008. Danger: High power! – exploring the statistical properties of a test for random forest variable importance, In *Proceedings of the 18th International Conference on Computational Statistics*: Porto, Portugal; 59–66.
The MathWorks Inc. 2009. Matlab version 7.9.0, Natick, Massachusetts. http://www.mathworks.se/products/matlab.
Venables WN, Ripley BD. 2002. Modern Applied Statistics with S. Springer: New York.

APPENDIX

Table A1. The number of correctly classified individuals for each classification method

	n	MLP	KNN	NB	RBF	SVM	RF	RBA^{*}_{ave}	LDA
Ameletus inopinatus	54	47	40	33	44	48	48	46	46
Arctopsyche ladogensis	24	13	16	18	18	19	16	16	16
Asellus aquaticus	180	141	84	33	152	143	123	140	129
Baetis muticus	146	113	105	85	126	116	118	125	105
Baetis niger group	89	42	46	66	67	73	72	82	55
Baetis rhodani	67	35	26	38	39	38	42	55	41
Bithynia tentaculata	136	134	133	101	133	132	130	131	122
Caenis rivulorum	29	16	5	22	12	19	15	20	22
Callicorixa wollastoni	40	38	33	37	38	33	39	35	36
Ceratopsyche silfvenii	116	84	78	84	84	94	88	89	65
Ceratopogonidae	33	32	22	30	31	32	30	30	28
Cheumatopsyche lepida	57	52	32	34	43	51	40	49	49
Diura sp.	122	104	89	80	104	98	100	98	90
Elmis aenea	92	82	77	66	82	77	81	86	77
Ephemerella aurivillii	103	65	55	53	67	73	68	73	82
Seratella ignita	51	42	31	34	45	39	47	49	39
Ephemerella mucronata	29	25	20	24	23	22	25	25	22
Habrophlebia sp.	41	25	10	8	17	21	16	24	25
Heptagenia dalecarlica	100	23	12	1	32	50	25	30	45
Hydraena sp.	61	60	59	61	60	61	61	59	60
Hydropsyche pellucidula	59	26	7	8	18	28	17	24	19
Hydropsyche saxonica	86	29	32	27	39	34	40	47	40
Hydropsyche siltalai	120	76	71	76	88	89	92	102	79
Isoperla sp.	267	205	207	152	230	223	217	216	187
Leuctra sp.	74	47	50	43	53	52	54	55	53
Limnius volckmari	96	87	72	59	90	92	82	92	89
Micrasema gelidum	44	35	37	42	40	43	42	44	41
Micrasema setiferum	151	146	140	107	147	146	145	147	138
Myxas glutinosa	108	100	89	96	106	104	102	108	98
Nemoura sp.	134	24	46	26	67	82	37	77	77

Table A1. Continued.

	n	MLP	KNN	NB	RBF	SVM	RF	RBA$^{*}_{ave}$	LDA
Sphaeridae sp.	29	28	20	29	28	28	25	27	26
Protonemura intricata	168	84	84	78	112	116	99	117	115
Protonemura meyeri	55	8	9	9	16	31	18	40	31
Rhyacophila nubila	114	96	55	51	99	93	79	100	79
Taeniopteryx nebulosa	332	314	225	140	310	291	309	322	253
	3407	2478	2117	1851	2660	2691	2542	2780	2479

MLP,multilayer perceptron; KNN, k-nearest neighbor; NB, naïve Bayes; RBF, radial basis function; SVM, support vector machine; RF, random forest; RBA, random Bayes array; LDA, linear discriminant analysis.

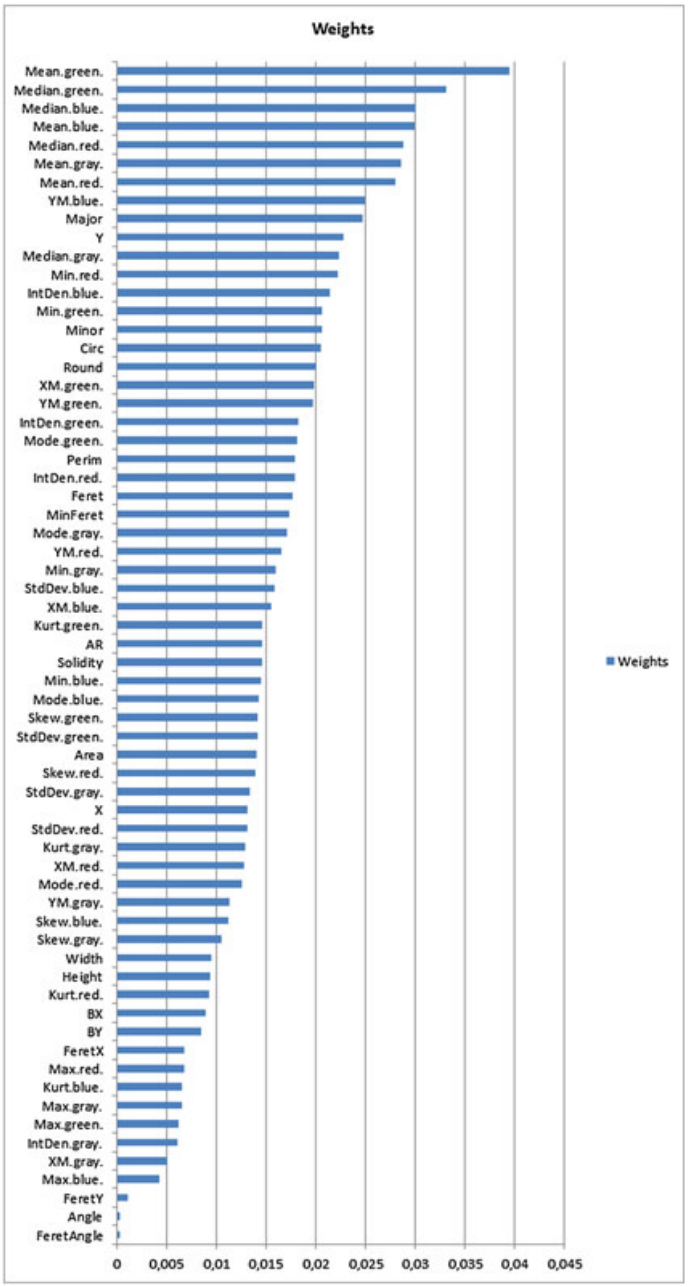

Figure A1. Weights for the features calculated from the variable importance

IV

Ärje, J., Kärkkäinen, S., Meissner, K., Iosifidis, A., Ince, T., Gabbouj, M. & Kiranyaz, S. The effect of automated taxa identification errors on biological indices. Submitted.

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]

[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 most biomonitoring, the target is to determine the status of ecosystems based on several biological indices. Taxa identification errors can, however, have strong effects on the indices and thus on the determination of the ecological status. To increase cost-efficiency, computer-based taxa identification for macroinvertebrate image data has recently been developed. We examine the effect of eleven classification methods in the case of macroinvertebrate image data and show how those errors propagate into different 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.

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 address this problem, 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.

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 which of the indices are least biased by classification errors.

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	$C_\lambda = \frac{2\sum_{h=1}^{c} X_h Y_h}{(D_x + D_y)nm}$	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 use 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.

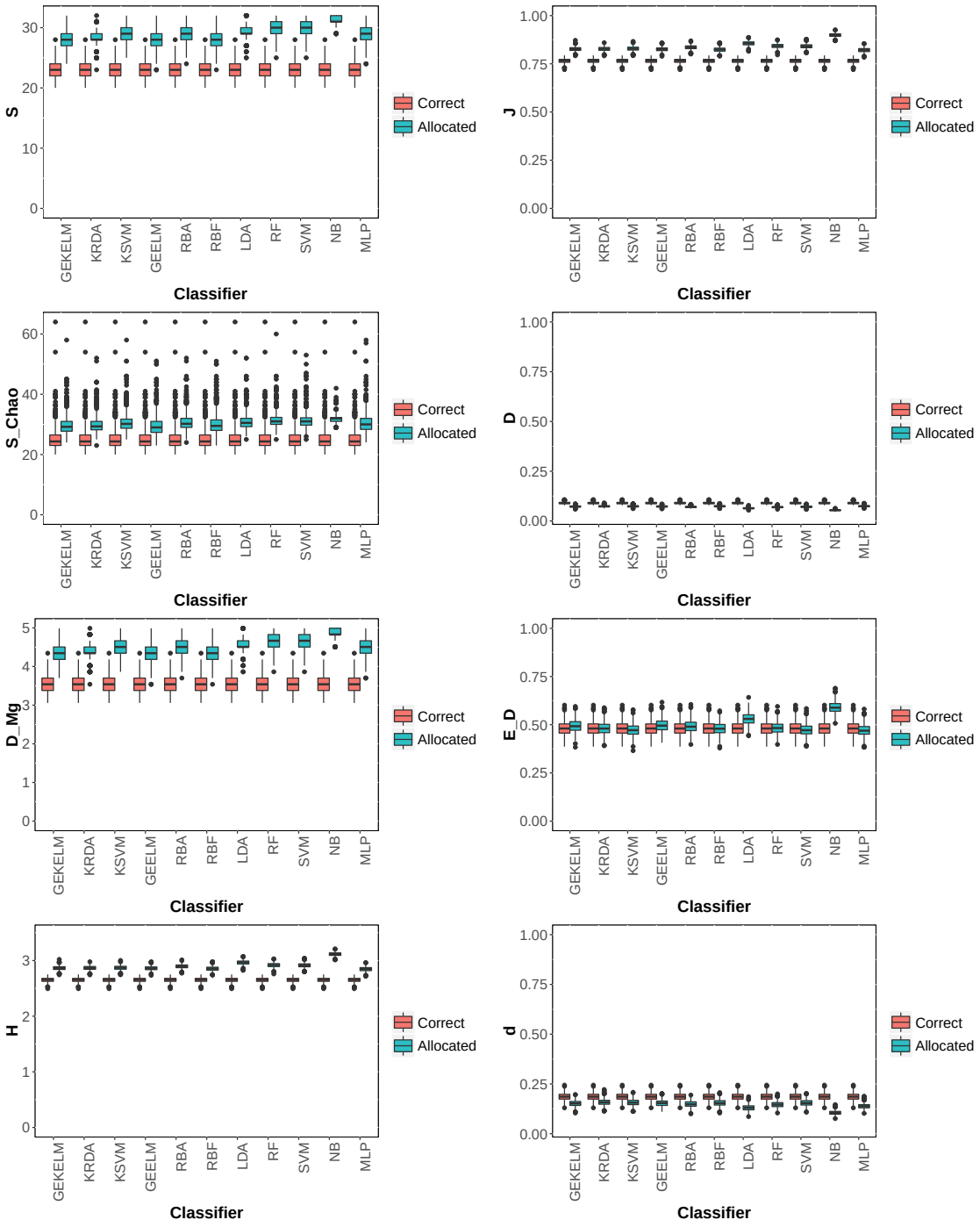

Figure 1: The effect of classification errors on richness, diversity, evenness and dominance indices for medium-sized non-reference peatland streams in southern Finland. Here, $\mathbf{X} \sim M(500, \mathbf{p})$. The red boxplots represent the correct index value distributions. The blue boxplots represent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

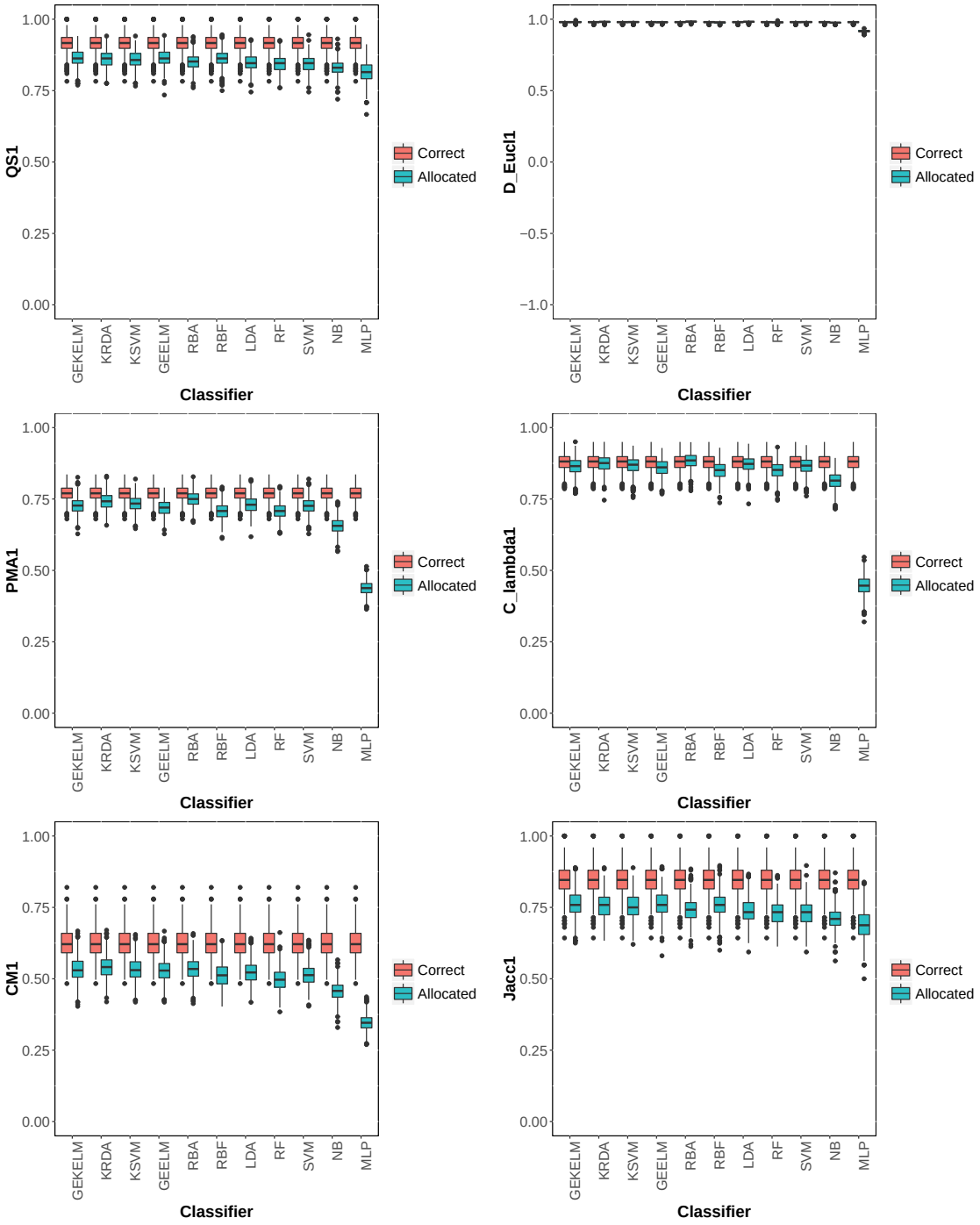

Figure 2: The effect of classification errors on similarity indices for medium-sized non-reference peatland streams in southern Finland when the reference sample is assumed to be known. Here, $\mathbf{X} \sim M(500, \mathbf{p})$ and $\mathbf{Y} \sim M(500, \mathbf{q})$. The red boxplots represent the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

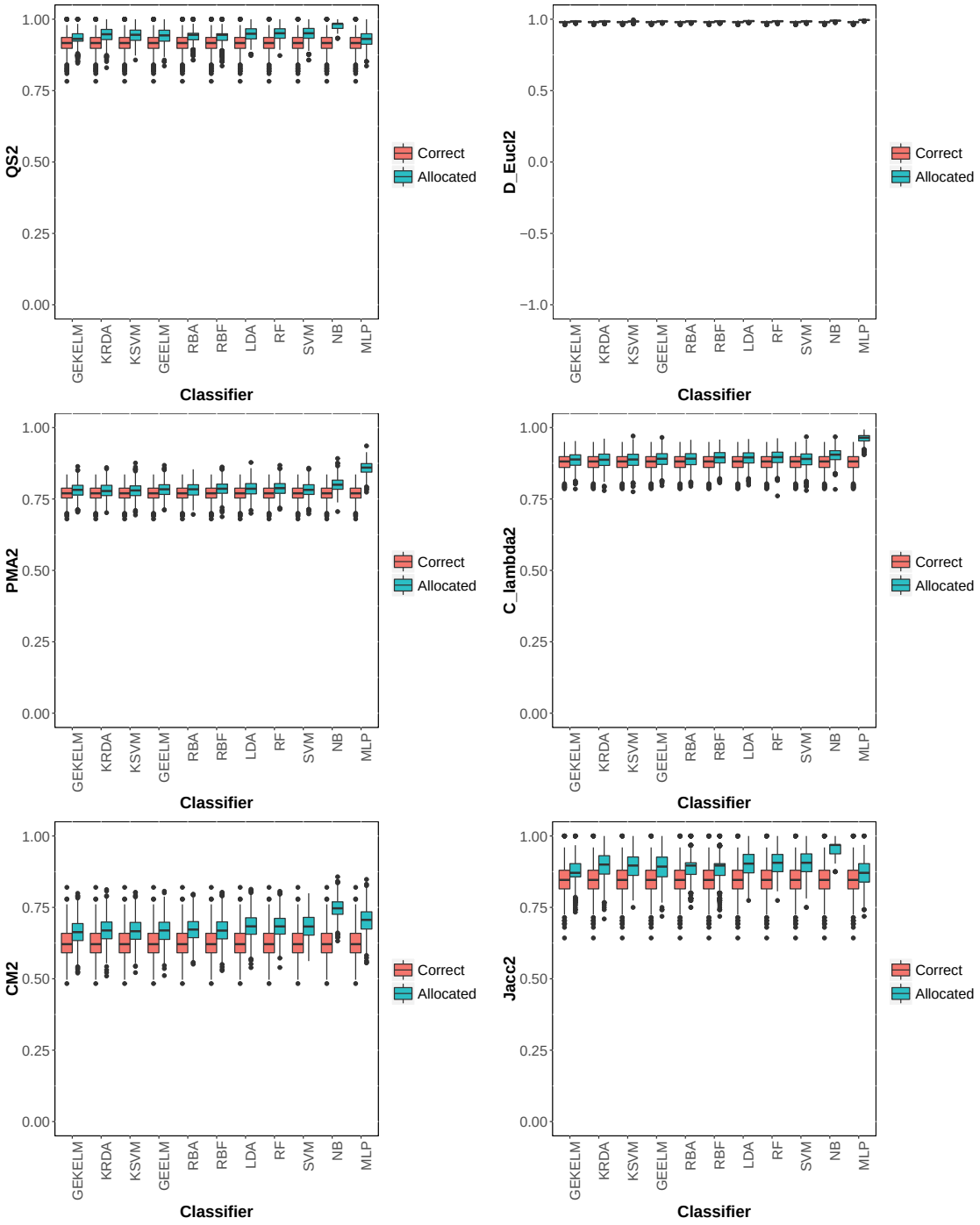

Figure 3: The effect of classification errors on similarity indices for medium-sized non-reference peatland streams in southern Finland when both samples may contain classification errors. Here, $\mathbf{X} \sim M(500, \mathbf{p})$ and $\mathbf{Y} \sim M(500, \mathbf{q})$. The red boxplots represent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

larity 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.

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.

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

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.

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.

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.

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

— (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.

— (2014b) On the kernel extreme learning machine classifier. *Pattern Recogn. Lett.*, **54**, 11–17.

— (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.

— (1975) *Ecological diversity*. New York: Wiley InterScience.

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.

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.

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.

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	0.36	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